

# JSTOR PLANTS Handbook



1. Introduction.....	2
1.1. About the Project.....	2
1.2. Joining, Training, and Support.....	2
1.3. Setting up the Scanning Station.....	4
1.4. Location of Scanning Station.....	4
1.5. Room Lighting.....	5
1.6. Unpacking the HerbScan Frame.....	5
1.7. Unpacking the Scanner.....	6
1.8. Assembling the HerbScan Frame and Scanner.....	7
1.9. Testing.....	8
1.10. Installing EpsonScan.....	8
1.11. Connecting the Scanner.....	11
1.12. Installing Adobe Photoshop.....	11
1.13. Workflow.....	13
2. Specimen Databasing.....	13
2.1. Introduction.....	13
2.2. Barcoding.....	14
2.3. Databasing Software.....	15
2.4. Databasing Methods.....	16
2.5. Database Fields.....	16
3. Imaging of Specimens.....	22
3.1. Standard Specimen Sheet Layout.....	22
3.2. Image Format/Output.....	24
3.3. File Naming Conventions.....	25
3.3.1. One Specimen on a Single Sheet.....	25
3.3.2. Detailed Capture or Additional Scans of a Single Sheet.....	25
3.3.3. Multiple Specimens on a Single Sheet.....	26
3.3.4. One Specimen on Multiple Sheets.....	26
3.4. Non-standard Specimens.....	27
3.4.1. Packet/Capsule/Envelope.....	27
3.4.2. Reverse of Specimen Sheet.....	27
3.4.3. Additional Materials.....	28
3.5. Step-by-Step Scanning Instructions.....	28
4. Image Quality Control.....	35
4.1. During Scanning.....	35
4.2. Quality Control Post Scanning.....	35
4.2.1. Initial Rapid Checks.....	35
4.3. Check for Duplicates.....	36
4.4. Check for Components.....	36
4.5. Check Images for Scanning Artifacts.....	36
4.5.1. Pixilation.....	36

4.5.2.	Vertical Lines .....	37
4.5.3.	Color Separation .....	38
4.5.4.	Glue on Scanner Glass .....	39
4.5.5.	Green Cast .....	39
4.5.6.	Other Artifacts (acceptable).....	40
4.6.	Check Focus.....	42
4.7.	Check Scanning Settings .....	42
4.8.	Following Up.....	43
4.9.	Creating B&W GIF Images .....	43
5.	Export.....	48
5.1.	Introduction.....	48
5.2.	Principles .....	49
5.3.	XML Schema.....	49
5.4.	GPI XML Generator .....	50
5.5.	Batches, Datasets, and Units .....	50
5.6.	XML File Name.....	51
5.7.	General XML Formatting Rules .....	52
5.8.	XML Schema Fields.....	53
5.9.	Validating the GPI XML file .....	65
6.	Transfer to JSTOR .....	69
6.1.	Test Batch of Images and Specimen Data.....	69
6.2.	Hard Drives .....	70
6.3.	File Directory Structure .....	70
6.4.	Shipping .....	71
6.5.	Schedule .....	71
7.	Xumba.....	71
7.1.	Access .....	71
7.2.	Checking Reports.....	Fehler! Textmarke nicht definiert.
7.3.	Data and Image Corrections .....	76

## Appendix X - GPI specimen data XML examples

### 1. Introduction

#### 1.1. About the Project

This document serves as the partner handbook for the Global Plants Initiative (GPI) project. It covers all aspects of the project including scanning, databasing, export, and quality control. The keeper of this document is JSTOR. If you have any comments or corrections, please send them directly to [deirdre.ryan@jstor.org](mailto:deirdre.ryan@jstor.org).

GPI's long-term goal is to build a comprehensive online research tool aggregating and linking presently scattered scholarly resources about plants, thereby dramatically improving access for students, scholars, and scientists around the globe. Partners digitize all plant specimens classified as types (includes flowering plants, algae, fungi, lichens, and bryophytes).

#### 1.2. Joining, Training, and Support

Institutions can join by submitting a proposal to the Andrew W. Mellon Foundation. Please send your inquiries to the Smithsonian Tropical Research Institute (Gloria Jovane [jovaneg@si.edu](mailto:jovaneg@si.edu)), or to

the Mellon Foundation directly (Doreen N. Tinajero [dnt@mellon.org](mailto:dnt@mellon.org)). It takes 3 to 9 months to join depending on the location and formalities of your institution and the Foundation.

Once your proposal is accepted by the Andrew W. Mellon Foundation, your institution will receive information on receiving funds, equipment, and training. Depending on your preference and location you may attend training at any one of the training centers at Royal Botanic Gardens Kew, Smithsonian Tropical Research Institute, New York Botanical Garden, Missouri Botanical Garden, or Instituto de Botánica Darwinion. GPI coordinators are ready and able to help you before, during and after your training as you begin production.

Royal Botanic Gardens, Kew  
Richmond, UK  
Kathryn Beck  
(Email: [k.beck@kew.org](mailto:k.beck@kew.org))

New York Botanical Garden  
Bronx, New York USA  
Kat DeWitt  
(Email: [kdewitt@nybg.org](mailto:kdewitt@nybg.org))

Missouri Botanical Garden  
Saint Louis, Missouri, USA  
Rafael Barron  
(Email: [rafael.barron@mobot.org](mailto:rafael.barron@mobot.org))

Smithsonian Tropical Research Institute  
Panama City, Panama  
Nelly Florez  
(Email: [FlorezNA@si.edu](mailto:FlorezNA@si.edu))

Instituto de Botánica Darwinion  
Buenos Aires, Argentina  
Manuel Belgrano  
(Email: [mbelgrano@darwin.edu.ar](mailto:mbelgrano@darwin.edu.ar))

Partners can also find resources online (<http://plants-partners.jstor.org>).

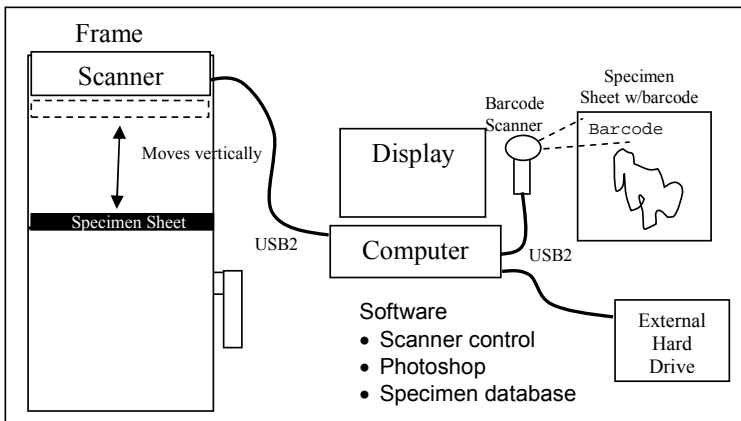


### 1.3. Setting up the Scanning Station

For the digitization of specimens the project utilizes the HerbScan system. Created by the Royal Botanic Gardens, Kew, the HerbScan system includes a mobile frame in which a standard flat-bed scanner is held in an inverted position, with a rising bed mechanism that brings flat upright specimens to the scanning surface. This simple technology has made it possible to create high quality digital scans of flat herbarium material while avoiding damage to the specimens.

The HerbScan was engineered to create high resolution specimen images with minimal damage. The scanning area of the HerbScan is 12.2 x 17.2 inches or 310 x 437 millimeters. It consists of a scanner mounted upside-down in a vertically moveable frame which enables digital scanning without inverting the specimen sheets. A HerbScan system includes:

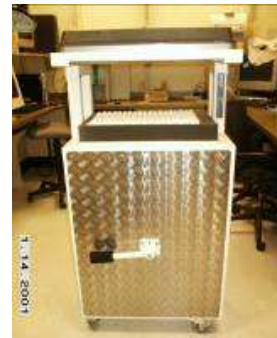
- HerbScan frame
- Scanner with cable
- Computer
- Display
- Barcode reader
- External hard drive



Each Partner will have a different scanner workspace, number of scanners, and team members based on the allocation of resources in your grant. The following sections contain suggestions to help you achieve an efficient workflow once you begin digitization.

### 1.4. Location of Scanning Station

Ideally the scanning facility will be located close to where the specimens are stored. A convenient location for the HerbScan is in a centralized area in the herbarium. If the scanning station cannot be set up in the herbarium you will have to consider transporting the specimens to the station, and storage for them while they are being processed.



Some Partners have found it beneficial to locate the scanning station in a separate room to aid concentration and reduce the likelihood of distractions. You should choose a room or workspace which is well ventilated—the computer and scanning equipment generate heat.

**TIP:** If possible only use the computers connected to the scanner for scanning—this reduces the risk of a virus attack through e-mails/internet, and therefore reduces the risk of lost scanning time due to equipment failure.

Do not store scanned images on the computer. This can slow scanning time considerably. Use a network server or external hard drive to store images. Some Partners have discovered that antivirus software interferes with the scanning process by increasing scan time significantly.

### 1.5. Room Lighting

When choosing a position for the scanner, make sure that it is positioned away from direct light, which could interfere with the quality of the scans. Make sure there is uniform light surrounding the scanner, and that the room is dimly lit. Partners may also find it beneficial to use some form of blackout blind if locating the scanners in a room with windows to avoid sunlight directly shining on the equipment—this will also help with reviewing images on the computer monitor.

**TIP:** Equipment should be kept clean and dust free at all times. The foam bed of the scanner, in particular, should be cleaned of any bits of plant material that may appear over time. The recommended method is to use a compressed air spray rather than cloths or dusters. This will ensure no dust lines appear on the images (vertical thin colored lines due to light reflecting off dust particles).






### 1.6. Unpacking the HerbScan Frame

Partners need to work closely with the Coordinators to receive their equipment in a timely and efficient manner.

The metal HerbScan frame is manufactured by HerbScan Engineering in London, UK and is distributed by different institutions depending on your location:

For Europe: Contact [lapi@kew.org](mailto:lapi@kew.org) at Royal Botanic Gardens, Kew

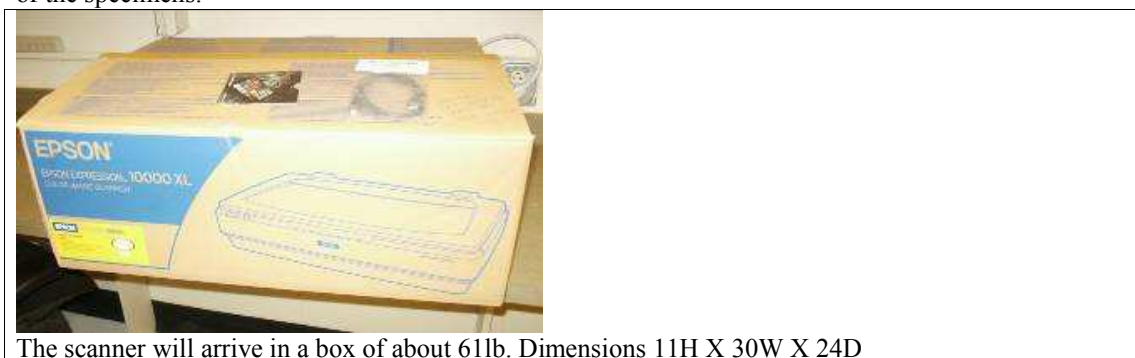
For Latin America: Contact Gloria Jovane ([JovaneG@si.edu](mailto:JovaneG@si.edu)) and Nelly Florez ([FlorezNA@si.edu](mailto:FlorezNA@si.edu)) at Smithsonian Tropical Research Institute

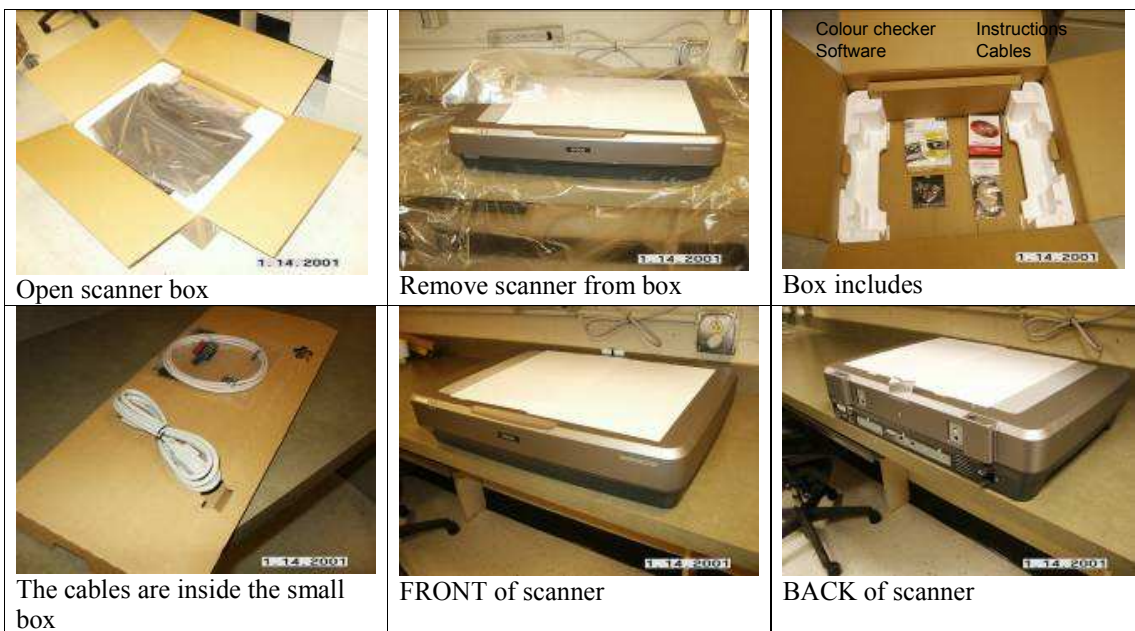
	<p>The HerbScan frame is in one piece and does not require any assembly.</p>	
<p>The HerbScan will arrive in a box of about 265lb. Dimensions of 60H X 29 W X 23 D</p>	<p>Dimensions of the HerbScan frame once uncrated 55 H X 26 ½ W X 21 D</p>	
		
<p>1. Remove top cover</p>	<p>2. Remove side cover</p>	<p>3. Remove HerbScan frame</p>

### 1.7. Unpacking the Scanner

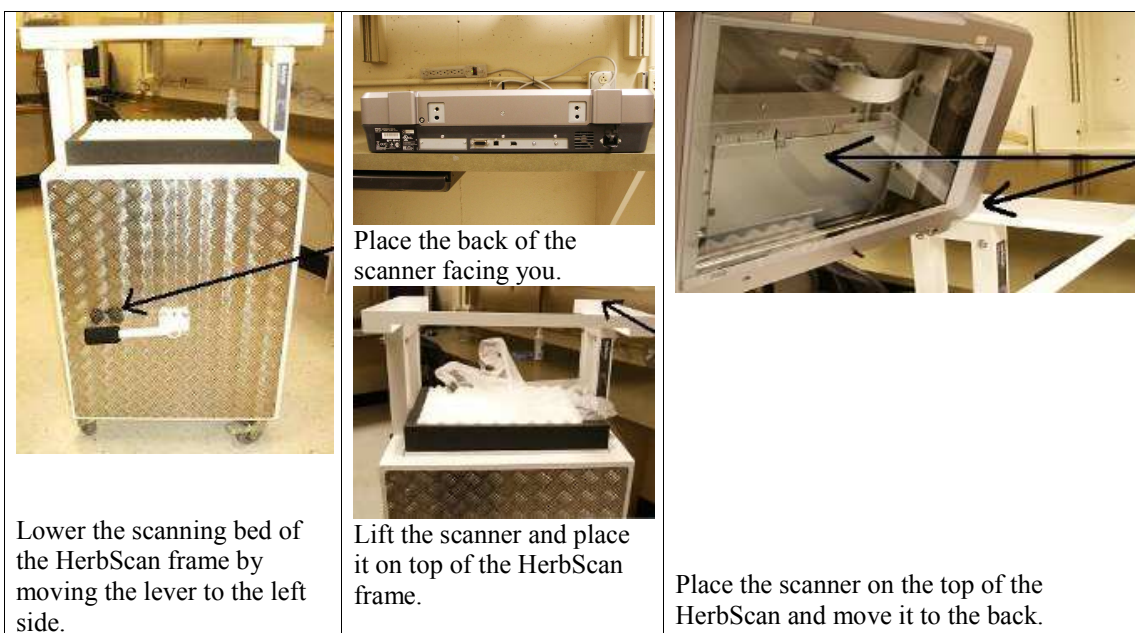
The standard high resolution flat-bed scanner used with the HerbScan is the Epson Expression Model 10000XL, Graphic Arts, USB2 and Firewire interfaces. Either a USB or Firewire cable is required depending upon which interface is utilized. The USB2 interface is standard for GPI. The scanner must be modified for inverted use by someone qualified to perform the modification. Modified scanners are provided with the HerbScan frames by the same distribution system.

Note: Do not assemble the top cover of the scanner. The top cover will not be used for the scanning of the specimens.





**1.8. Assembling the HerbScan Frame and Scanner**





Lower the scanner



FRONT view



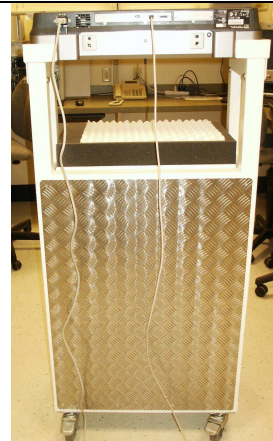
BACK view



*White arrow:* Plug the scanner's power cord into the AC connector (Note: You must slide the transportation lock up to the UNLOCK position).  
*Black arrow:* plug the USB cable into the scanner's USB port.



FRONT view



BACK view

## 1.9. Testing

All equipment must be tested to ensure that it is working properly and was not damaged during shipment. If the HerbScan frame and/or scanner are defective, please contact the GPI Coordinator from which you received them immediately.

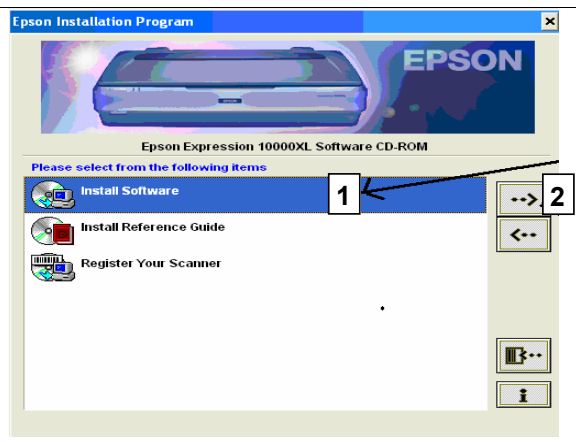

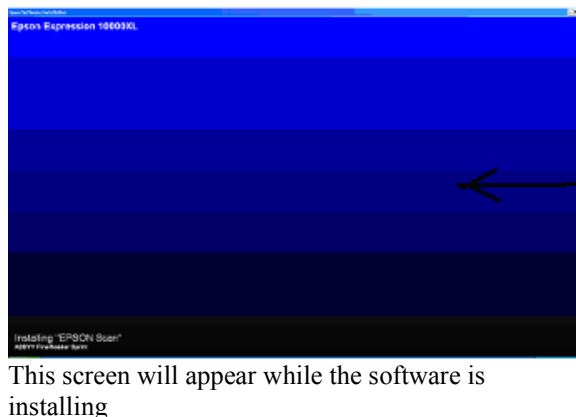

If the equipment or software you purchased is defective contact the source promptly to make corrections. Please note that scanning time should be between 3-5 minutes. If scanning a specimen takes longer than this there is a problem with your set-up that should be corrected before proceeding.

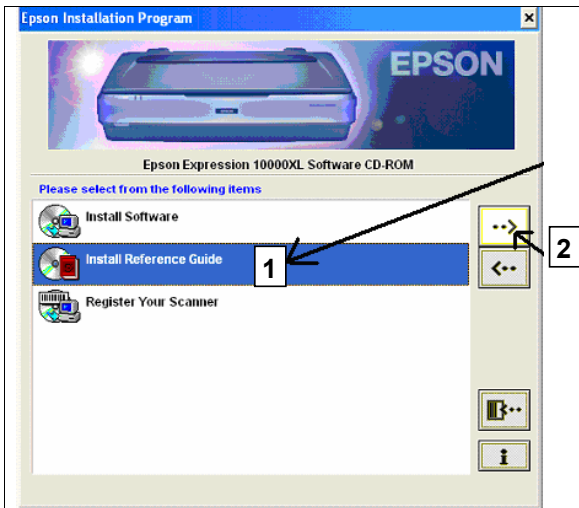
## 1.10. Installing EpsonScan

The software for the scanner has to be installed before the scanner is connected to the computer. EpsonScan software is provided in the box with the scanner (CD-ROM). This software will work

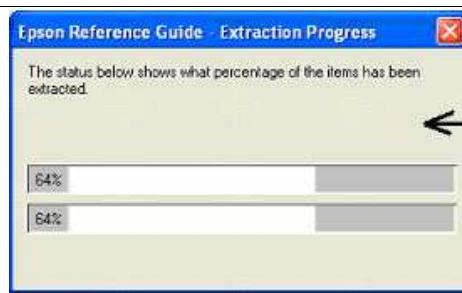


with Windows 2000 or later. Insert the scanner's software (white CR-ROM) in your CD-ROM or DVD drive.

 <p>1. Select Install software 2. Click install button</p>	 <p>Click install</p> <p>Read license agreement and click "Accept" and the software will install automatically</p>
 <p>This screen will appear while the software is installing</p>	 <p>When installation is complete, click "OK."</p>



1. Select "Install Reference Guide"
2. Click the arrow button

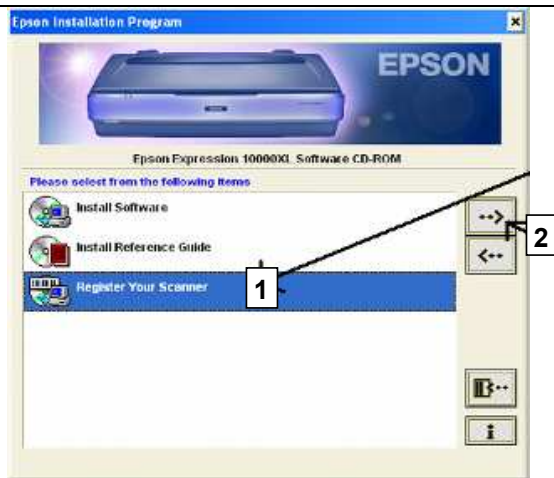


Status progress appears



Expression 10000XL Reference Guide.lnk

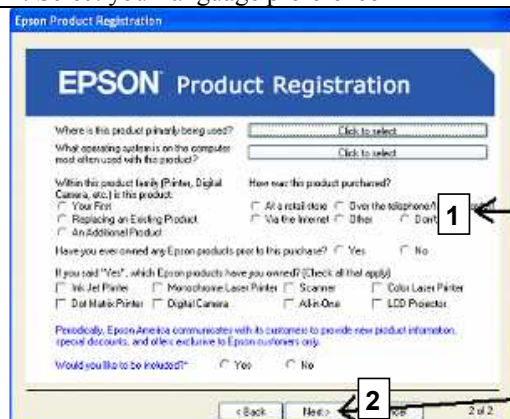
Once the installation is complete, an icon for the guide is placed on your desktop.





1. Select "Register Your Scanner"
2. Click the arrow button



1. Select your country
2. Select your language preference



Enter the information required

Enter the information required	Select Next
 <p>EPSON Product Registration</p> <p>You are now ready to register your product. Please select the method you prefer to submit your registration by selecting one button below.</p> <p>Select registration method:</p> <ul style="list-style-type: none"> <li><input checked="" type="radio"/> Send over the internet</li> <li><input type="radio"/> Modes using 800 number</li> <li><input type="radio"/> Print form for faxing</li> <li><input type="radio"/> Print form for mailing</li> </ul> <p>&lt; Back Register Cancel</p>	 <p>EPSON Product Registration</p> <p><b>Thank you for registering.</b></p> <p>To answer all of your product questions or to download updated drivers and view online manuals, please visit Epson Technical Support at:</p> <p><a href="http://support.epson.com/">http://support.epson.com/</a></p> <p>For information about Epson supplies and accessories developed specifically for your product, please visit the Epson Store at:</p> <p><a href="http://www.epsonstore.com">http://www.epsonstore.com</a></p> <p>Done</p>
Select register	Select Done, and close the Epson Installation Program screen and remove the CD-ROM.

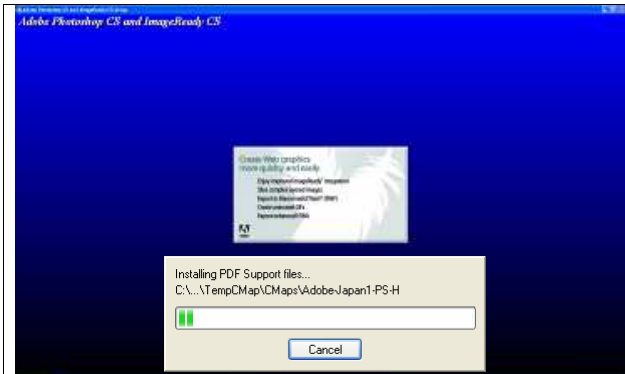
### 1.11. Connecting the Scanner

If the scanner's software was installed successfully then you are ready to connect the scanner to your computer.

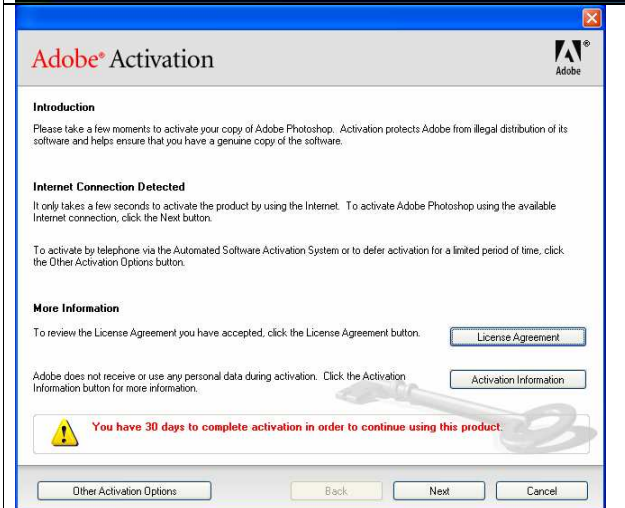


### 1.12. Installing Adobe Photoshop

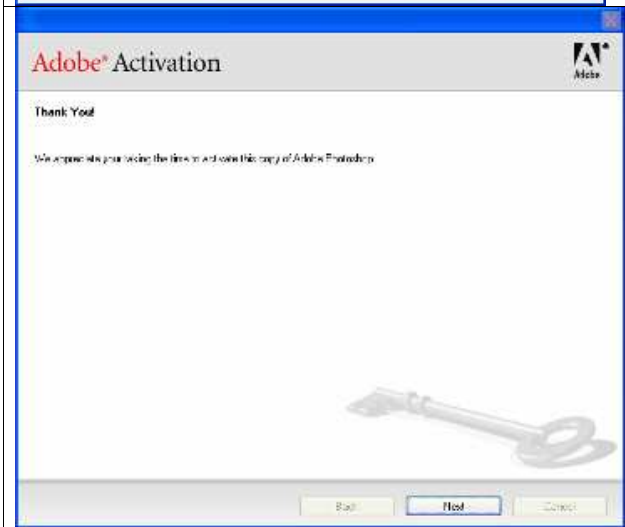
Insert your Adobe Photoshop CS CD-ROM in your CD-ROM drive.



Screen preparing for the Software installation.



To activate the license click next.



Activation is complete.

**TIP:** Another useful tip is to compartmentalize your PC by partitioning the hard drive. Then, you can assign the scratch disks in Photoshop to a drive partition other than that used by the program files. This will speed up processes in Photoshop. When an image is opened or scanned Photoshop retains a

“memory” of it on the scratch disks so after several scans this will affect the speed of Photoshop processes – if the scratch disks are re-assigned as indicated above, no effect on speed will be noticeable. Re-assigning the scratch disks can be done in Photoshop by going to Edit/Preferences/Plug-ins & scratch disks – ALL should be re-assigned to the newly partitioned drive if available.

### 1.13. Workflow

The following questions are important to consider in preparing for the project. We assume that many of these questions have already been considered in the process of preparing the proposal to Mellon Foundation for your participation in the project.

- How much time and labor will be required to gather your type specimens?
- How is your herbarium organized? Will you scan your specimens in the same order in which they are organized?
- Are some/all/none of your type specimens data based?
- What is the most efficient order in which to process your types?
- Where is the most efficient place to set up the HerbScan station?
- How will you handle specimens on loan? Tell staff at your institution about the project — ideally any outgoing loan material can be data based and imaged before it leaves the institution. Similarly, incoming loan material can be brought to the team’s attention and also be captured.

**NOTE:** Please keep in mind that the project is intended to deal only with specimens that have already been identified as types. Curation of specimens is expected to be the responsibility of your institution.

It should take between 3-5 minutes for a specimen to be scanned. Longer scan times indicate a problem with the setup that should be corrected before proceeding with the project.

**TIP:** When not using the scanner, please leave the scanner in the raised position to reduce the accumulation of dust. Also please turn the scanners off if being left unused for periods of more than 30 minutes to reduce the heat stress on the equipment.

## 2. Specimen Databasing

### 2.1. Introduction

The two main tasks of the project are the databasing and scanning of herbarium specimens. Each specimen must be barcoded and the label information for each specimen entered into a spreadsheet or database system. Each institution will have its own method for databasing specimens but these data will have to be exported into a common XML (Extensible Markup Language) format designed for the project for delivery to JSTOR. This format is defined by an XML schema.

This section will cover general guidelines and recommendations for capturing specimen label data, but as each partner will use their own method for capturing data, this section is inherently generic in scope. However, the quality of the data produced in the project can be improved by standardizing data entry as much as possible, correcting spelling mistakes that may exist on labels, and using online or print authority files to verify information before it is sent to JSTOR.

The Export section will specifically address how to work with data outside of your institutional database and format it to fit the XML schema.

### **Type specimens**

Specimens included in the project fall under a wide umbrella of what are or could possibly be type specimens. Many institutions have traditionally put type or historical material in “red folders” to designate their importance in the collection. Since most institutions have not verified all of their type specimens and funding is not available to research the status of every specimen, anything marked as a type or possible type can be included in this project.

### **Specimens vs. Collections**

Most herbarium specimens consist of a dried plant specimen glued to a single sheet of paper with a label describing the collection event. A collection event is typically defined by a collector in the field and given a unique collection number as part of a series for that collector. A single collection event can produce a number of duplicates that are sent to other institutions. It is also possible that a single specimen of that event can take up one or multiple herbarium sheets. Dealing with these various cases is described below.

## **2.2. Barcoding**

### **Barcoding Specimens**

Adding a barcode to each specimen is mandatory for the GPI project and each image file must be named with the corresponding barcode code number. However, this is not necessarily the same idea as adding accession numbers to specimens as traditionally done in herbaria. For most institutions, barcode numbers and accession numbers have nothing in common.

*For a single sheet with one specimen*

In most cases, a single herbarium sheet will contain one specimen and receive one barcode.

*For one sheet with more than one specimen*

In cases where there is more than one specimen on a sheet, we encourage partners to barcode each specimen on the sheet. This also makes naming image files easier when specimens are scanned.



**BM0000001**



**BM0000002**



**BM0000003**

*For two or more sheets containing one collection*

For a collection that makes up more than one sheet, each sheet should be given its own unique barcode but the sheets should be indicated as associated or related in the database record. This provides each sheet with a unique identifier and is useful when the parts are separated from each other. For example, you might have a palm specimen which due to size is mounted on 3 sheets—each one has a unique barcode but they are the same collection.

The images below represent one collection which was split onto three sheets with each sheet having a unique barcode.



**BM0000123**



**BM0000124**



**BM0000125**

Note: in this example, some institutions will use the same barcode number for all three sheets. How to name the associated image files will be covered in the digitization section.

**2.3. Databasing Software**

A computer database is needed to store the specimen data for each digitized image. The choice of software for databasing specimens depends on the partner institution, but it must be capable of exporting the required data into the GPI standard format. The two most widely used programs by herbaria are:

BRAHMS (<http://herbaria.plants.ox.ac.uk/BOL/home/default.aspx>)  
SPECIFY ([www.specifysoftware.org](http://www.specifysoftware.org))

BRAHMS is used extensively throughout Latin America and an export has been developed to automatically export data from BRAHMS to the proper XML schema for the GPI project.

For institutions with only a small numbers of specimens, Microsoft Excel or compatible spreadsheet software can also be used to record the specimen data. A spreadsheet following the XML schema can be obtained from Rafael Barron or the GPI XML Generator can be used for data entry. (See Export section for more information.)

## **2.4. Databasing Methods**

There are two methods generally used by partners to digitize and database their herbarium specimens. The first is to database the specimens and then scan them, the second is to scan the specimens and then database from the image of the specimen. The method that works for each partner will vary by institution.

### *Database First*

It often easier for decisions about the specimen (such as multiple specimens on a single sheet or multiple sheet specimens) to be made before scanning the specimens. This is generally the method used when specimens are not clearly labeled and more experienced curatorial staff can be used to make decisions on what specimens should be databased and scanned for the project. This prevents less experienced scanning technicians from making difficult decisions about the specimens and they can focus instead on scanning specimens more quickly.

### *Scan First*

Another method is to scan the specimens first and then database the specimen from the scanned image later at a workstation dedicated to that task. This allows the HerbScan unit(s) to be placed where it is most convenient for the institution, most often in the herbarium itself, which reduces the movement of the specimens in and out of the herbarium. The data workstation ideally consists of a computer with two screens. One screen is a large format flat panel mounted vertically to display the specimen while the other screen displays the data entry form. This is ideal for situations where specimens have already been clearly verified as types and decisions about what is on the sheet don't have to be made by the scanning technician.

### **Data Entry**

Data entry will vary by institution based on the database software in use. The software should contain fields that conform to items in the XML schema covered below. Please note that most database systems will NOT have fields identical to the XML schema, but almost all will have fields that overlap the schema in some way. During the export, it will be necessary to export the data and convert it to the fields that make up the schema. This is covered in the Export section of the document.

## **2.5. Database Fields**



## **Fields from the XML Schema**

The following are fields that make up the GPI XML Schema. Recommendations are given here to encourage standardized data entry for content among the partners. The Export section will cover standards for the format of the data in XML.

### **UnitID**

UnitID refers to the barcode number for the specimen. This field should be made up of the institution's Index Herbariorum acronym and barcode number. This number must be unique for every specimen. The image file associated with the specimen must have the same name as the UnitID.

### **DateLastModified**

This date refers to the last time any part of the label information for this Unit or specimen image was changed in the database.

### **Identification**

Identification is a series of fields that cover the identification or determination of the specimen. Since each specimen could have multiple determinations, this field can be repeated in the XML export. In these cases, each identification must be identified as being the "filed as" name or not. (This is covered more extensively in the Export section.) The fields include:

#### **Family**

The family name for the taxonomic name based on the scanning institution's own taxonomic decisions or as shown on the sheet.

#### **Genus**

As recorded by the scanning institution. First letter should be uppercase.

#### **Species**

As recorded by the scanning institution. Should be all lowercase.

It is recommended that scientific names be verified for spelling in a taxonomic authority file. Some recommendations include:

International Plant Names Index (IPNI) <http://www.ipni.org/index.html>

Tropicos (<http://www.tropicos.org/>)

World Checklist of selected plant families <http://apps.kew.org/wcsp/home.do>

#### **Author**

The author of the species name including basionym author and ex/in following standard format. Standard author abbreviations should be used based on the Authors of Plant Names maintained by the Royal Botanic Gardens, Kew at <http://www.ipni.org/ipni/authorsearchpage.do> If the species author is missing or unknown, use "Not on sheet".

#### **Identifier**

The identifier refers to the name of the person who made the determination of this Identification. Often called "Determiner" in other databases. Use "Not on sheet" if the identifying/determining person is not known.

#### **IdentificationDate**

The date recorded by the scanning institution for when the determination of this Identification was made. In the XML Schema, the date must be separated into distinct Day, Month, and Year fields, however most institutional databases store this information in one field.

### **TypeStatus**

The formally recognized type status of specimens placed in ‘red folders’ in herbaria or scanned for GPI varies, and it not always easy to determine. It is recommended that herbaria only use the currently accepted set of type categories which form part of our formal nomenclatural system as represented in the International Code of Botanical Nomenclature (ICBN) (see Appendix x for Article 9 of the Vienna Code) with the addition of certain extra terms as defined below.

<b>GPI Standard Term</b>	
Holotype	See ICBN (Article 9.1.)
Epitype	See ICBN (Article 9.7.)
Isoepitype	Duplicate of an epitype
Lectotype	See ICBN (Article 9.2.)
Isolectotype	Duplicate of a lectotype
Neotype	See ICBN (Article 9.6.)
Isonotype	Duplicate of a neotype
Paratype	See ICBN (Article 9.5.)
Isoparatype	Duplicate of a paratype that is not cited in the protologue
Syntype	See ICBN (Article 9.4.)
Isosyntype	Duplicate of a syntype that is not cited in the protologue
Isotype	Duplicate of a holotype. See ICBN (Article 9.3.)
Type	Type category uncertain
Type ?	Possible, putative, probable, uncertain type material
Original material	See ICBN (Article 9.2, Note 2.)
—	Determined not to be a type

In cases where a specimen is in a type folder but its status as a type is unknown, then the type status should be recorded with a dash [-], and a comment should be added to the notes field stating ‘in type folder’ (or similar wording).

In cases where a supposed type has been determined to be *definitely not a type*, then the type status should be recorded with a dash [-], and a comment should be added to the notes field providing the relevant information. This would include cases where an annotation on the specimen or the type folder refers to a taxon that has not been validly published, or when an incorrect interpretation of the typification of the taxon has resulted in an incorrect annotation of a sheet or an incorrect citation of the type of a taxon in the literature.

If the policy of the herbarium is to enter type status exactly as appears on the type sheet or red folder, when a non-standard term is encountered this should also be converted into a corresponding standard term as indicated before the data is sent to JSTOR.

Examples:

<b>Non-standard term</b>	<b>GPI Standard term</b>
Clonotype	-
Cotype	Type ?
Co-type	Type ?
HOLO	Holotype
Holotype fragment	Holotype when held in same herbarium as holotype; isotype when held in another herbarium
Holotypus	Holotype
ISO	Isotype
ISOLECTO	Isolectotype
Isolectotypus	Isolectotype
ISONEO	Isonotype
Isonotype	Isonotype
ISOPARA	Isoparatype
ISOSYN	Isosyntype
Isotype?	Type ?
Isotypus	Isotype
LECTO	Lectotype
Lectotypus	Lectotype
NEO	Neotype
No type?	Type ?
PARA	Paratype
Paratypus	Paratype
SYN	Syntype
Syntyp	Syntype
Syntypus	Syntype
Type fragment	Type
Type material	Type
Type material ?	Type ?

#### **GenusQualifier**

Qualifier expressing doubt about the genus epithet (eg. cf)

#### **SpeciesQualifier**

Qualifier expressing doubt about the species epithet (eg. cf)

#### **Infra-specificRank**

Rank based on ICBN and as recorded by the scanning institution. Should be all lowercase.

#### **Infra-specificEpithet**

As recorded by the scanning institution. Should be all lowercase.

#### **Infra-specificAuthor**

Follow the same guidelines as for the Author field.

#### **PlantNameCode**

An optional code that is meaningful to the scanning institution for the name given. Often a tracking number. May be used to provide feedback from JSTOR to the scanning institution.

### Collectors

This field is just a text string listing the collector or collection team for this Unit. The senior or primary collector should be listed first. Many databases store the primary collector and collection team in separate fields and these will need to be merged before creating the XML file.

Information about botanists, proper spelling of their names, dates, information about their collections, types, and publications, can be found at:

Gray Herbarium (GH), Harvard University's Index of Botanists:  
[http://asaweb.huh.harvard.edu:8080/databases/botanist\\_index.html](http://asaweb.huh.harvard.edu:8080/databases/botanist_index.html)

Stafleu, F. A. & Cowan, R. S. 1976-1988. *Taxonomic Literature*, 2<sup>o</sup> ed. (TL2) Vols. 1-7. *Regnum Veg.* 94: 1-1136; 98: 1-991; 105: 1-986; 110: 1-1214; 112: 1-1066; 115: 1-926; 116: 1-653. Bohn, Scheltena & Holkena, Utrecht.

Stafleu, F. A. & Mennega, E. A. 1992-2000. *Taxonomic Literature Supplementum*. Vols. 1-6. *Regnum Veg.* 125: 1-453; 130: 1-464; 132: 1-550; 134: 1-614; 135: 1-432; 137: 1-518. Koeltz Sci. Books, Königstein.

Stafleu, F. A. & Mennega, E. A. 1992-2000. *Taxonomic Literature Supplementum*. Vols. 1-6. *Regnum Veg.* 125: 1-453; 130: 1-464; 132: 1-550; 134: 1-614; 135: 1-432; 137: 1-518. Koeltz Sci. Books, Königstein.

Dorr, L. J. & Nicholson, D. H. 2008. *Taxonomic Literature Supplementum*. Vol. 7. *Regnum Veg.* 149: 1-469. Gantner Verlag.

### CollectorNumber

This is generally the number assigned by the senior (or primary) collector to the specimen. This field can contain any characters found on the sheet, including prefixes or suffixes. Where there is no collector number for the Unit, the value "s.n." should be used in the field.

Examples:

1234  
LW-1234  
LW-1234a

*Note:* In some cases, a person other than the Collector gives a Collection Number to the specimen, or gives an additional one (different from that given by the collector). In these cases, the scanning institution decides which Collector and Collection Number to enter. If the scanning institution verifies the type status of the specimen against the protologue, using the same Collector and Collection Number indicated there is recommended.

### CollectionDate

This date recorded by the scanning institution for when the collection was made. Occasionally the collection date is made up of a date range (May – June 1904) or other text (Spring 1865). These can be accommodated in the XML but must be split into separate Start Date and End Date fields, each comprising separate Day, Month, Year fields. An Other Text field is available for any other text used to describe the date.

**ISO2Letter**

ISO refers to the International Organization for Standardization. The GPI project uses their standard for country name, represented by a 2-letter ISO 3166-1 code. For example, the 2-letter code for Argentina is AR. This code in the schema refers to the country where the specimen was collected. The ISO 3166 master list is available at

<http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html>

Many databases do not contain a field for ISO2 Letter. This field can be populated from the Country Name field after the data is exported from the institution's database. The XML Generator will also automatically populate this field if the country name field is populated.

**CountryName**

This refers to the country in which the specimen was collected. Often older specimens, especially type specimens, have country names that are no longer officially recognized. In these cases, it is recommended to use the current country name for the region and to put the country name as written on the label in the notes field. However, it is acceptable to use this field for whatever information you have.

*Note:* Many times in old collections geographical Latin names are used to indicate the country or other part of the locality. To assign the proper/actual country of origin of the specimen the following work can be consulted:

Stearn, W. 2004. *Botanical Latin* 4<sup>o</sup> ed. 1-560. Timber Press, Portland. (*search at the Sections "Latinization of Place-Names" and "Geographical Names"*).

Example:

Brasilia meridionalis (refers to Southern Brazil)

**Locality**

This is the literal string of text that was recorded for "locality" describing where the specimen was collected. No other itemization of location is available in the schema. State, County, Municipio, latitude and longitude, etc. must all be concatenated and added to the Locality field.

Example:

Provincia de Misiones: Dpto. San Javier, Ruta Provincial 2, bordeando el Río Uruguay 27°27'10''S 54°38'04''W

**Altitude**

This is the altitude (or altitudinal range) where the specimen was collected, indicated in meters or feet. Please enter 'meters' or 'feet' and not just the first character 'm' or 'f'.

Examples:

100 meters  
25.5 feet  
3500-4000 meters  
1500-2000 feet

**RelatedUnitID**

This tag can be filled with the UnitID/s of another specimen/s (this is a multiple occurrence tag, so multiple UnitIDs for related Units can be included). There is no attribute to describe or classify the nature of the relation to the other specimen/s, just the existence of a relation.

It can also be used to relate or link the duplicates or additional sheets for the treated UnitID, or to relate it to other syntypes, paratypes, etc.

#### **Notes**

This field can be any text and has no other constraint. It can also be used to enter any additional information for which fields are missing in the XML schema.

Examples:

Shrub 1.2 m. Flowers deep blue.

Specimen cited in Ann. Missouri Bot. Gard. 89: 101.

Close to a narrow water stream. Flowers arranged in clusters. Commonly know as “calabaza del monte”

### **3. Imaging of Specimens**

The purpose of digitizing specimens is to create high-resolution archival master images for long-term preservation in the form of a TIFF file. These master files will be deposited at JSTOR and will be displayed on the web. The archival images will be stored to tape twice at JSTOR and will be kept in two separate locations in the US for preservation.

A digital master file, created through direct scanning or imaging of an object, should accurately represent the visual information in the original object. This type of accuracy is best achieved by adjustments to equipment settings before digitization. An archival quality scan should not require intensive enhancement or processing after scanning.

#### **3.1. Standard Specimen Sheet Layout**



This section reviews the requirements for the digitization of standard specimen sheets.

For a specimen to be accepted for GPI it is necessary to scan the specimen with three items on the sheet:

### 1. Specimen barcode

A barcode, starting with the unique herbarium code from Index Herbariorum (<http://sweetgum.nybg.org/ih/>)

### 2. Institutional logo and scale

A measurement scale (with your institution's logo).

### 3. Color Target

The required color target is the Gretag Macbeth Mini Color Checker. The checker contains an array of 24 scientifically prepared colored squares that provides the standard for comparing, measuring and analyzing differences in color reproduction. ([http://www.xrite.com/custom\\_page.aspx?PageID=73](http://www.xrite.com/custom_page.aspx?PageID=73))



Care should be taken in positioning of both color target and scale bar. The required color target is the Gretag Macbeth Mini ColorChecker. The color checker, as seen in the examples below, may be placed anywhere on the specimen sheet.



**TIP:** The size of the checker is 3.25 x 2.25 in (8.25 x 5.7 cm)—in some cases the entire color target might not fit on the specimen sheet. If so, please use common sense and place the checker where it will fit. In some instances this may cover some small part of the specimen. This is relatively rare and when necessary it is usually possible to avoid covering any critical diagnostic part of the specimen.

Do not cut the target into strips—it will make it difficult for users to perform automatic image calibration.

**NOTE:** No color calibration or filtering of the image is performed during the image creation process.

### 3.2. Image Format/Output

Each type at your institution must be digitized according to the following specifications:

Resolution:	600 pixels per inch	Color Space:	Adobe RGB (1998) <sup>1</sup>
Color Depth:	24-bit		
File Format:	Uncompressed TIFF files (Tagged Image File Format)		
Layout:	Portrait (not landscape)		

The resulting file size will be approximately 200MB for sheets 310mmx437mm or 11x17inches.

One high-resolution image for each type specimen is required. If you wish, you may submit more than one image per type specimen. Situations in which you may wish to create more than one image per type specimen include:

- when there are details of a specimen you would like to scan at a higher resolution
- when there are contents in a packet/envelop that you do not want to scan with the main image or that cannot fit with the main image
- when a specimen sheet contains additional plant material, annotations and/or illustrations

<sup>1</sup> RGB and sRGB (which sometimes is set as the default color space) are not the same.



### 3.3. File Naming Conventions

The file naming conventions are a crucial aspect of the production process and output. Please follow these conventions closely. If you have any questions please contact JSTOR directly. The file naming convention for a specimen image must ALWAYS be the **institution's Index Herbarium code concatenated with the barcode number**. There are some variations to accommodate additional images and non-standard specimens.

#### 3.3.1. One Specimen on a Single Sheet

The file naming convention for the specimen image must be the institution's Index Herbarium code concatenated with the barcode number.

Example: **P123456789.tif**

Herbarium Code	=	<b>P (Paris)</b>
Specimen Bar Code	=	<b>123456789</b>
File Extension	=	<b>.tif</b>

The same barcode number must also be included in the metadata submitted with the image file.

**NOTE:** The following naming conventions are NOT acceptable:

<b>123456789.tif</b>	<b>p123456789.a.tif</b>
<b>P 123 456 789.tif</b>	<b>P123456789</b>

#### 3.3.2. Detailed Capture or Additional Scans of a Single Sheet

To capture details of a particular specimen or to scan additional items from or with a sheet (such as an illustration, capsule contents, etc.) an additional scan can be captured at the same resolution or higher (1200ppi). This additional scan must include the scale bar and color checker. The file naming convention must include the same barcode as the original image followed by an underscore and a lowercase letter (starting with "a"):

Example: **P123456789\_a.tif**

Herbarium Code	=	<b>P</b>
Specimen Bar Code	=	<b>123456789</b>
Specimen Detail	=	<b>_a, _b, _c, etc.</b>
File Extension	=	<b>.tif</b>



P123456789.tif



P123456789\_a.tif

**NOTE:** The following naming conventions for additional/detailed scans are NOT acceptable:

P123456789 a.tif

P123456789.a.tif

P123456789.1.tif

P123456789\_1.tif

P123456789 1.tif

### 3.3.3. Multiple Specimens on a Single Sheet

In cases where there is more than one specimen on a sheet, we encourage partners to barcode each specimen on the sheet. The Digital Technician can scan the sheet once, and then save the master image several times, naming each image with the relevant barcode.



BM0000001.tif



BM0000002.tif



BM0000003.tif

### 3.3.4. One Specimen on Multiple Sheets

Each sheet should be given its own unique barcode but the sheets should be indicated as associated or related in the database record for each. For example, you might have a palm sheet which due to size is mounted on 2 sheets—each sheet has a unique barcode but they are the same specimen. For more information on the data aspects please see the –Export section in this document.

The images below represent one type specimen which was saved on three sheets. Each sheet has a unique barcode.



BM0000123.tif



BM0000124.tif



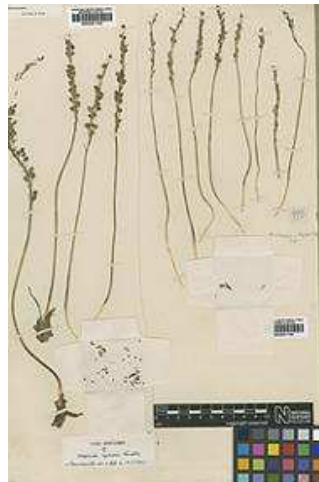
BM0000125.tif

### 3.4. Non-standard Specimens

#### 3.4.1. Packet/Capsule/Envelope

Some partners image the contents of a packet/capsule/envelope separately from the main image/sheet. This is not a required practice and is left to the discretion of the partner.

You may choose to paste the image of the capsule on top of the main image unless the capsule itself carries important annotations, in which case to the side of the capsule would be appropriate.



#### 3.4.2. Reverse of Specimen Sheet

Some specimen sheets may carry important annotations on the reverse of the sheet. These may be included by inverting the specimen, imaging the annotation, and then following the procedures for multiple sheet types or by using a “normal” scanner (one placed the right way up) in your institution. Whether or not to include them is at the discretion of the partner.

### 3.4.3. Additional Materials

Some type specimen sheets carry additional plant material, annotations, illustrations and/or notes.

We recommend scanning these items separately when they obscure the specimen or labels. The additional material can be “folded” out of the way for one scan and then imaged separately and included as associated images using the file naming conventions.

For GPI this situation will be handled by scanning each sheet/item and naming the image files with the same barcode number appended with “\_a”, “\_b”, “\_c”, etc. The use of different barcodes for each sheet/item is also acceptable.

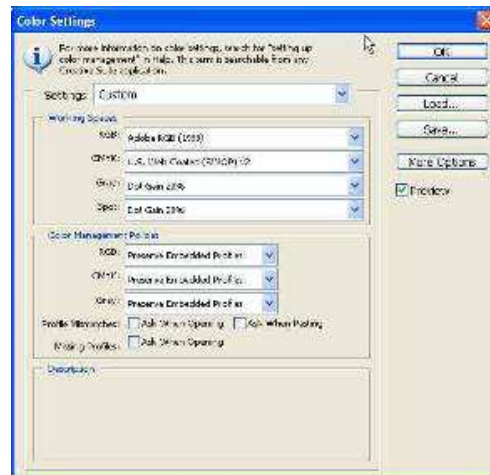


### 3.5. Step-by-Step Scanning Instructions

**NOTE:** Please use Windows Explorer to handle all files. In Windows Explorer you can change the views on the list of files. Go to the View menu and toggle between Thumbnail view or Details view. This is very useful when checking for file names, file sizes, etc.

1. **The software that controls the scanner, Epson Scan, has to be used through Photoshop.** Open Photoshop. The first step is to configure the software (this should be done before scanning and checked at the start of every session to ensure nothing has changed).

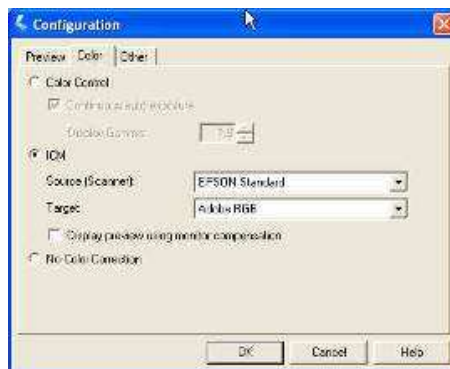
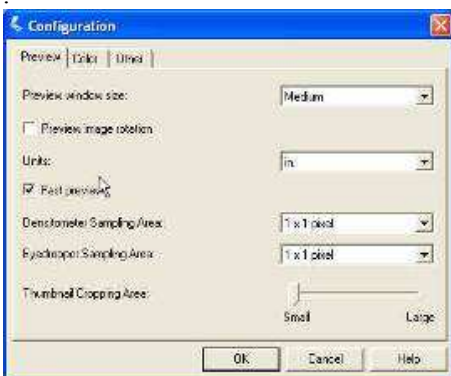
- **Go to the Edit menu** and select Color Settings
- In the Settings drop-down menu choose Custom.
- Next, select Adobe RGB (1998) in the RGB drop-down menu. Hit OK.



2. **Turn on the scanner.** Go to the File menu and select Import > Epson Expression 10000XL. This will open the Epson Scan software. If the software is not opening, the scanner is probably still warming up. Be sure that the software is set at Professional Mode. It should look like the image below. If not the Mode can be switched at the top of the window.



3. **Click on the Configuration button** (located at the bottom of the window) and select the Preview tab. Make sure that the Preview Image Rotation box is not ticked. In the same Configuration window, go to the Color tab. Select ICM, select EPSON Standard as the Source (Scanner), and select Adobe RGB as the Target. Click OK when done.



4. Be sure that all options on the Epson Scan window are the same as on the image on step #4.
  - Document Type: Reflective

- Document Source: Document Table
- Auto Exposure Type: Photo
- Image Type: 24-bit Color
- Scanning Quality: Best
- Resolution: 600 dpi
- Scale: 100%

Make sure that all boxes in the Adjustments section are unticked.

5. ***Place the barcoded specimen in the HerbScan.*** Be sure to align the specimen sheet on the foam. Place the institutional logo scale and the GretagMacbeth color card taking care to not cover any writing or part of the specimen. If there is no space for either, place blank herbarium sheets under the specimen and pull out, to expand the “canvas”.

**NOTE:** Important notes on specimen handling and image capturing:

- When handling a specimen, be sure to grab the sheet by the underside with both hands. This will avoid any bending of the sheet.
- Beware of any loose parts, be sure to put it inside the specimen packet.
- Be sure that there is enough fertile material visible on the sheet. If no fertile material is visible, remove any pieces from inside the specimen packet.
- If fruits or flowers inside packet are too brittle or in several pieces, **see step 10.**
- If necessary, take two or more images to show every label or important fertile specimen part. Beware of overlapping labels!!! (See file naming section of this manual **and step 10).**

6. **Move the HerbScan lever** to lower the scanner onto the specimen. Click on the Preview button. Make sure that the entire specimen is represented and that the scale and color card are visible.



7. **Select the area of the preview to scan.** Move the cursor so that it is over the image of the specimen. The cursor will turn into a cross, and while right-clicking and holding the button, drag it across the screen until the entire area is selected within the outline. Be sure that every part of the specimen shows in this view and there are no pieces outside the frame. After it is selected, click the “Auto Focus” button.
8. **Press Scan on the Epson Scan window.** This will take approximately 5 minutes. When it has completed scanning the image will appear in Photoshop behind or in front of the EpsonScan window. Be sure that you don’t run any other programs on the computer or bump or move the HerbScan system. This could potentially cause glitches and artifacts on the image.







10. **Checking Image for focus and problems.** Thoroughly check the first image and every 10 images or the last one during the session if less than 10 are scanned (this is for each Herbscan used at a time). Bring the image up to 100% view. This can be done various ways:

Use the loupe icon until the window title reads 100%, right click on the image and select Actual Pixels, or click on the Actual Pixels button on top of the window. Press CTRL+ALT+0.

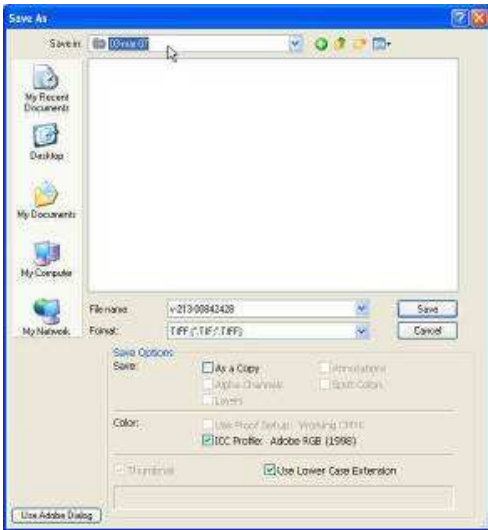
Look at the image thoroughly, every corner, by using the Photoshop Hand tool. Look for any artifacts.

**NOTE:** If any artifacts are found rescan the specimen. Repetition of artifacts might mean that the scanner is breaking down. Be sure to also check the previous images if some were done during that same session.



11. **Saving TIFF.** Once you're comfortable with the quality of the image save the file to the external hard drive. Be sure to follow the File Naming Standards.

- From the File menu select Save or Save As.. (or hit CTRL+S).
- Select TIFF as the format (the file extension in the filename will change to TIFF)
- Name the image following the File Naming Standards.
- Click on Save. Set image compression as none and select IBM PC as the byte order.
- Verify that the file was saved in the correct folder.



## 12. File Naming Standards

NY12345678.tif

NY – Institution's Index Herbariorum code

12345678 – 8-digit barcode number

.TIF File type

It is very important that all your images are saved as this file format.

If more than one image is captured per barcode number then the underscore and an "a", "b", or "c" is added for the next consecutive images.

**NY12345678.tif**

**NY12345678\_a.tif**

**NY12345678\_b.tif**

**NOTE:** If two or more specimens to be scanned are on the same sheet, scan the sheet once as usual and save a separate copy for each barcode number – e.g. if there are four barcoded specimens on the

sheet, you will save four copies of the same image. This is done by using the Save As... command in the Photoshop File menu.

Remove the specimen from the scanner and be sure to return any particles to the packet. As an institution, it would be helpful to mark specimens that have already been scanned.

Double check that the image or images were saved to the correct folder. You should also note that these images will generally result in a 200MB file size or larger. To do this be sure to have the Details view on Windows Explorer. You can do this for all files at the end of your session. If there is a discrepancy of file sizes, check your settings. This is also a good way to catch any mistakes while naming the images.

#### **4. Image Quality Control**

The purpose of this section is to provide quality control guidelines for the images of type specimens. While JSTOR does perform Quality Control on a random 10% sample for every batch received, we have found it to be beneficial for onsite quality control to be performed as well. The two stages of quality control occur during scanning and post scanning and are outlined below.

##### **4.1. During Scanning**

Check the first image and every 10<sup>th</sup> image in each scanning session or the last one during the session if less than 10 are scanned. Bring the image up to 100% view and move systematically through the image looking for scanning artifacts.

To zoom into the image:

- Use the loupe icon until the window reads 100%, right click on the image and Select Actual Pixels, or click on the Actual Pixels button on the top of the window.
- Press CTRL+ALT+0.

Below are descriptions and examples of problems that may arise during this first initial check.

Insert the artifact section here.

##### **4.2. Quality Control Post Scanning**

###### **4.2.1. Initial Rapid Checks**

Certain details can be checked rapidly in Windows Explorer. These are:

###### ***File Size***

An uncompressed full specimen image should be in the region of 180Mb (+/- 40Mb). The size of files varies according to the area captured in an image. If it is significantly out of this range this may indicate a problem.

Any files that are around twice this size are likely to be unflattened. These should be opened in Photoshop and flattened as soon as possible.

Smaller sized images may be found due to extraneous material (i.e.: letters, artwork, cryptogams).

### ***File Type***

Check that the file format of all image files is TIFF. If the format is not TIFF but is another uncompressed format – e.g. PSD or BMP – the specimen does not need to be rescanned – it can simply be resaved to the correct format. If the format is a compressed format (e.g. JPEG) then the specimen needs to be rescanned and saved as a TIFF.

### ***File Name***

In Windows Explorer it is possible to quickly check that all filenames are of the correct length and format (e.g. K123456789). If filenames have been hand-typed then occasionally digits are incorrectly saved. If filenames have been entered using a barcode reader then the herbarium code (e.g. K) may be saved as lower case if the Caps Lock button was on when the filename was entered. All suffix images should be in the format 'K123456789\_a'. Often the underscore is left out or replaced by a space. In these cases the files should be renamed according to the correct format. (insert where this is outlined-see section ####)

#### **4.2.2 Check for Duplicates**

It is useful to keep a list of all barcodes sent to JSTOR in order to make sure that the same barcode is not sent twice. Please check for duplicated images before sending new batches to JSTOR. If you find duplicate images remove them from the disk.

#### **4.2.3 Check for Components**

All images should be checked for presence of all components (scale, color checker, plant parts, and labels) and that all critical information on the specimen is clear and visible. These checks can either be done using derivative GIF images (as in the examples below) or JPEG images.

- Check that ColorCheckerTM color/greyscale, scale bar and barcode are fully visible.
- All relevant labels should be visible – if not then there should be a suffix image which shows the full image with the label visible.
- No relevant part of the specimen has been obscured.

If the necessary components are not present in the image or parts are being obscured, the image is rejected and the specimen will need to be rescanned.



### **4.3. Check Images for Scanning Artifacts Should be inserted above and re-numbered**

#### **4.3.1. Pixilation**

The artifacts may appear on the image as areas which are more pixilated or blocky than the surrounding area of the image. This is due to a hardware problem with the scanners which seems to develop over time.

Some examples of Pixilation:

- The area will look pixelated or “blocky” with jagged edges..
- In areas where there is not much of a color contrast, the image will look fuzzy compared to the surrounding areas.
- The color may appear to be broken up causing a rainbow-like effect along the edges.

	
<p>An example of pixilation. Notice that the specimen detail is blurred and the edges appear to be jagged.</p>	<p>An example showing pixilation on the edge of a specimen leaf.</p>

**ACTIONS:**

- *Image:* Reject the image and rescan the specimen.
- *Hardware:* If pixilation shows up after rescanning, there is a potential problem with the scanner. Check all previously scanned images to see if it is an isolated event.
- *Reporting:* Report the problem with the scanner to IT support and your project manager, noting the asset number of the scanner. If the problem continues, report it to the institution that provided you with the scanner or any of the LAPI Coordinators.

**4.3.2. Vertical Lines**

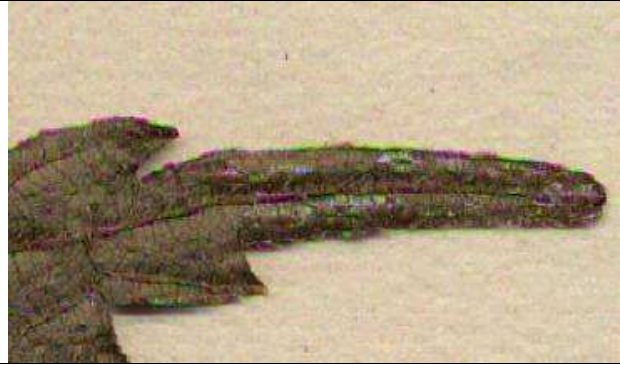

Vertical lines are sometimes caused by dust on the scanning head. These appear as vertical lines running through parts of or the entire specimen and are slightly colored (yellow or green).

	<p><b>ACTIONS:</b></p> <ul style="list-style-type: none"> <li>• <i>Image:</i> Reject the image and rescan the specimen</li> <li>• <i>Hardware:</i> If the line continues to appear, there is a potential problem with the scanner. Check all previously scanned images to see if it is an isolated event . Clean the scanner head if the line/s persists.</li> <li>• <i>Reporting:</i> Report the problem with the scanner to IT support and your project manager, noting the asset number of the scanner. If the problem continues, report it to the institution that provided you with the scanner or any of the LAPI Coordinators??</li> </ul>
--	---

### 4.3.3. Color Separation

The artifacts may show up on the image as varying degrees of bright streaks of color.. This is almost always seen in conjunction with pixilation. Some examples are show below. This should not be confused with light reflection.

<p>Note in the middle of this image there are bright colors around the edge of the specimen.</p>	<p>In this image, the colors appear in the shadow of the specimen and the problem is not immediately noticeable.</p>


	
<p>The bright streaks of color on this image are noticeable on both the edge of the specimen and on the specimen itself.</p>	<p>In this image notice the color effect in the shadow of the label.</p>

**ACTIONS:**

- *Image:* Reject the image and rescan the specimen
- *Hardware:* If color separation continues to occur there is a potential problem with the scanner. Check all previously scanned images to see if it is an isolated event .
- *Reporting:* Report the problem with the scanner to IT support and your project manager, noting the asset number of the scanner. If the problem continues, report it with the institution that provided you with the scanner or any of the LAPI Coordinators??

**4.3.4. Glue on Scanner Glass**

On occasion, glue applied to the barcode is transferred to the glass of the scanner. For this reason it is recommended that databasing is done before scanning or that the glue completely dry before the image is scanned.

	<p>On this image there is a patch of glue on the edge of the barcode label. It is only a problem if it affects all or part of the specimen.</p>
---	---

**ACTIONS:**

**If the specimen is affected or any label is illegible**

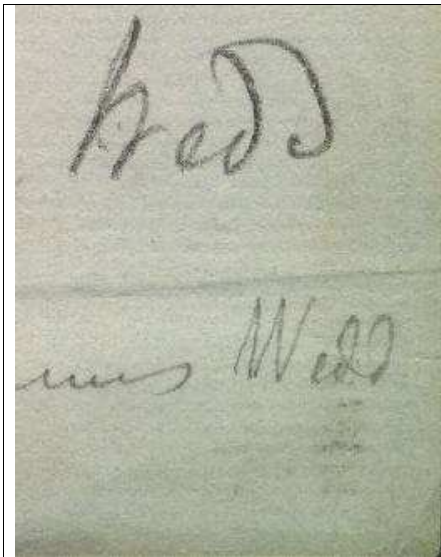
- *Image:* Reject the image and rescan after cleaning the scanner
- *Hardware:* Clean the scanner glass using recommended cleaning materials
- *Reporting:* Report the problem to your project manager.
- 

**If the specimen and labels are not affected**

- *Image:* Accept the image
- *Hardware:* Clean the scanner glass using recommended cleaning materials
- *Reporting:* Report the problem to your project manager.

**4.3.5. Green Cast**

Some of the scanners produce a green cast in the bottom right hand corner of the image. The cast will always appear in the same corner of the image.



In this image the color of the label becomes increasingly green towards the bottom right hand side of the image. If the specimen is not in the area of the green cast then it is not a problem. If the specimen is affected, it will need to be rescanned on a different scanner.

#### **ACTIONS:**

**If any part of the specimen is within the area of the green cast, or the label is illegible, or the color/greyscale targets are affected:**

- *Image:* Reject the image and rescan the specimen, placing it on the scanner in such a way that none of the important parts of the image are affected.
- *Hardware:* There is a problem with the scanner. Report the problem to support (see below) and check all images produced on the scanner in future.
- *Reporting:* Report the problem with the scanner to IT support and to your project manager, noting the asset number of the scanner. If the problem continues, report it with the institution that provided you with the scanner or any LAPI Coordinator??

**If no important component of the specimen or labels is affected in the ways described above**


- *Image:* Accept the image.
- *Hardware:* There is a problem with the scanner. Report the problem to support (see below) and check all images produced on the scanner in future.
- *Reporting:* Report the problem with the scanner to IT support and to your project manager, noting the asset number of the scanner. or any LAPI Coordinator??

#### **4.3.6. Other Artifacts (acceptable)**



Other artifacts that may occur but where image is Accepted:

***Light reflections from shiny surfaces:*** Light reflections will happen when the scanner light hits reflective/shiny surfaces on the specimen, such as glue and, sometimes, the barcode label. It will look like a rainbow effect, but should not be confused with the color separation problem.





	<p>On this image you can see bright patches of discoloration. This is due to the light in the scanner reflecting off the glue and does not indicate a problem with the image.</p>
<p><b>Actions - there is no problem with the image so no action is necessary.</b></p>	

**Parts of the image which may look like artifacts:** Sometimes a genuine part of the specimen image may look like an artifact. Some examples are below.

	
<p>In the middle of the image there appears to be a row of pixels that are missing. This is actually the light in the scanner hitting one of the very fine spines on the specimen which is close to the scanning glass.</p>	<p>As with the image above, the light in the scanner hits a spine close to the scanning glass.</p>
<p><b>Actions - there is no problem with the image so no action is necessary.</b></p>	

**Light on edge of objects:** There is a light source in the scanner which hits the specimen at an angle as it is scanned. Sometimes this can cause what looks like an artifact, but is not.

	
<p>In this image you can see what looks like a lighter colored iridescence at the top of the capsule. This is not a problem and the image is acceptable for our archives.</p>	<p>In this image you can see a white strip at the top of the label which is caused by light reflection. This is not a problem with the image.</p>
<p><b>Actions - there is no problem with the image so no action is necessary.</b></p>	

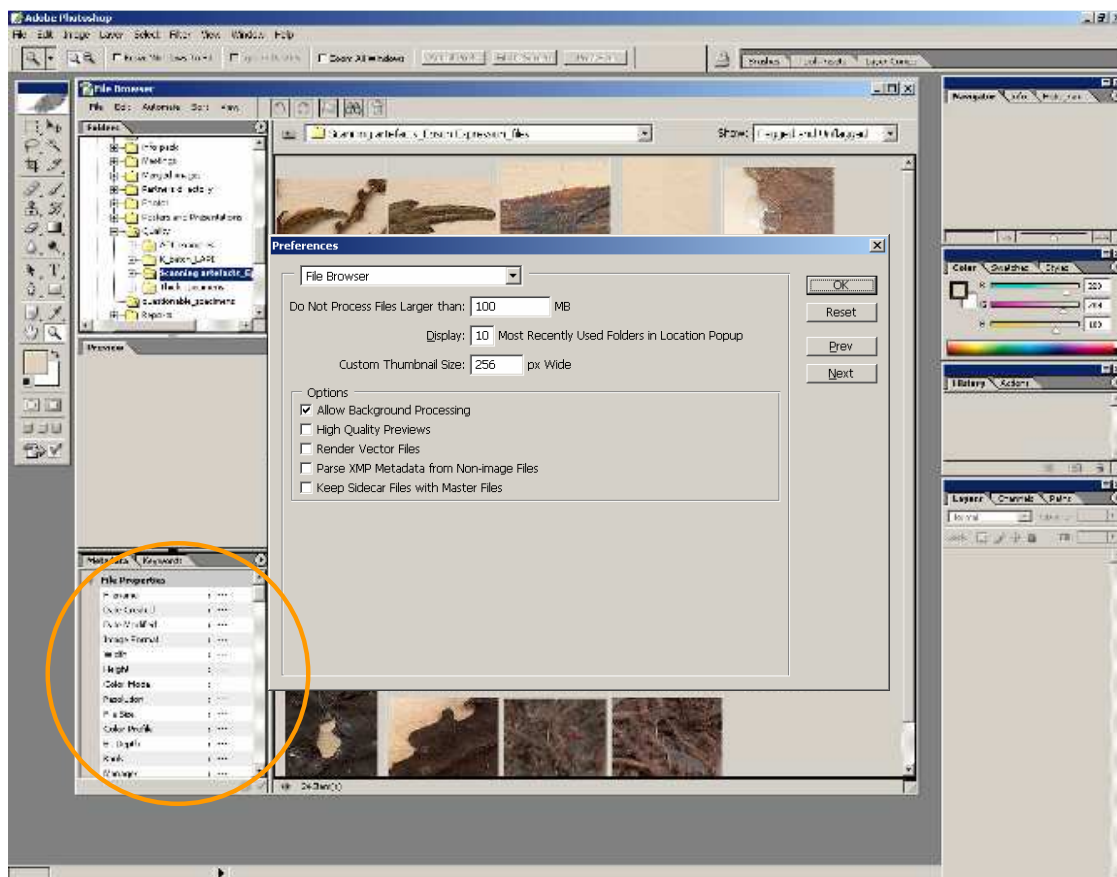
#### 4.4. Check Focus

Checking the focus should be done when viewing the images at 100% (actual pixel size) while moving across the entire image. You will be able to check most effectively by concentrating specifically on the following areas:

- Areas where there is a contrast in thickness between different parts of the material (e.g. stems or seeds against leaves or mounting paper)
- Edges – e.g. edges of leaves-do they look sharp?
- Labels – is the writing clear?

If there is a problem with focus on an image, the specimen will need to be rescanned.

#### 4.5. Check Scanning Settings



*To see the integral metadata included in an image file, select the thumbnail and wait for the information to load up in the lower left hand panel.*

*Note:* The following can be checked using the 'Browse' facility in Photoshop. Go to File 'Browse' and then browse to your chosen directory. This facility will load up thumbnails of all images in the directory unless the settings are changed to avoid this. We recommend changing the settings to prevent it loading up thumbnails for images over 100MB, otherwise the computer is likely to slow down. To change settings go to Edit > Preferences > File Browser and type 100 into the box 'Do not

process files larger than.’ You can also tick ‘Allow background processing’ and untick ‘High quality previews’. To see the integral metadata included in an image file, select the thumbnail and wait for the information to load up in the lower left hand panel.

#### **File Name**

Check that the file name matches the barcode.

#### **Resolution**

The resolution of the image file should be 600dpi, sometimes 650dpi if scanning on an old 1640XL Epson scanner.???(don’t know about this-needs to be verified)Anything below 600dpi is unacceptable and indicates that the scanner settings are incorrectly configured.

#### **Color**

Check that the image is in Adobe RGB (1998). If you have Adobe 98 RGB (1998) selected in your color settings, then an error message will appear when you open an image that has not been created in Adobe RGB (1998).

### **4.6. Following Up**

If a problem with an image is found, it may be necessary to check other images created by that same equipment and/or digitizer.

Where there is a problem with an image that cannot be corrected (e.g. corrupted file, pixilation, barcode missing or invisible), then this image must be removed from the hard disk. It is helpful to keep track of these problematic images in whatever manner works for the particular institution-I don’t like this sentence but something to this effect.

If a problem appears to be repeating with a particular digitization staff then the digitiser further training might be needed. If a problem appears to be repeating with a particular workstation then that workstation must be checked. Report the problem with the scanner to the institution that sent the equipment or any LAPI Coordinator.

### **4.7. Creating B&W GIF Images**

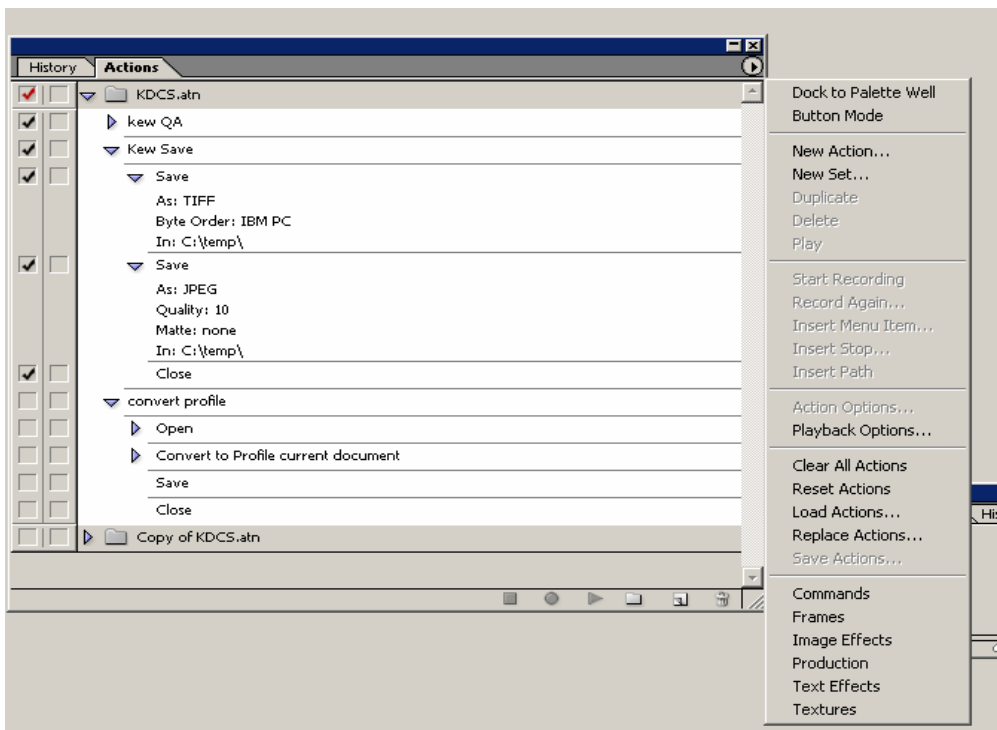
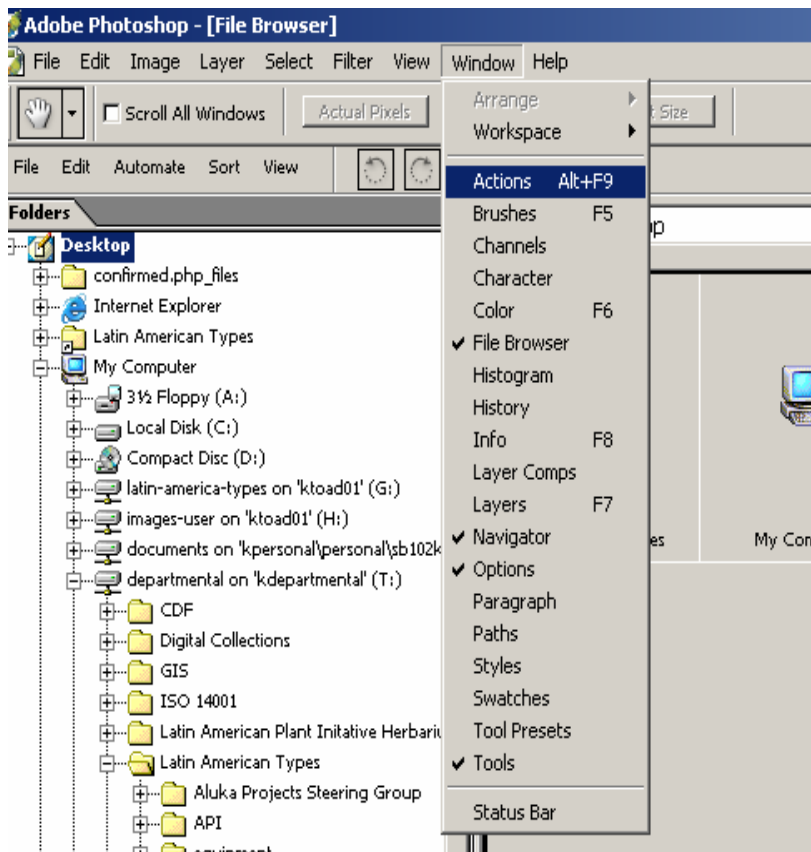
Kew Gardens created a process (using an “action” in Photoshop) for digitizers to check for artifacts and to ensure that digitization standards have been followed. This process includes converting the TIFFs to black and white GIFs to highlight potential problems on the images. The steps are outlined below:

To create GIF images the original TIFF images are run through a preloaded ‘action’ on Photoshop, thus creating an additional GIF image.

- 1) The action for highlighting scanning artifacts can be downloaded from the external GPI Partners website <http://plants-partners.jstor.org/>. Save a copy of the action (KCDS.atn) to a known location.
- 2) The action needs to be loaded into Photoshop.

How to load the QC action in Photoshop

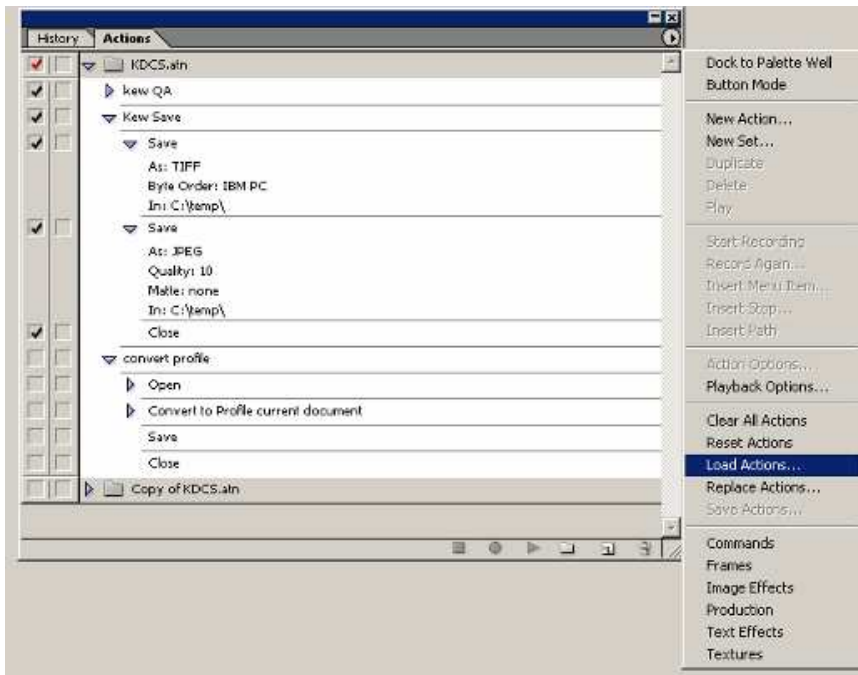
- 1) Open Adobe Photoshop CS.
- 2) Go to Window and choose the tab Action.



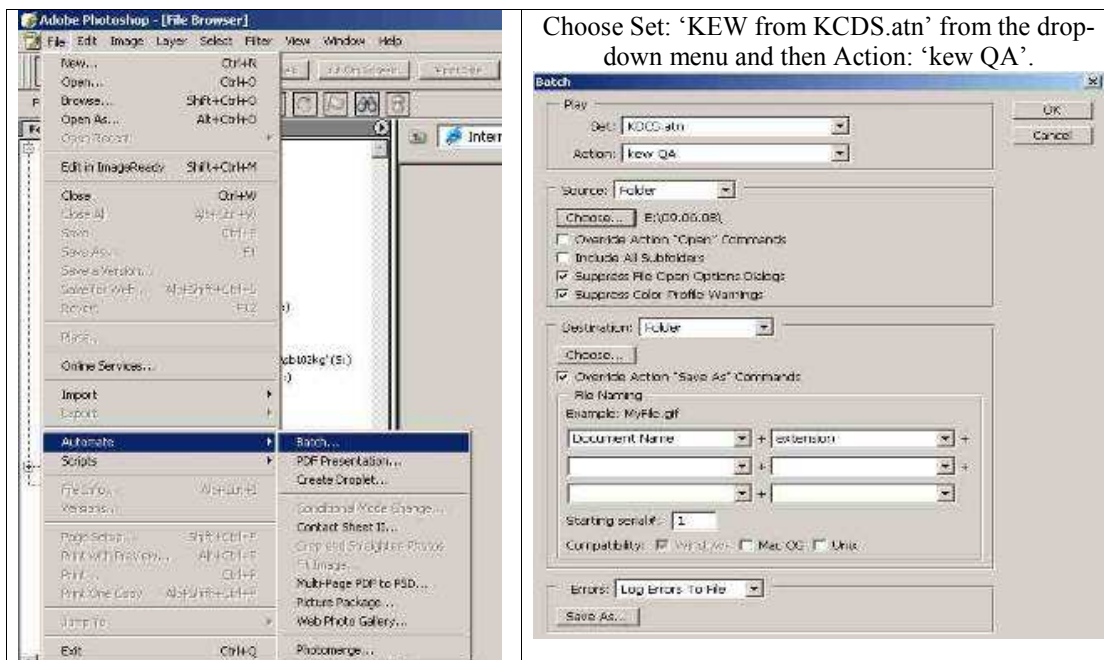
3) Click on

the arrow on that appears on the right of the tab :

- 4) Choose the option Load Actions.



- 3) You should have sufficient space to hold the GIF files. The GIF files are about 1/10 the size of the originals TIF images, so approximately 25MB of space will be needed overall.
- 4) To run the batch process in Photoshop choose File > Automate > Batch. This opens the Batch process dialogue box.



- 5) In Source ensure 'Folder' is selected and click 'Choose'. Navigate to the folder containing your images.
- 6) Uncheck the boxes 'Override Action "Open" Commands' and 'Include All Subfolders' (unless you have several subfolders and wish to check images from all of them, in which case ensure that you have chosen the folder containing these sub-folders in the previous step. Your GIFs will all be created in a single folder). This needs rewording
- 7) The remaining boxes should be ticked.
- 8) In Destination choose 'Folder' and navigate to the folder where you would like to save the GIF images resulting from this batch process. Create a new folder if necessary and name it appropriately (e.g. K\_batch\_1\_GIFs).
- 9) Leave the file-naming box as set at default.
- 10) In the Errors section choose 'Log errors to file' and then click 'Save As'. Navigate to an appropriate folder (e.g. the same as the GIFs folder) and choose an appropriate name for the error file (e.g. K\_batch\_1\_GIFerrors.txt).
- 11) Click 'OK' and the batch process will begin. On a high spec. computer (e.g. Dell Precision 370) this process takes around 90 seconds per image, therefore approximately 30 hours in total for a disk containing 1200 images. It can be left to run overnight and for a disk this size can usually be completed in two overnight sessions. This process uses most of the processing memory of the computer so should not be run in the background as you work.
- 12) In order to stop the process it seems necessary to force Photoshop to close via 'Ctrl+Alt+Delete' and End Task. In principle it should be possible to press the 'Stop' button in the Actions window, in practice the computer does not have thinking time to process this instruction!

13) If the batch process is stopped before the entire TIF files have been processed, temporarily move all images that have been processed into a new folder and carry out the same process the following night on the remaining images.

### Checks on GIFs

Once the GIF images have been created with the above process each image needs to be checked for artefacts and to ensure that digitization standards have been followed. We recommend using free download software called *IrfanView*. (<http://www.irfanview.com/>) for checking the GIF images. By using this software it is possible to click rapidly through the black and white GIFs in ‘full screen’ view or to set up a slide show that automatically scrolls through the images every 3-4 seconds. For each image check for:

#### Corrupted files

Corrupt TIFFs will interrupt these processes—you will need to investigate the original image if this occurs.

#### Vertical lines

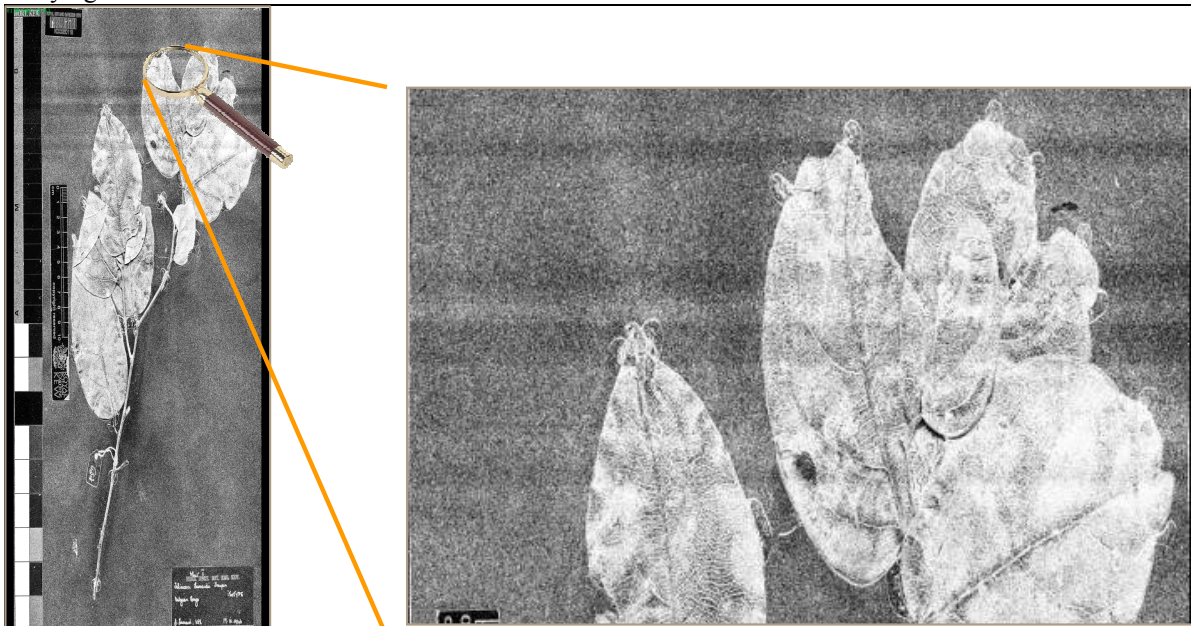
Vertical lines are sometimes caused by dust on the scanning head. These appear as vertical lines through the whole or part of the specimen and are slightly colored (yellow or green).



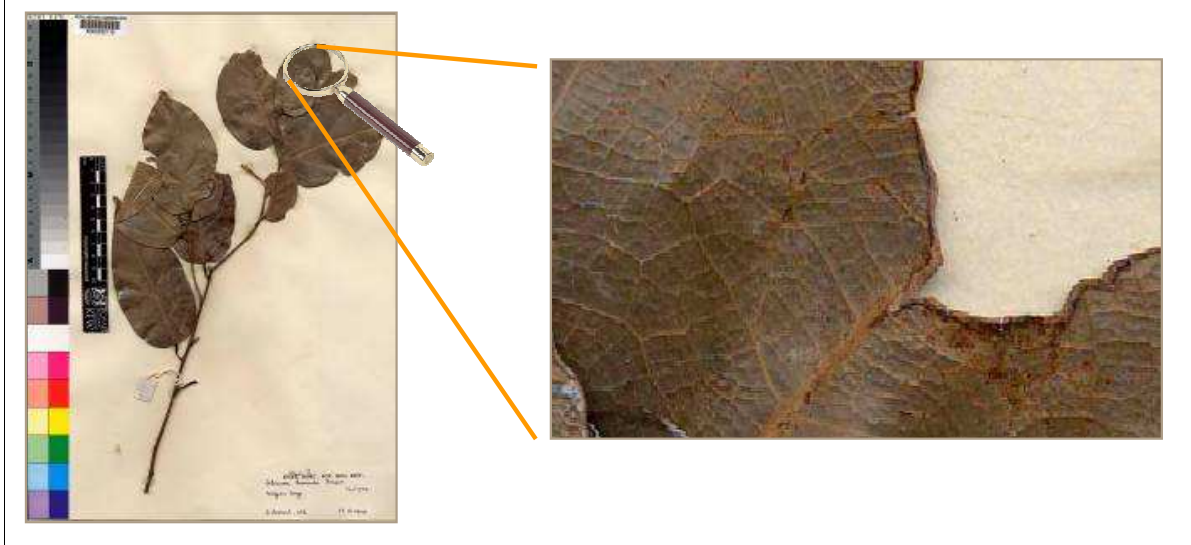
GIF image and corresponding TIF showing a vertical line due to dust in the scanner head.  
REJECTED IMAGE.

## Scanning artifacts

Scanning artifacts (pixilation or color separation) tend to appear as dark or light horizontal bands of varying thickness.



**Example of a scanning artifact due to pixilation:** these horizontal bands in the GIF (see above) reflect pixilation problems in the original image (see below)



## 5. Export

### 5.1. Introduction



Images of type specimens supplied to JSTOR must be accompanied by their corresponding specimen label metadata. The metadata is typically captured in a specimen database or spreadsheet but no specific database software is mandatory for the GPI project. Instead, each institution can use their preferred software but must export the specimen metadata into a standard XML (Extensible Markup Language) format designed for the project. This format is defined by an XML Schema. This document provides an explanation of the fields that make up the XML schema and guidelines for creating and validating the XML prior to delivery. Each partner must develop their own method for databasing their specimens and exporting the data into the highly structured GPI XML file format.

## **5.2. Principles**

### **Why Standardize?**

Each GPI partner may utilize their own method for the capture of the specimen metadata for each of the scanned specimens. However, these different metadata sources must be joined into a single repository at JSTOR. To enable the joining of the disparate data sources, they must be unified into a single, uniform database. By standardizing on a common XML schema, all partners have a pattern to follow to transform their original data into a standard form that can be combined. However, conforming to a standard generally means that each partner will have some extra data transformation work to produce the standard XML metadata file. But, the effort to produce the standardized output will enable the combined repository to be created in way that enables each partner's data to be presented accurately and completely in the combined repository. Effectiveness of searching can be significantly enhanced through the existence of rich, consistent metadata about the specimens.

### **Metadata**

For the GPI project, specimen metadata refers to the specimen label data recorded about a scanned specimen. This data is generally recorded in a database, but sometimes can be in a spreadsheet or text document. Examples of metadata are the barcode number, determination, collector, locality, collection dates, etc.

For the GPI project, metadata must be recorded for each digitally imaged type specimen. For each batch of image files submitted to JSTOR, an XML file must be submitted containing metadata for each image and the file must comply with standards established for the project.

*Note:* Specimen metadata must be shipped in the same batch as the corresponding images. JSTOR will not begin processing the batch until both have been received. Only when both the images and the XML file for the corresponding images are in hand will we complete ingestion and quality control.

## **5.3. XML Schema**

The format of the file to be used for exporting GPI data is XML. A standard XML schema originally developed for the African Plants Initiative is used for the Latin American Plants Initiative. The schema defines the required and optional elements of the XML and constrains the content and structure of the metadata being exported.

Partners should not create individual XML files for each specimen. One XML file must be generated for all specimens contained in one batch. JSTOR will parse the XML file into individual records.

All XML files must follow a standard structure to be readable by an XML viewer. The rules of this structure are defined in an "XML Schema" document. The standard schema for the GPI project is AfricanTypesv2.xsd which is available at the JSTOR website.

More information about XML and XML Schemas can be found at:

XML Tutorial: <http://www.w3schools.com/xml/default.asp>

O'Reilly's XML Tutorials: <http://www.xml.com>

XML Schema Primer: <http://www.w3.org/TR/xmlschema-0/>

### **Data Export from Institutional Database**

Depending on the database software used at each institution, the methods for extracting data will vary. It is up to each partner to figure out the best method for mapping the fields of their database to the fields needed for the XML schema, modifying (if necessary) the fields to fit the structure of the XML schema, and generating the XML file.

#### **5.4. GPI XML Generator**

Rafael Barron, the XML Coordinator for the GPI project, has created an Access database that will automatically create the XML necessary for delivery to JSTOR from a simple Excel spreadsheet or Access table. This GPI XML Generator removes the need for institutions with limited IT support to fully understand how XML and XML schemas work. However, it does require that an institution is able to export data from their in-house database and reformat the data to fit the fields in the schema.

An Excel spreadsheet (or Access table) with the fields necessary to use the GPI XML Generator is provided with the system and the user just needs to reformat their data to fit the spreadsheet. Once the data is in the correct format in the spreadsheet or table, the user only then needs to import it into system. GPI XML Generator will identify problems in the metadata, check taxonomic names against TROPICOS, automatically export the metadata into the proper XML format, and keep track of sent batches. More details can be obtained from Rafael Barron at [rafael.barron@mobot.org](mailto:rafael.barron@mobot.org). There is also a PowerPoint presentation on the project website that describes the use of the system.

#### **5.5. Batches, Datasets, and Units**

Digitized images will be submitted to JSTOR in groups called Batches. A Batch consists of approximately 1,000-1,200 TIFF files that have been created by the HerbScan device, one for each digitized specimen image. These files will be placed on an external hard drive to be shipped to JSTOR.

The term Dataset refers to the metadata for a Batch. A Dataset is an XML file which contains the metadata for all of the image files in the Batch. There will be one Dataset on each external hard drive when it is submitted. A one-to-one match is required between the specimen metadata records/UnitIDs in the Dataset and the specimen image files. Each metadata record/UnitID must match an image file, and each image file must have a metadata record.

#### *Exception*

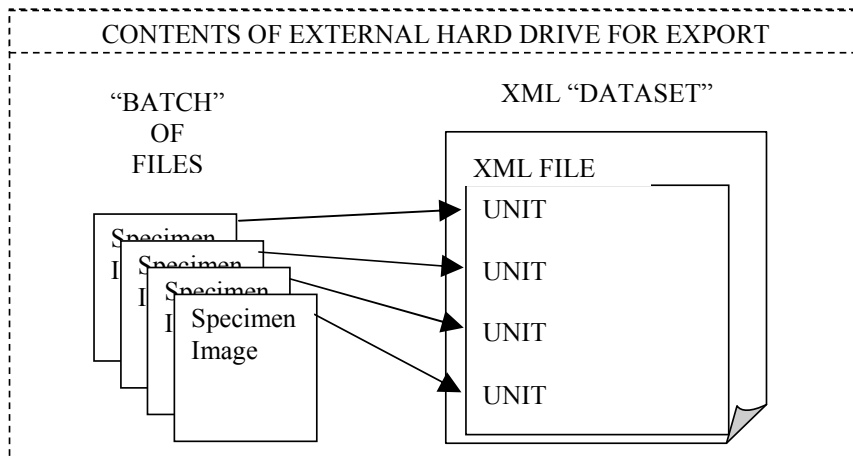
Where multiple images are made for single specimen, such as a more detailed scan of a part of the sheet or multiple sheets for single specimen, there can be multiple files for the same UnitID, but the files must be named UnitID\_a, UnitID\_b, etc.

#### *Note*

The term “dataset” is used generically to refer to the XML file. There is also an XML tag called DataSet that is part of the structure of the XML file.

The term Unit refers to a single specimen. All of the metadata for a single specimen image file is associated with that Unit. There is an XML tag called Unit that is part of the structure of the XML file.

The following diagram illustrates these concepts:



Missouri Botanical Garden will be playing an important role in helping partners to export the data about the specimens correctly. Please contact Rafael Barron ([rafael.barron@mobot.org](mailto:rafael.barron@mobot.org)) for assistance before you export your data.

### 5.6. XML File Name

The XML file for the GPI metadata must be named according to your institution’s Index Herbariorum acronym (<http://sweetgum.nybg.org/ih/>), batch number and date using the following filename format:

institution\_batch\_YYYYMMDD.xml

Batches are numbered sequentially starting from 0 for the first test batch, and then 1, 2, 3 etc. for each batch sent to JSTOR.

Example:

Institution: Missouri Botanical Garden (IH code: MO)

Batch number: 2 (the second batch ever sent to JSTOR)

Date: 1 November 2006

Filename would be: MO\_2\_20061101.xml

### XML Header (Mandatory)

Every XML file must have a header section. The header must be the first line of the file and contain the following:

```

<?xml version="1.0" encoding="UTF-8" ?>
<DataSet xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance":noNamespaceSchemaLocation="http://plants.jstor.org/xsd/AfricanTypesv2.xsd">
<InstitutionCode>NY</InstitutionCode>
<InstitutionName>The New York Botanical Garden</InstitutionName>
<DateSupplied>2009-05-08</DateSupplied>
<PersonName>Melissa Tulig</PersonName>
<Unit>

```

**Kommentar [DRyan1]:** Add file to partner site

## 5.7. General XML Formatting Rules

The following general XML formatting rules must be followed in constructing the GPI XML metadata file. The XML document:

- must begin with the XML declaration (header)
- must have one unique root element (“DataSet” for GPI)
- all start tags must match end-tags
- XML tags are case sensitive (unlike HTML!)
- all elements must be closed
- all elements must be properly nested
- all attribute values must be quoted
- XML entities must be used for reserved characters

### XML Entities for Reserved Characters

The following text characters are reserved by the XML structure and cannot be included in any metadata values. The “XML Entities” must be substituted for these characters before being included in the exported XML file.

Reserved Character	XML Entities
Greater than >	&gt;
Less than <	&lt;
Ampersand &	&amp;
Quote “	&quot;
Apostrophe ‘	&apos;

#### Note

If these reserved characters are inadvertently included in the XML value data, the XML file will produce unpredictable results when viewed with a browser.

For Example:

```

<Locality>Koopmansfontein: Agricultural Research Station; Golden Rock. Pan. 28°11'49.7"S
24°06'17.9"E</Locality>

```

Should look like:

```

<Locality>Koopmansfontein: Agricultural Research Station; Golden Rock. Pan.
28°11&apos;49.7&quot;S 24°06&apos;17.9&quot;E</Locality>

```

Correct Spelling of Tags

All GPI XML tags must be spelled correctly using the correct letter case.

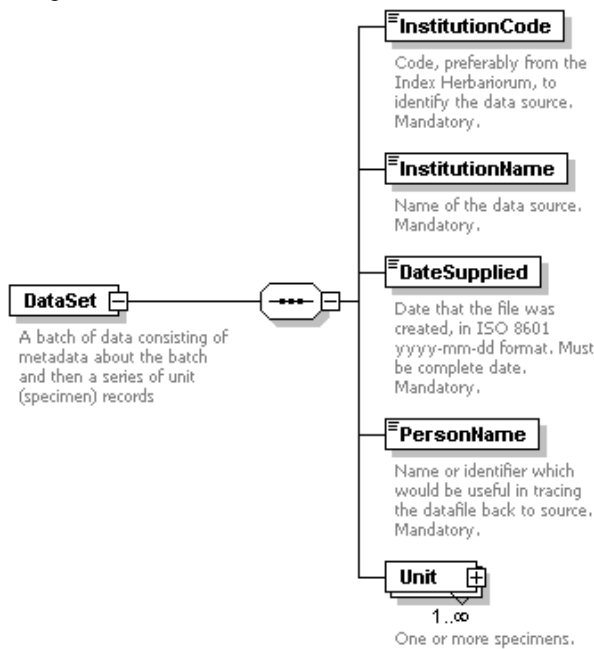
Correct Tag	Incorrect Tags
DataSet	Dataset, dataset
StoredUnderName	Storedundername, storedundername
InstitutionCode	Institutioncode

### 5.8. XML Schema Fields

The GPI XML Schema consists of a “DataSet” tag with 5 main tags:

- InstitutionCode
- InstitutionName
- DateSupplied
- PersonName
- Unit (repeats, one for each specimen image file)

This is a diagram of the top-level schema structure:



#### DataSet Tag (Mandatory)

The DataSet tag is required for the GPI schema. Its form is:

```
<DataSet xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://extranet.aluka.org/xsd/AfricanTypesv2.xsd">
.
.
.
</DataSet>
```

This tag must be spelled “DataSet” and not “Dataset”, “dataset” or “dataSet”.

### **InstitutionCode, InstitutionName, DateSupplied and PersonName Tags (Mandatory)**

Each of these four tags is mandatory.

1. The InstitutionCode value must be from Index Herbariorum (<http://sweetgum.nybg.org/ih/>). If there is no IH code for the institution then a code will be assigned for GPI.
2. The InstitutionName is the name of your institution.
3. The DateSupplied is the date of the creation of the metadata file.
4. The PersonName is a contact at the Institution for potential follow-up.

An example of the form of these tags is:

```
<InstitutionCode>K</InstitutionCode>  
<InstitutionName>Royal Botanic Gardens, Kew</InstitutionName>  
<DateSupplied>2004-04-01</DateSupplied>  
<PersonName>John Jones</PersonName>
```

### **Unit Tags**

There must be at least one Unit tag for each GPI XML file. But an unlimited number of Unit tags can be included. Each Unit tag usually represents one image file in a Batch. The UnitID and the image filename name must match based on the barcode number.

#### *Exception*

Where multiple images are made for a single specimen, such as a more detailed additional scan of a part of the sheet or multiple sheets for a single specimen, there can be multiple image files for the same UnitID, but the files must be named UnitID\_a.tif, UnitID\_b.tif, etc.

Since there are many potential metadata values associated with a specimen image, the Unit tag has many sub tags available to be used. Some are required and some are optional.

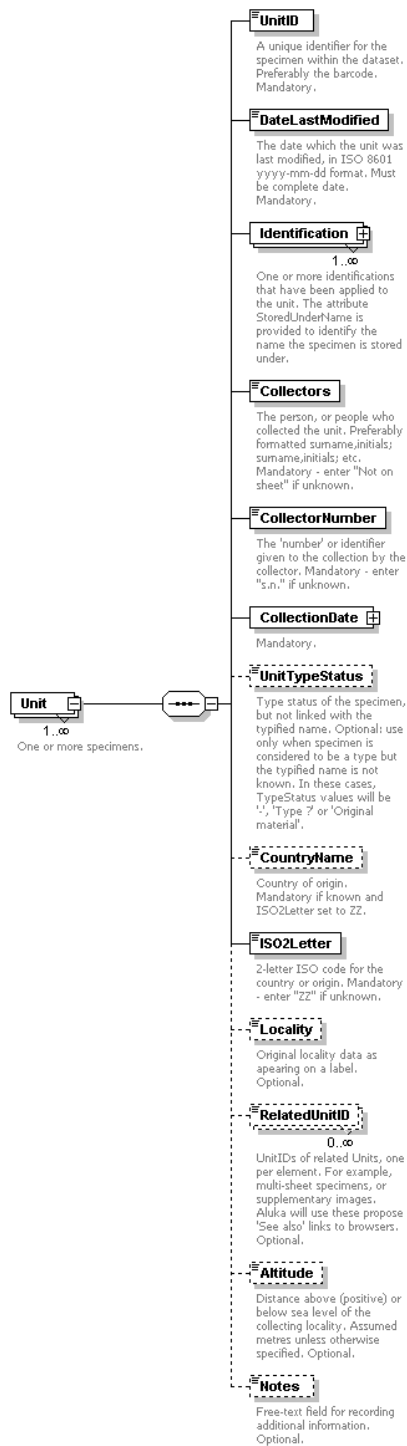
There are 7 required tags for each Unit:

1. UnitID
2. DateLastModified
3. Identifications (at least one)
4. Collectors
5. CollectorNumber
6. CollectionDate
7. ISO2Letter

There are 6 optional tags for each Unit:

1. UnitTypeStatus
2. CountryName
3. Locality
4. RelatedUnitID
5. Altitude
6. Notes

Here is a diagram of the Unit tag and its subtags:



## Explanation of Required Tags for Unit Tag

### UnitID (Mandatory)

This is a unique identifier for the Unit or specimen image. It is required to be the same as the InstitutionID concatenated with the barcode on the specimen and not preferable as stated in the schema description. The UnitID is also required to be the same as the name of the specimen image file for the Unit. One of the quality checks for data submission is to compare the UnitIDs in the XML dataset with the filenames on the hard drive for an exact match (except for the multi-image files with \_a, \_b extensions that have been noted above.)

Example in XML:

```
<UnitID>K10000081</UnitID>  
<UnitID>US00987634</UnitID>
```

Invalid UnitIDs

```
<UnitID>12121212</UnitID> (No Index Herbariorum acronym)  
<UnitID>34567ABC</UnitID> (Index Herbariorum acronym should be at start of number)  
<UnitID>L444555666</UnitID> (Where L is not the institution code)
```

*Note*

Do not include any spaces in the UnitID value

If the barcode on the specimen does not include the institution's code as a prefix, the UnitID and the image filenames can be created on output from the institution's metadata by concatenating the institution code with the barcode.

### DateLastModified (Mandatory)

This date refers to the last time any part of the metadata for this Unit or specimen image was changed. This "last change date" is generally recorded in the primary database of the scanning institution. Note that the format is yyyy-mm-dd, including the dashes.

Example in XML:

```
<DateLastModified>2006-06-23</DateLastModified>
```

### Identification (Mandatory)

Each Unit can have multiple Identification tags. The intent of this is to enable recording of multiple names associated with a specimen sheet rather than just one determination. For instance, a single specimen could have multiple determinations by different specialists, plus it could have different names for which it is a type, and it could have a name it is filed by.

*StoredUnderName Attribute*

The Identification tag has a mandatory attribute called StoredUnderName. This StoredUnderName attribute is either "true" or "false". One and only one Identification tag can have a "true" value for StoredUnderName, because no specimen can be stored or filed under more than one name.

In other words, the value of "true" for StoredUnderName should occur only once for all Identification tags within a single Unit. All other Identifications within a single Unit must have a value of "false" for StoredUnderName. This rule is checked as part of the quality control of submitted GPI XML metadata files.

*Note*



Do not use “0” or “1” or “Yes” or “No” or “True” or “False” for the value of StoredUnderName. The only acceptable values are either “true” or “false”. Only one true and as many false as needed can be used.

*How is StoredUnderName intended to be used?*

In some herbaria, specimens are placed inside folders or organized by one name, the “stored under” name, but may also have a different determined name or type name on the sheet. Using this attribute allows the name chosen by the institution as the stored or filed name to be distinguished from any others associated with the specimen.

*What if we do not use “Stored Under” Names?*

In many herbaria, no distinction is made between the determined name and the way it is stored or filed. And some herbarium databases do not record multiple names for a single specimen. In either of these cases, a single name would be recorded in the specimen database for the sheet.

If only one name is recorded in the database, then the StoredUnderName attribute must be “true” for that name, since that one name is serving the purpose of a stored or filed name.

Example in XML:

<Identification StoredUnderName="true"> *Note: should occur only once for any Identification tag within a single Unit.*

.  
</Identification>

Or

<Identification StoredUnderName="false"> *Note: All other Identification tags within a single Unit must have a value of “false”*

.  
</Identification>

**Identification Required Subtags**

Each Identification has a further 7 required subtags:

**Family (Mandatory)**

Based on the scanning institution’s own taxonomic decisions or as shown on the sheet. It is recommended that the family is entered in all uppercase letters.

Example in XML:

<Family>ASCLEPIADACEAE</Family>

**Genus (Mandatory)**

As recorded by the scanning institution. First letter should be uppercase.

Example in XML:

<Genus>Secamone</Genus>

**Species (Mandatory)**

As recorded by the scanning institution. Should be all lowercase.

Example in XML:

<Species>grandiflora</Species>

### **Author (Mandatory)**

The author of the species name including basionym author and ex/in authors following standard format. Standard Author Abbreviations should be used based on Authors of Plant Names maintained by RBG Kew at <http://www.ipni.org/ipni/authorsearchpage.do>

Example in XML:

```
<Author>Klack.</Author>
```

### *Note*

If species author is missing or unknown, use “Not on sheet”.

### **Identifier (Mandatory)**

The name of the person recorded by the scanning institution who made the determination of this Identification.

### *Note*

Use “Not on sheet” if the identifying/determining person is not known.

Example in XML:

```
<Identifier>Not on sheet</Identifier>
```

### **IdentificationDate (Mandatory)**

The date recorded by the scanning institution for when the determination of this Identification was made. The date value is not entered directly under this tag. Rather the standard Date subtags must be used.

### *Date Subtags*

For CollectionDate and IdentificationDate 7 subtags are used to specify the date value:

1. StartDay
2. StartMonth
3. StartYear
4. EndDay
5. EndMonth
6. EndYear
7. OtherText.

Only one of these subtags is mandatory. However, it is expected that if a Day is recorded, there must also be a Month and Year. And, if a Month is recorded, there must be a Year. A value for Year can be provided without Month or Day. And, there should not be an end date value if there is no start date value.

### *Note*

At least one subtag of Date must have a value.

If there are no date values on the sheet, then “Not on Sheet” must be inserted into OtherText.

The Start and End tags are designed for date ranges most often associated with the Collection Date, not the Identification Date. For this reason, it is mostly likely you will only use the StartDay, StartMonth, and/or StartYear subtags for Identification Date.

Example:

```
<IdentificationDate>
  <StartDay>27</StartDay>
  <StartMonth>01</StartMonth>
  <StartYear>1992</CollectionDate>
</IdentificationDate>
```

Or where there is no date:

```
<IdentificationDate>
  <Other Text>Not on Sheet</OtherText>
</IdentificationDate>
```

### TypeStatus (Mandatory)

The type status of this Identification. Use “-” if the Identification is not a type name.

Use *ONLY* one of the following values for type status:

Holotype	Epitype
Isoepitype	Lectotype
Isolectotype	Neotype
Isoneotype	Paratype
Isoparatype	Syntype
Isosyntype	Isotype
Type	Type ?
Original material	-

Any other type status recorded by the institution must be converted to one of these.

#### Note

*If a specimen is known or thought to be a type specimen, but the name for which it is a type is unknown, then all Identifications will have TypeStatus of “-”. If you are working with non-type specimens, also use “-” for this field..*

Example in XML:

```
<TypeStatus>Holotype</TypeStatus>
```

Each Identification has a further 6 optional tags:

1. **GenusQualifier – (Optional)** Qualifier expressing doubt about the genus epithet (eg. cf)

Example in XML:

```
<GenusQualifier>cf</GenusQualifier>
```

2. **SpeciesQualifier – (Optional)** Qualifier expressing doubt about the species epithet (eg. cf)

Example in XML:

```
<SpeciesQualifier>cf</SpeciesQualifier>
```

3. **Infra-specificRank – (Optional)** Rank based on ICBN and as recorded by the scanning institution. Should be all lowercase.

Example in XML:

```
<Infra-specificRank>var.</Infra-specificRank>
```

4. **Infra-specificEpithet – (Optional)** As recorded by the scanning institution. Should be all lowercase.

Example in XML:

```
<Infra-specificEpithet>alba</Infra-specificEpithet>
```

5. **Infra-specificAuthor – (Optional)** Follow the same guidelines as for the Author subtag.

Example:

```
<Infra-specificAuthor>Wild.</Infra-specificAuthor>
```

6. **PlantNameCode – (Optional)** An optional code that is meaningful to the scanning institution for the name given for this Identification. Often a tracking number. May be used to provide feedback from JSTOR to the scanning institution.

Example in XML:

```
<PlantNameCode>123ABC</PlantNameCode>
```

### **Collectors (Mandatory)**

This is just a text string listing the Collector or Collecting Team for this Unit. The preferred way of listing a Collecting Team is:

Surname1, Initials1; Surname2, Initials2; Surname3, Initials3

using a semi-colon to separate the individual collectors. The Senior Collector should be listed first. If no collector data is available, then the value “Not on sheet” must be manually inserted in the XML. This tag cannot be left blank.

Example:

```
<Collectors>Beentje, H.J.; Quansah, N.</Collectors>
```

Or if there are no collectors recorded

```
<Collectors>Not on Sheet</Collectors>
```

### **CollectorNumber (Mandatory)**

This is generally the number assigned by the senior collector to the specimen. But, it can contain letters if needed. Where there is no collector’s number for the Unit, the value “s.n.” must be inserted in the XML. This tag cannot be left blank.

Example:

```
<CollectorNumber>4559</CollectorNumber>
```

Or

```
<CollectorNumber>4559, 4560</CollectorNumber>
```

Or if there is no collector number

```
<CollectorNumber>s.n.</CollectorNumber>
```

**CollectionDate (Mandatory)**

The date recorded by the scanning institution for when the collection was made. The date value is not entered under this tag. Rather the standard Date subtags must be used. An explanation of the subtags can be found under the section for the IdentificationDate tag.

Example:

```
<CollectionDate>  
  <StartDay>27</StartDay>  
  <StartMonth>01</StartMonth>  
  <StartYear>1992</CollectionDate>  
</CollectionDate>
```

Or if there is no collection date

```
<CollectionDate>  
  <OtherText>Not on Sheet</OtherText>  
</CollectionDate>
```

**ISO2Letter (Mandatory)**

This is the 2-letter ISO 3166-1 code for the country where the specimen was collected. The ISO 3166 master list is available at

<http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html>

The following table lists all of the ISO 3166-1 Caribbean and Central and South American country codes:

COUNTRY NAME	ISO CODE	COUNTRY NAME	ISO CODE
ANTIGUA AND BARBUDA	AG	GUYANA	GY
ARGENTINA	AR	HAITI	HT
ARUBA	AW	HONDURAS	HN
BAHAMAS	BS	JAMAICA	JM
BARBADOS	BB	MARTINIQUE	MQ
BELIZE	BZ	MEXICO	MX
BOLIVIA	BO	NETHERLANDS ANTILLES	AN
BRAZIL	BR	NICARAGUA	NI
CAYMAN ISLANDS	KY	PANAMA	PA
CHILE	CL	PARAGUAY	PY
COCOS (KEELING) ISLANDS	CC	PERU	PE
COLOMBIA	CO	PUERTO RICO	PR
COSTA RICA	CR	SAINT KITTS AND NEVIS	KN
CUBA	CU	SAINT LUCIA	LC
DOMINICA	DM	SAINT VINCENT AND THE GRENADINES	VC
DOMINICAN REPUBLIC	DO	SOUTH GEORGIA AND THE SOUTH SANDWICH ISLANDS	GS
ECUADOR	EC	TRINIDAD AND TOBAGO	TT
EL SALVADOR	SV	TURKS AND CAICOS ISLANDS	TC
FALKLAND ISLANDS (MALVINAS)	FK	URUGUAY	UY
FRENCH GUIANA	GF	VENEZUELA	VE
GRENADA	GD	VIRGIN ISLANDS, BRITISH	VG
GUADELOUPE	GP	VIRGIN ISLANDS, U.S.	VI
GUATEMALA	GT		

When the country is missing or unknown for a specimen, the 2-letter code “ZZ” must be inserted into the XML.

*Note*

Where the institution has not utilized ISO codes in its data system, a conversion will need to be made from the country coding system used by the institution to the ISO2 system before inclusion in the

XML metadata file. The GPI XML Generator will automatically add ISO2 codes if the CountryName field is populated in the import file.

Example in XML:

```
<ISO2Letter>BR</ISO2Letter>
```

### **CountryName (Optional)**

This name is not needed if an ISO2Letter value has been provided. Only provide a value for this tag if there is no ISO2 code for the country and the value of “ZZ” has been inserted for ISO2Letter. This tag allows an unusual country name not recognized by ISO to be assigned to the Unit or specimen image.

Example in XML:

```
<CountryName>AnUnusualCountryName</CountryName>
```

### **Locality (Optional)**

This is the literal string of text that was recorded for “locality” describing where the specimen was collected. No other atomization of location is available in the schema. State, County, Municipio, latitude and longitude, etc. must all be concatenated and added to the Locality field.

Example in XML:

```
<Locality> AMBANJA: Manongarivo Special Reserve, Bekolosi Mt.; in open montane forest. </Locality>
```

The original data needs to be carefully examined before export to XML to replace any of the XML reserved characters - <, >, &, ‘, and “. Also, if locality data includes accented or extended ASCII characters refer to the UTF-8 information discussed earlier on section What is UTF-8?

### **UnitTypeStatus (Optional)**

This tag is only used by RBG Kew. All other institutions will omit it.

#### *Note*

*If a specimen is known or thought to be a type specimen, but the name for which it is a type is unknown, then for GPI nothing is to be recorded in the XML metadata for UnitTypeStatus.*

### **RelatedUnitID (Optional)**

This is a multiple occurrence tag, so multiple UnitIDs for related Units can be included. This will be the UnitID (IH code concatenated with barcode) of another specimen. There is no attribute to describe or classify the nature of the relation to the other specimen, just the existence of a relation.

Example in XML:

```
<RelatedUnitID> K0123456</RelatedUnitID>
```

### **Altitude (Optional)**

This is a number in meters of the altitude above sea level where the specimen was collected. Please enter ‘meters’ or ‘feet’ and not just the first character ‘m’ or ‘f’.

Example in XML:

```
<Altitude>100 meters </Altitude>  
Or  
<Altitude>25.5 feet</Altitude>
```

### Notes (Optional)

This can be any text and has no other constraint. Be cautious of the reserved characters and extended characters as with other text. See section on XML Entities for reserved characters.

Example in XML:

```
<Notes>This can be any text except reserved characters like &apos; </Notes>
```

### Handling of Missing Data Review

For the GPI project, specific data values have been chosen to be used when no data is available for required XML tags, rather than just leaving them empty. This enables standardization of the data for smoother integration into the JSTOR repository.

Using “Not on Sheet”

The text string “Not on Sheet” is required to be used for the following required tags, if no data is available. This value is not required to be recorded in the institution’s database or regular specimen data format. Each institution is free to record missing or blank data as it chooses. But, on output of the XML metadata, the recorded values must be converted to “Not on Sheet” in the XML file.

Use “Not on Sheet” for missing, empty or blank values for:

- Unit/Identification/Author
- Unit/Identification/Identifier
- Unit/Collectors
- Date/Other Text – if there is no Month, Day, or Year

Use “s.n.” for missing, empty or blank values for:

- Unit/CollectorNumber

Use “ZZ” for missing, empty or blank values for:

- Unit/ISO2Letter:

### Save as UTF-8

If any Windows extended characters are included in the metadata export, steps must be taken to ensure that the XML file is converted to UTF-8 encoding. This will not happen by default with Microsoft Office tools.

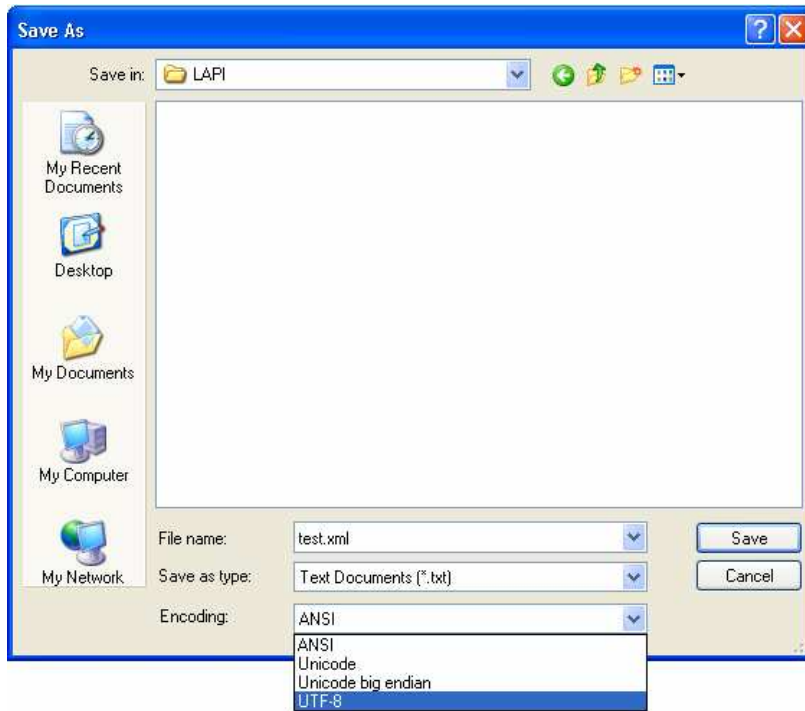
UTF-8 refers to the underlying method to be used for the text characters included in the XML file. UTF-8 is one way of displaying Unicode; there are others, but UTF-8 is used for GPI.<sup>2</sup>

One simple method is to use Microsoft Notepad, Version 5, or later to open the exported XML file and then select Save As and choose UTF-8 as the Encoding at the bottom of the window as in this example:

---

<sup>2</sup> The main impact of specifying UTF-8 encoding for an XML file is that it actually *excludes* a very common form of data encoding used in Western countries, namely the Microsoft “symbols” which can be inserted in the Western versions of Microsoft Word, Excel and Access.





## 5.9. Validating the GPI XML file

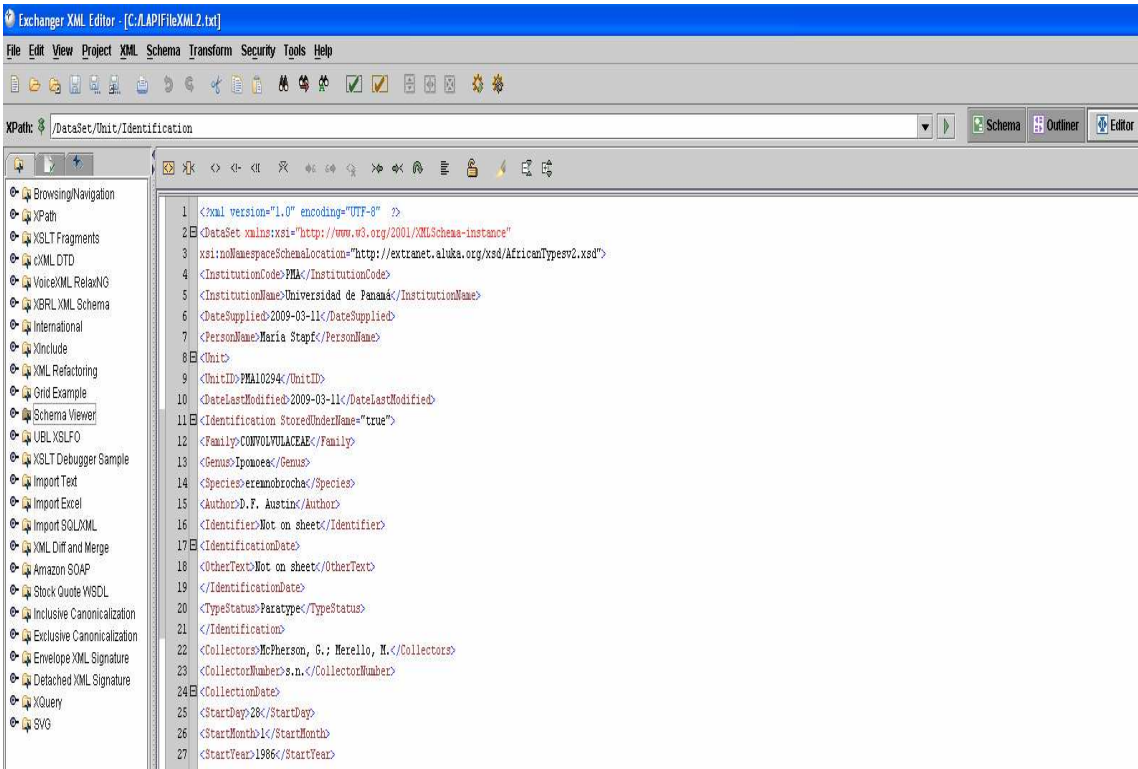
The GPI Metadata XML file must be validated before sending to JSTOR. Validated means that the file content and structure conform to the basic XML formatting rules as well as the GPI standard schema.


### *Using Exchanger XML Lite 3.2 Editor to Validate an XML file*

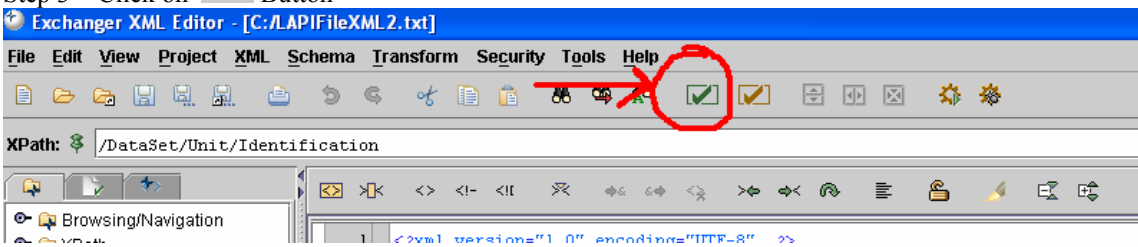
One easy way to validate an XML file is by using Exchanger XML Lite 3.2. This is a free program that can be downloaded here: [www.freexmleditor.com](http://www.freexmleditor.com). The following provides an example of using Exchanger XML Lite 3.2 to validate.

*Step 1* – Run **Exchanger XML Lite 3.2**.

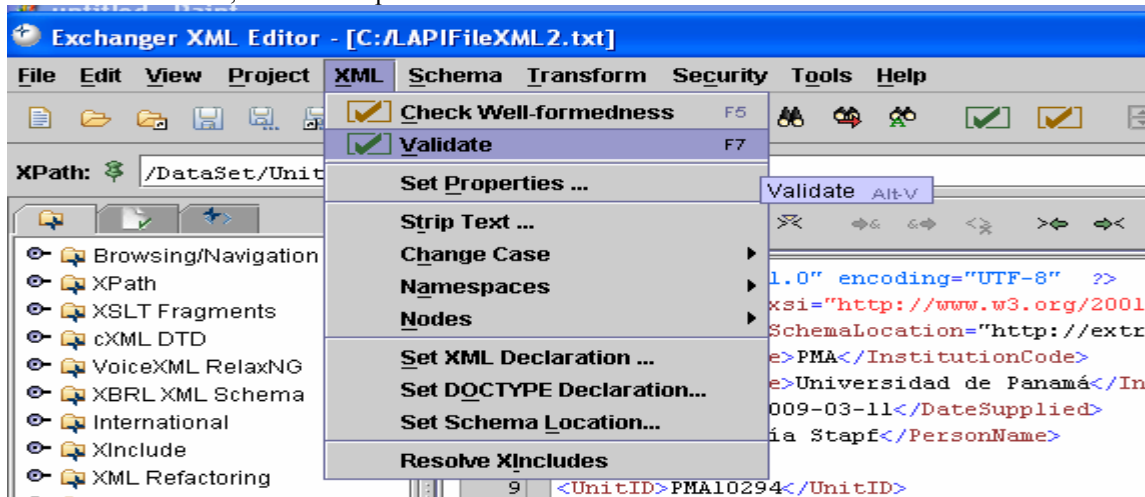
*Step 2* – Click on **File, Open** and select the Metadata XML file. The contents of the XML file will be shown like this example:



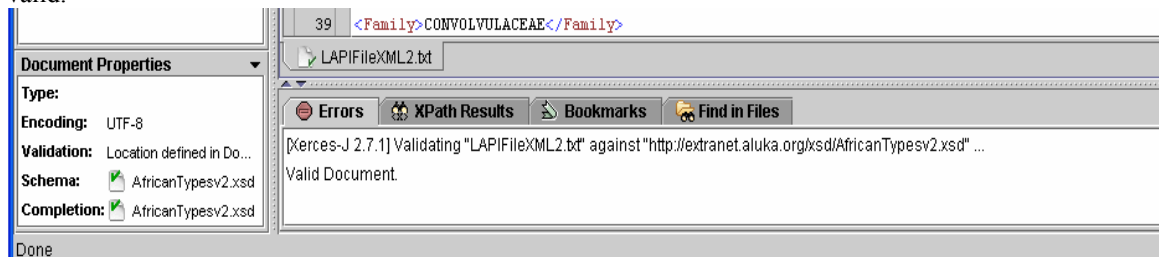
Step 3 – Click on  Button



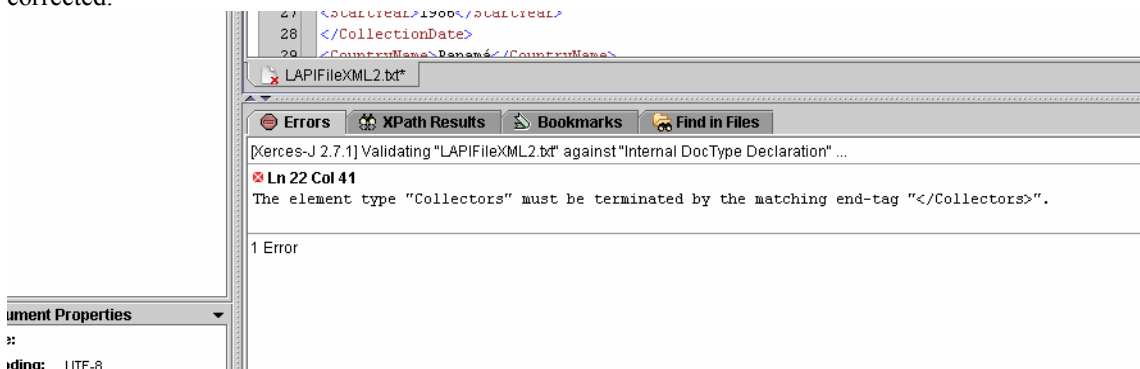
OR – Click on **XML, Validate** or press **F7**.



Step 4 – A message will appear at the bottom of the screen indicating whether the file is valid or not valid.




Here is an example of a not valid message, caused by a mismatch end tag. All such errors must be corrected.

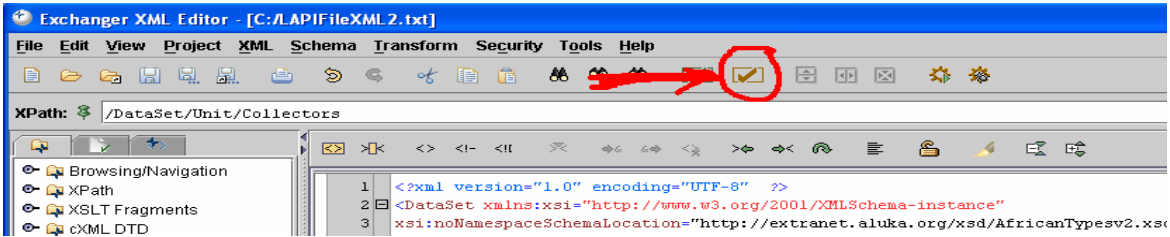


The file must be valid before it can be sent to JSTOR.

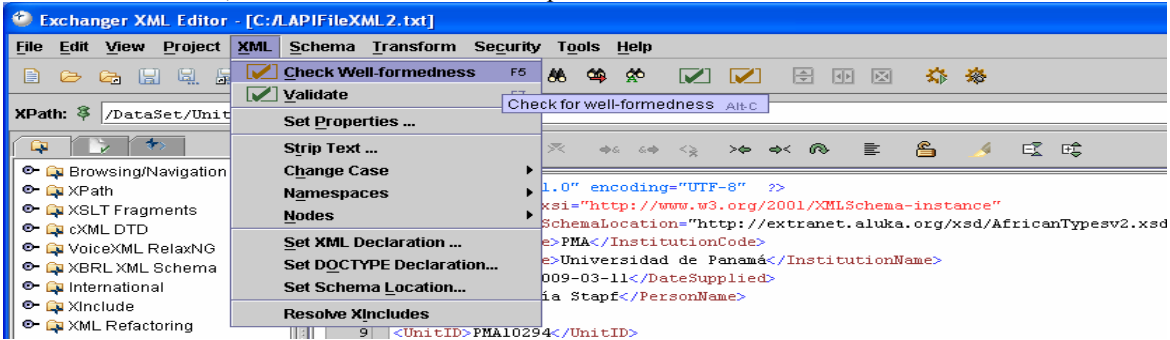
Using *Exchanger XML Lite 3.2* to see if it's **Well-Formed**.

Use the **Well-Formed** button to see if the active document is well-formed XML.

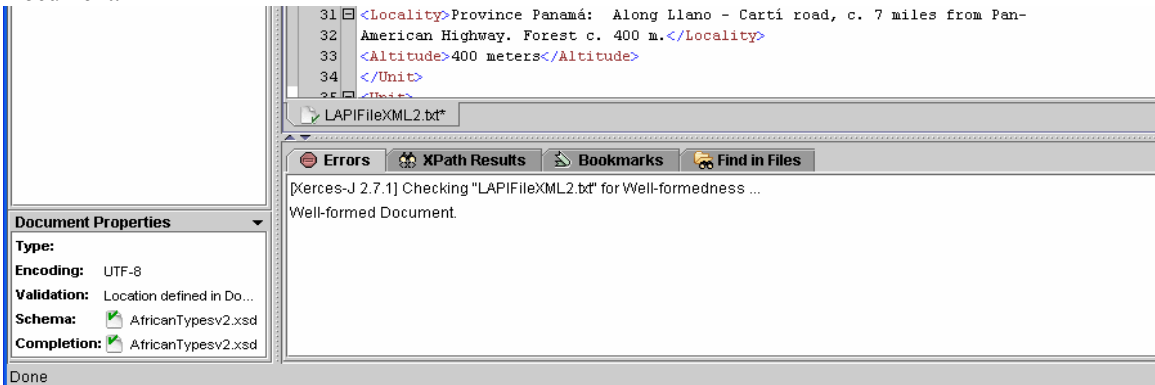
Step 1 – Click on  Button



OR – Click on XML, Check Well-formedness or press F5.



Step 2 – A message will appear at the bottom of the screen indicating whether the file is well-formed Document.



### Other XML File Validation Tools/Websites

There are other websites or software tools that can be used to validate the XML Metadata file:

XML Cooktop – free download – Windows only - <http://www.xmlcooktop.com/>

XML Fox – free download – Windows only - <http://xmlfox.com/download.htm>

XML Buddy – free download – Eclipse-based - <http://www.xmlbuddy.com/>

Website - <http://www.xmlvalidation.com>

Website - <http://www.validome.org/xml/>

## 6. Transfer to JSTOR

A copy of the images and specimen metadata must be sent to JSTOR. Regular communication with JSTOR about the scheduling and submission of images & specimen data is critical. Once you have become an official partner you will receive an email from JSTOR welcoming you to the project.

**NOTE:** Submissions are tracked by partners and JSTOR by batch number. A batch typically contains 1,000-1,500 images. The first submission (test batch) is labeled batch 0, the first official batch is batch 1, and so on. The batch number is noted on all correspondence about shipping, receiving and processing your images and data. Partners need to start submissions for GPI at batch 01.

### 6.1. Test Batch of Images and Specimen Data

#### *Test Batch: Images*

Before starting the digitization process, you are REQUIRED to send a test batch of 10-15 images with their corresponding databased information for initial review via CD-ROM, DVD or FTP to JSTOR.

1. CD-ROMs / DVDs

Please post CD-ROMs with sample images to the following address:

Ithaca  
JSTOR Production  
100 Campus Drive, Suite 100  
Princeton, NJ 08540  
Attention: Production/Plants

2. FTP

Images can be zipped and uploaded to our public FTP site:

ftp.aluka.org  
Username: aluka  
Password: upload101

**NOTE:** If images do not meet the GPI standards the sample will be rejected. You will be required to digitize and submit another sample within one month. We will work closely with your team to achieve a high-level of quality to prevent extra work.

#### *Test Batch: Specimen data*

Missouri Botanical Garden provides a critical role in helping partners to export the specimen data correctly. Please contact Rafael Barron ([rafael.barron@mobot.org](mailto:rafael.barron@mobot.org)) for assistance before you export your data.

Once you have sent a sample to Rafael Baron please also send a sample XML file via FTP or CD-Rom with your sample test batch of images. Associated specimen data must be sent with the images or within 30 days of sending the images. JSTOR can not evaluate the test batch until both have been received.

**NOTE:** The majority of problems partners experience in submitting their content to JSTOR do not result from image or data quality issues. Rather, they result from discrepancies between the specimen data supplied in the XML and image files. Often images are missing specimen data and/or vice versa. It is critical to always run queries/checks prior to dispatching hard drives to ensure specimen record barcodes and image name barcodes match.

## 6.2. Hard Drives



For data transfer it is recommended that image and metadata files be delivered on external hard drives and shipped to the JSTOR Princeton office. JSTOR can purchase and ship the hard drives to your institution if they were not included in your original proposal.

JSTOR recommends the Western Digital's Passport FireWire/USB Compatible drives. If you plan on using another type of external hard drive please contact us. DVDs or CD-ROMs are acceptable for herbaria with relatively small collections.

## 6.3. File Directory Structure

The images should be organized on the hard drive by saving them in one folder labeled "images". Please do not place images in separate or nested folders—doing so creates numerous problems for ingestion and backup of your drive once it is received at JSTOR.

Likewise the XML file containing the specimen data and the excel worksheet containing the image technical metadata should be placed in a folder labeled "data" without additional subfolders.

Example:

```
images/BR13974009.tif  
BR25028757.tif  
etc.  
data/BR_2_20090511.xml
```

**NOTE:** Please do not place specimen images in various subfolders (i.e. by date, digitizer's name, plant family, etc.)

If a drive contains more than one batch, create a folder for each batch. Within that batch folder place the images and XML.

## 6.4. Shipping

Please ensure that the external hard drive is labeled with your institution's name and note the drive's serial number for your records.

For JSTOR's DHL or FedEx account information please contact JSTOR.

Shipping address:

Ithaca  
JSTOR Production  
100 Campus Drive, Suite 100  
Princeton, NJ 08540

Tax ID # 30-0152775

For some countries, you must also include a customs letter (on your institution's letterhead) declaring that the drive is for the purpose of sharing botanical information and is NOT for sale. Include your Tax-exemption number.

Please also send an email alert to JSTOR citing the date of shipment, contents of drive (number of image and data files) and the hard drive serial number. Please check that the number of primary images match the number of metadata records in the XML file.

## 6.5. Schedule

Our goal is to ensure a timely and efficient workflow between your institution and JSTOR. When your first batch is received at JSTOR you will receive a confirmation email. Within 4 weeks you will receive an email containing a Quality Control Image and Data report as well as a link to Xumba from your Production Assistant. If the images and data do not meet the accepted standards, the shipment will be rejected. It is crucial to perform in-house quality control of all digitized images and metadata before sending a shipment to JSTOR.

Please ship a hard drive for every 1,000-1,200 images. JSTOR will off-load, check and return hard drives in a timely fashion. If possible partners should submit a proposed schedule for the submission of their batches for the year.

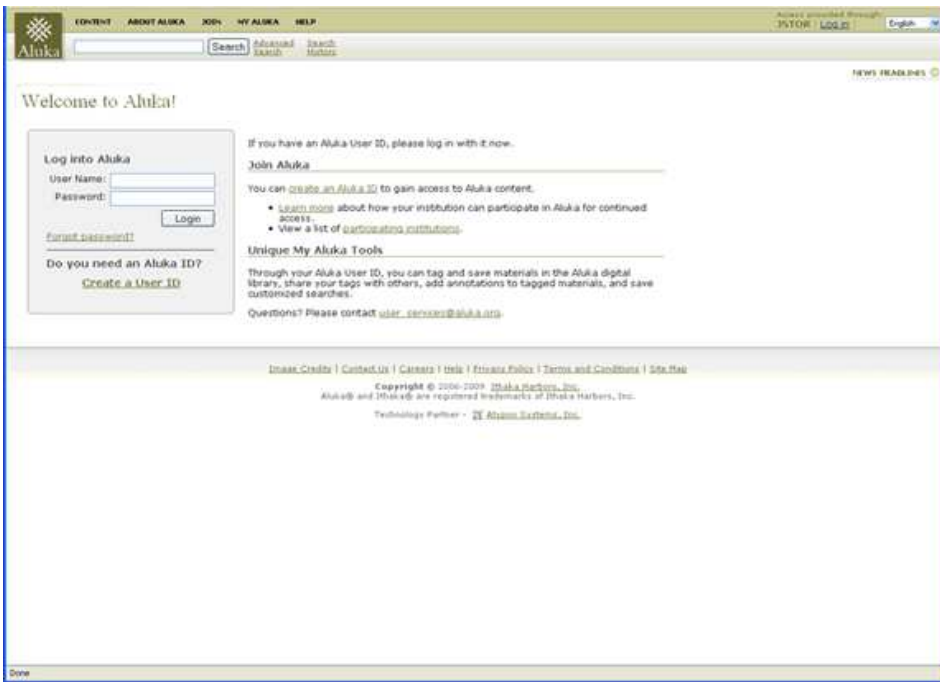
**Note:** If your institution has less than 500 specimens please send them in a timely manner.

## 7. Xumba

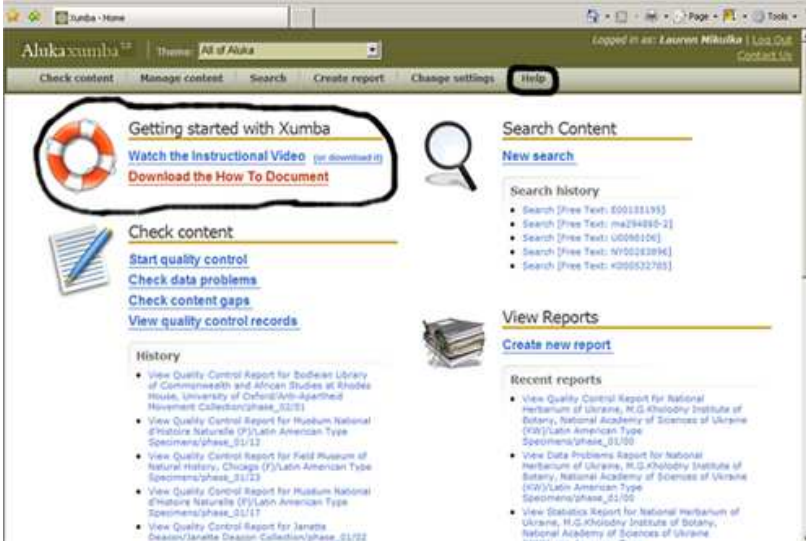
Xumba is a web interface that provides access to the ingestion, metadata, quality control and publishing status reports for your institution's content. It acts as a staging area for your images prior to their publication to <http://plants.jstor.org/>. You can use Xumba to analyze the current status of your content, isolate problematic objects and correct metadata. You can also track changes and resubmissions from ingestion, through the QC process, to publication on the website.

### 7.1. Access

You need a username and password to access your institution's content. If you haven't received one or are unable to access content in Xumba as expected, please contact [plants@jstor.org](mailto:plants@jstor.org) for assistance. Once you have a username and password visit <http://www.aluka.org/xumba> and log in.



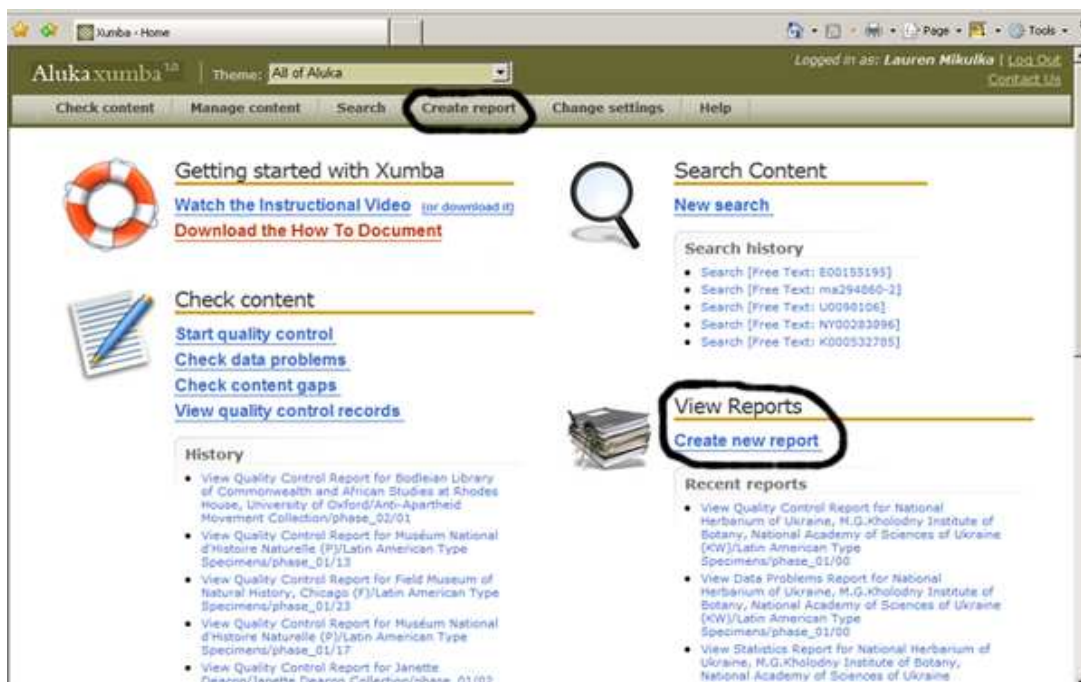
After you log in there will be the home page. If this is your first time on Xumba it is helpful to watch an instructional video and download the How-To Document as a quick reference. If at any time you need additional information or assistance there is a help tab on the right of the tool bar.



## 7.2. Checking Reports

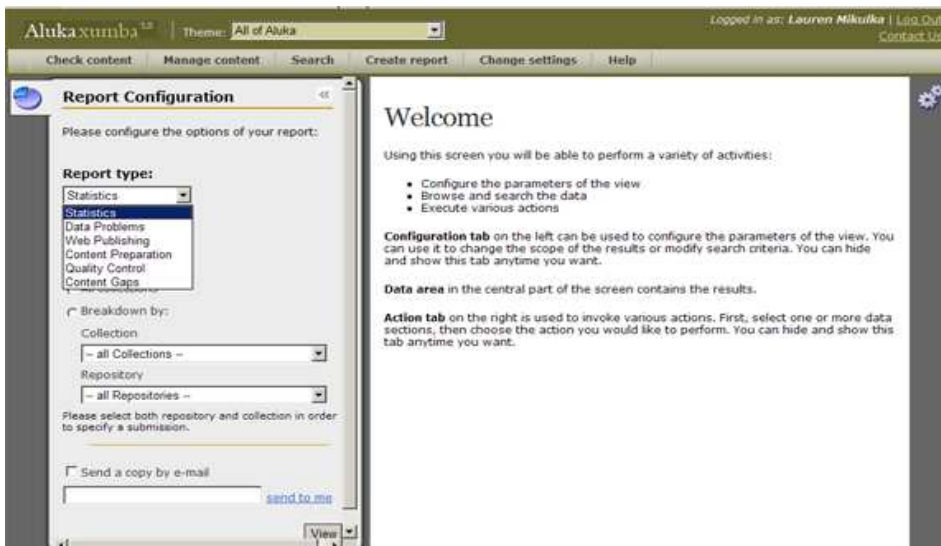
Xumba has the option to “**Create new report.**” This will allow you to check on your institutions image and metadata status for all batches.





The Report Configuration screen will appear on the left. You can select from 6 types of reports:

- **Statistics:** View the number of objects submitted to JSTOR and the number of unique documents. (Note: If a given document has multiple images, the number of objects will be higher than the number of unique documents since each image is counted separately.)
- **Data Problems:** Isolate metadata problems in the object data for your institution's content.
- **Web Publishing:** Get a quick look at the total number of objects either published or not published to <http://plants.jstor.org>.
- **Content Preparation:** Check the status of previously submitted, recently submitted or corrected content for your institution.
- **Quality Control:** View the current status and progress of content as it moves through the QC process.
- **Content Gaps:** Check for missing images or corresponding files for the content from your institution.








You can refine the scope of your report in several ways:



- All repositories (you will only have the ability to view your institution)
- All collections (you will only have the chance to view your collections)
- Breakdown by
  - all Collections (or a specific collection)
    - TYPSP will give you access to API submissions
    - LTYPSP will give you access to Global Plants submissions
  - all Submissions (or a specific submission)
    - Note: the submissions are broken down by phase and batch (example: phase\_01/02 . All batches submitted for plants are referred to as phase\_01. The 02 stands for the batch.

You can also email a copy of the report to yourself or to a colleague..

When viewing any report here is the necessary key to understand each report:

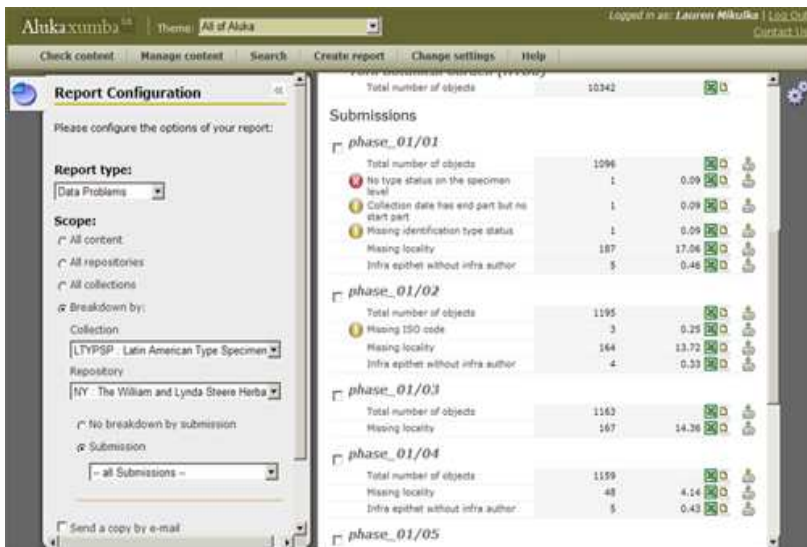
-  Error: Indicates an error in the metadata for the number of objects in the corresponding Values column. If this appears next to a value these objects will not be published until the issue is corrected.
-  Warning: Indicates some metadata may be problematic for the number of objects in the corresponding Values column.
-  Identifiers List: Download an excel file of the identifiers for the group of objects in the corresponding row.
-  Link to Records: View the list of records in Xumba, fitting the description of the corresponding row.
-  Download Batch Metadata: Allows the user to download the entire batch of metadata to view in spreadsheet form. Depending on the size of the file, download times will vary.


- [10.5555/AL.AP.SPECIMEN.NY00887871](#): Links to your specimen. Click the identifier and you will be able to see the Quality Control status, view the specimen (click on the Viewer tab on the right side of the screen), and view/edit the specimen metadata.

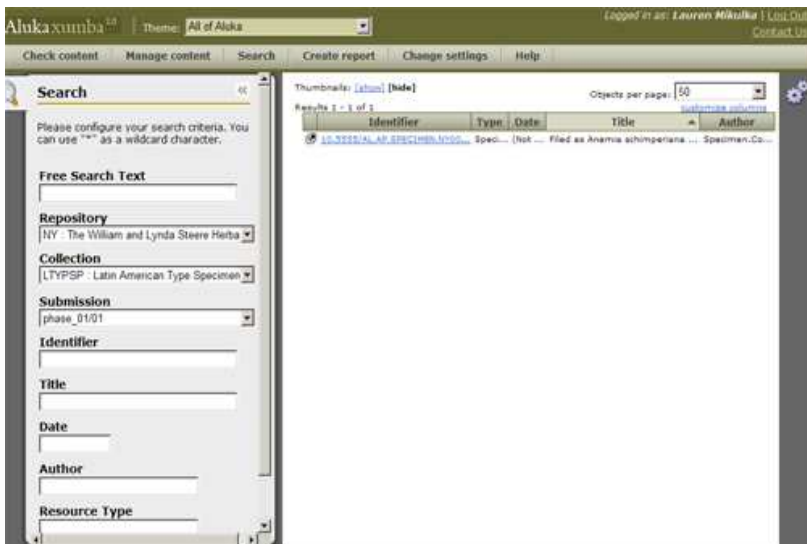
One of the more useful reports for all institutions to check is the Data Problems. You can check all submissions or you can check a single submission. To view a list of the objects in a specific category, such as **No Type Status on the Specimen level**, select the **See records icon** . This will display a list of all of the objects in this category. You can also select the **Download identifiers** Excel icon  in order to list the object identifiers on an Excel spreadsheet.

**Kommentar [D2]:** Partners will be confused by this report, most will be using QC to see what images need to be redone or fixed.

**Kommentar [D3]:** This doesn't export object data, it just provides the identifiers for each object. Exporting object data is done on a different screen.



The see records icon  will provide you with a list of identifiers. To view information on a particular object, you can click the **Identifier** link and bring up the object's data page.



When viewing the image screen you will have the opportunity to view the image submitted and the status of the image (approved or rejected).



You can also view the specimen metadata and the publishing status at the bottom of the page. Please note that this image is rejected for publishing because the “file is not found.” This does not mean that we do not have the image tiff file, as you should be able to view it on Xumba. However, it does mean that the image is not yet available to be viewed on <http://plants.jstor.org/>. This can take up to 4 weeks.

#### Publishing records

Status	Date	Who	Notes
Rejected	2009-09-28 11:34:12 AM	system_post	Object is rejected because of missing file (image or other) '/fpx/alukaplant/f/phase_01/f0004/f0071534f.fpx';

### 7.3. Data and Image Corrections

Partners can update their specimen data on a regular basis by resubmitting their data to JSTOR or by loading the metadata directly into Xumba. Individual records can be corrected directly on Xumba but multiple records can be submitted as an XML file via email to JSTOR. In addition, multiple records must be organized by batch. Images can be sent via ftp or sent on the next hard drive in a separate folder marked “Redo.” Once the data is resubmitted it can take up to 72 hours to show up on the <http://plants.jstor.org> site.

If editing an individual metadata record, open the image and scroll to where the specimen metadata is located. You can click on **Edit Specimen** and when you are done save your changes.

