Getting the Bioimages Collection Manager Software

Download the latest version from
https://github.com/kopolzin/bioimages-collection-manager/releases/latest
Click on the downloaded file to install and accept defaults.

Kinds of things that Bioimages tracks

Images
The primary focus of Bioimages is images of living organisms. Preferably, contributors photograph details of all parts of the organism including the whole organism itself. The images are grouped according to the organism they depict.

Organisms
Often images of the same "organism" are images of the same biological organism. However, the definition of "organism" that Bioimages uses is that of Darwin Core: "A particular organism or defined group of organisms considered to be taxonomically homogeneous." This is a practical definition, because sometimes it is difficult to tell where one organism ends and another begins, e.g. in a head of coral or clump of grass. As long as the entity is taxonomically homogeneous, it can be an "organism" used to group images in Bioimages. This includes small groups of animals or plants, e.g. a herd of deer photographed together, a tray in a greenhouse containing a single species of plants, or a roadside population of plants of which various individuals may be photographed without tracking each individual plant. The following controlled vocabulary terms are used to indicate the scope of the "organism": "biological organism" for actual individual organisms, "pack" for groups of animals, and "colony" for groups of plants.

Agents
An agent is a person, group of persons, or organization that plays some role in the generation of data in the Bioimages database. Agents can have the following roles:
- photographer (creator of images)
- determiner (assigner of taxonomic determinations to organisms)
- rights owner (holder of image copyright)
- provider (manages the submission of images to Bioimages, responsible for metadata quality, and ensures consistency of local identifiers)

Organizations that manage collections (botanical gardens, zoos, arboreta) are also considered to be agents.

In many cases, a single person will fill all of these roles with respect to a set of images. However, it is possible for someone other than the photographer to be the rights owner, someone other than the photographer to manage submission of images, or for multiple persons to assign taxonomic determinations to a set of organisms or even to a single organism.

Identifiers

Local identifiers vs. globally unique identifiers
Identifiers are a string (a sequence of characters) used to uniquely identify some kind of resource ("thing") so that it can be referenced and distinguished from other resources. An identifier can be a local identifier or a globally unique identifier.

A local identifier is unique within a certain context, such as a particular collection of images or the specimens in a museum's collection. There is no guarantee that a local identifier that is unique in one context will be different from some other local identifier that is unique in a different context. For example, a photographer may refer to an image in her collection using the local identifier "jb34692" to differentiate it from all other images in her collection. But a museum may have a specimen in its collection that also has the local identifier "jb34692". "jb34692" is unique in the specimen collection as well, but "jb34692" is not globally unique because "jb34692" can refer to two different things depending on the context.

There are a number of strategies for generating identifiers that are globally unique. Bioimages uses a strategy based on domain names. The IANA (Internet Assigned Numbers Authority) assigns Internet domain names in a way that ensures that no two organizations have the same one. This means that no other organization can use the domain "vanderbilt.edu" besides Vanderbilt University. The subdomain "bioimages.vanderbilt.edu" has been assigned to Bioimages, so we can be sure that no other organization can use that subdomain name. Appending "http://" to the front of the subdomain indicates that it's a domain that is Internet-accessible.

If Bioimages appends some local identifier to "http://bioimages.vanderbilt.edu/" it can be sure that identifier is globally unique as long as the local identifier is used to identify only one thing within the Bioimages database. However, just appending a contributor's local identifier may cause an identifier "collision" if two contributors use the same local identifier for different things. In the example above, "http://bioimages.vanderbilt.edu/jb34692" might refer to the image or the museum specimen. To avoid this problem, Bioimages uses "namespaces". A namespace is a local identifier assigned to a person or organization that takes responsibility for a set of local identifiers to ensure that they are locally unique and not reused in the context of the resources that person or organization manages. So if the local identifier "photographer1" were assigned to the photographer and "museum5" were assigned to the museum, the local identifier collision would be avoided by generating the globally unique identifiers: "http://bioimages.vanderbilt.edu/ photographer1/jb34692" and "http://bioimages.vanderbilt.edu/ museum5/jb34692".

Image identifiers
Bioimages generates local image identifiers semi-automatically based on the file names of images. It takes a certain number of characters from the end of the file name (not including ".jpg" or ".JPG") to use for a local identifier. This number is appended to an image namespace that is often related to the photographer's name. For example, "baskauf" is the namespace used for Steve Baskauf's images. The last five characters "35224" are pulled from the image filename "coob9-flfront35224.jpg" to create a local identifier that is unique in the context of Steve Baskauf's images, since he makes sure that no two images that he has contributed ever have the

same last 5 characters. By appended that identifier to the namespace "baskauf", an identifier "baskauf/35224" is formed that is unique in the context of Bioimages. A contributor can choose how many characters to use from the end of an image filename.

Because local identifiers are related to file names, they should be composed only of characters that are "safe" for filenames: letters, numerals, dash ("-"), and underscore "_". Although capital letters are allowed, it is best to use all lower case letters. File names/local identifiers must not contain spaces. Although underscores are fine to be included within filenames and local identifiers, local identifiers should not begin with an underscore - this may cause unpredictable effects in some systems. For example, if a filename is IMG_12345.jpg and the last 6 characters are used for the local identifier, that would result in a local identifier of "_12345", which would be unacceptable. It would be preferable to use the last 5 characters and have a local identifier of "12345".

Generating image identifiers in this way is consistent with a requirement for storing the images. Currently, all of the images from a particular provider are stored in the same folder, which is named using the namespace of the provider. The same rules that prevent collisions between local identifiers also ensure that two files in the same folder will not have the same filename.

When a set of images is imported into the Bioimages collection, the image import tool checks to make sure that the identifiers that will be generated do not collide with identifiers for any resources already in the database.

Organism identifiers
An image provider has two choices: manage its own identifiers for organisms or let the image import tool assign unique identifiers automatically. By default an image provider will have an organism namespace that is the provider's image namespace appended to "org-" (e.g. if the provider's namespace is "patel", the provider's organism namespace will be "org-patel". Under some circumstances, an image provider may track organisms using a system that requires one or more namespaces different from the default. In that case, the provider should contact Steve to discuss the alternative system. If the provider uses its own system for assigning organism local identifiers, it must make sure that its local identifiers are unique within the provider's namespace. If the provider doesn't care about generating and tracking its own organism identifiers, it can use the default namespace and let the software assign local identifiers.

When a new image is imported into the database, it can either be assigned to a new organism whose identifier is generated at that time, or it can be assigned to an organism that already exists in the database (e.g. if the organism is photographed again at a later time).

Agent identifiers
Each agent playing the role of photographer, determiner, or rights owner in Bioimages is assigned a locally unique identifier or code that is locally unique within Bioimages. Frequently this code is related to a person's last name (e.g. "andersonwb"), although it can also be an ORCID ID (e.g. " 0000-0002-5562-6882"). Image providers are associated with the namespaces they manage (see below). In many cases, the identifier value for "owner", "photographerCode", "identifiedBy", and the namespace will all be the same.

Each agent has a record that contains basic information about the agent, including name and contact information. The agent identifier is used in organism and image records to link to the agent's information record. Bioimages may also create a globally unique identifier for the agent, but there is not a systematic way that this is done. If the agent already has a globally unique identifier (such as the HTTP version of an ORCID ID), it is used rather than generating a new one.

Agent local identifiers are also tracked for organizations that manage collections. If there is a standard code for that collection, that code is used as the local identifier for the agent that manages the collection. Otherwise, a local identifier is created for the agent. This local identifier is used in the case where organisms that are imaged are an accessioned part of a collection (i.e. are part of an arboretum, botanical garden or zoo). The connection between the organism and the agent is made by using the agent's local identifier as the value of the collection code for the organism. Organisms with an assigned collection code are assumed to be living specimens, while those with no collection code are assumed to not be living specimens.   Such organisms typically also have a catalog number that serves to identify them locally within that collection. The collection code and catalog number assigned to a living specimen are independent of the namespace and local identifier used to create the identifier for the organism. The collection code may be the same as the organism namespace and the catalog number may be the same as the organism locally unique identifier, but this is not required.

General behavior of the Image Import Tool

If multiple thumbnails are selected, changes made to **image-related** properties will be applied to **all** images whose thumbnails are selected. For **organism-** and **determination- related** properties, changes to the values will be applied if **any** thumbnail depicting the organism is selected.

Workflow

1. Trim/modify image names using Bulk Rename Utility, NameWiz, or some other software. Useless prefixes like "IMG_" or "DSC_" are usually removed. It is preferable for all characters (including ".jpg") to all be in lowercase. Note: a certain number of the rightmost characters of the non-file extension part of the filename will be the local identifier part of the globally unique HTTP IRI identifier. Therefore, the image provider must ensure that those characters in the filenames are unique within their namespace. To prevent potential problems with future server resolution of the IRIs, the first letter of the local identifier must be alphanumeric. If the first character is underscore or dash, it will be removed by the software [this will actually be implemented in a future release of the software; in any case, don't do it].

2. Cull and rotate images using Windows Picture and Fax Viewer or some other utility. Note: some cameras like the iPhone just set a rotation bit in the metadata rather than actually rotating the images. This causes them to be displayed unreliably. This happens with vertical (portrait) images. There is a Python script at https://github.com/baskaufs/msc/blob/master/python/rotate_image_and_unset_orientation.py that

will rotate, unset the bit, and save the modified images in a new folder. It should be used on any phone images that are in portrait orientation.

Also, opening the image in Gimp will result in a dialog that asks whether GIMP should rotate it into the standard orientation. Allow it to do that, then Export and overwrite the existing images.

3. Agents management. Before images can be processed, unique (within Bioimages) identifiers must exist for all persons falling into the roles of photographer, rights owner, and determiner. If one of these identifiers does not already exist, a new one must be created. In addition, if images depict organisms that are an accessioned part of a collection a unique identifier must be assigned to the agent managing the collection. Typically, this would be a standard collection code if one exists. The agent identifier should be carefully chosen and consistently used. It should consist only of lower case letters, numbers, dashes, or underscores (NO spaces). It is best to discuss your choice of agent identifier with Steve before setting it. I think the software currently suggests a local identifier when agents records are created, but I think it can be overridden.

Launch the Bioimages Collections Manager software. If there are database updates available, a notification that says "Metadata updates are available. Click here to update." Click on this and wait for a while. This will sometimes bring up a conflicting record screen - usually accepting the suggestions by clicking "Finished" will be fine.

Click "Process New Images". Then in first screen, select the appropriate values for the agents falling into the various roles. The value given for "Your unique identifier" sets the default agent for the various roles that agents can play. The namespace corresponds to the agent managing the images and the photographer is the person to be credited for the images. In many cases these will all be the same.

4. Select images by clicking on the "Select image files" button. The correct namespace/photographer must be set when a particular set of images is loaded, but then other images can be added for a different namespace. For a typical image contributor, this won't be an issue since the namespace will usually be set to be the contributor's unique identifier in the first screen, and left at that. In the case where there are multiple photographers, set the photographer/namespace and select the images for that photographer. Then change the photographer/namespace and select the images for the next photographer. This can be repeated as often as necessary.

Similarly, if the images to be processed have different numbers of trailing characters to be used as the local identifier, set the number of trailing characters to be used, then select the images. You can then change the number of trailing characters to be used and select more images.
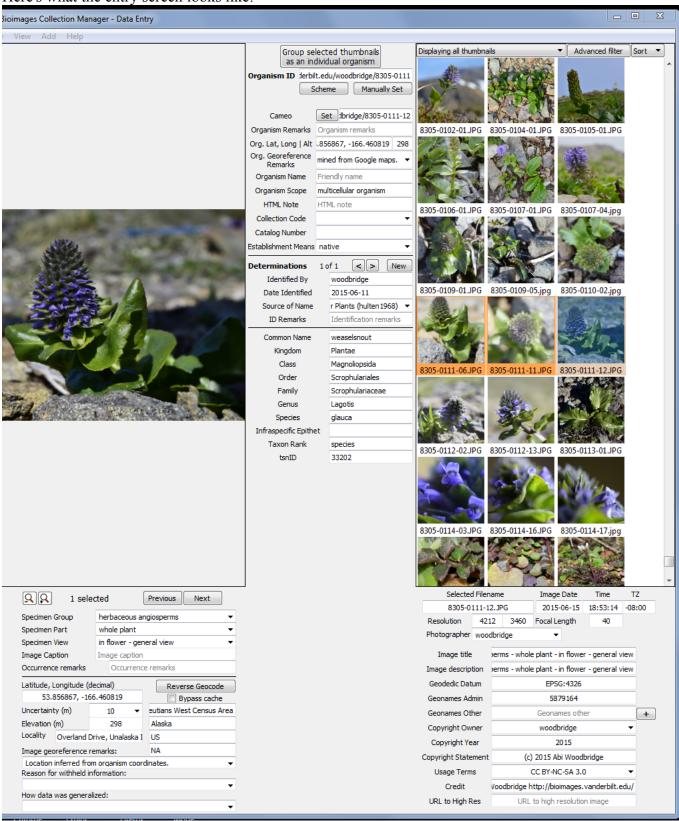
NOTE: if you accidentally include images that you don't intend to process, you can right-click on them in the Data Entry window and delete them. You can also clear the window by clicking on the "Clear all" button.

5. The first step in image processing is to group all of the images that depict the same organism. See the notes about what constitutes an "organism" in the background information above. If you
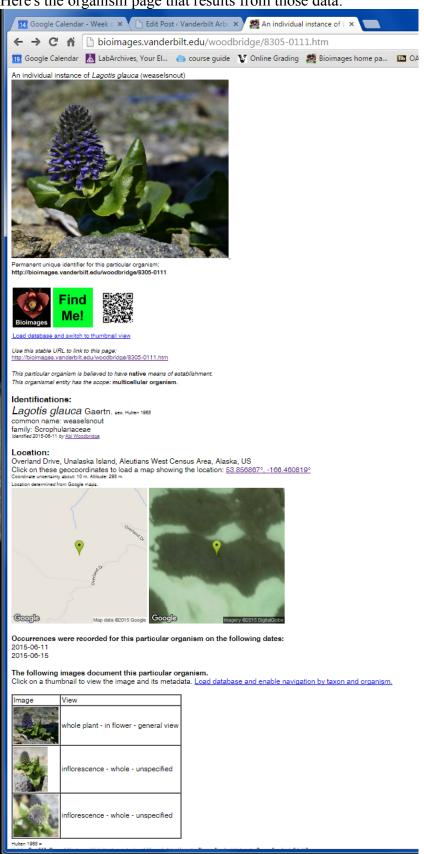
can't remember whether images depict the same biological organism or not, there are two courses of action. If they are from the same local population and you are sure that all of the images represent the same taxon, you can assign them to the same organism and set the value of Organism Scope to "colony" (for plants) or "pack" (for animals). (Yes, it is somewhat silly terminology to call a school of fish a "pack" or a grove of trees a "colony" but these are the values we have at the moment.) The other alternative is to assign every image to a separate organism and set the value of Organism Scope for each one to "multicellular organism".

Highlight all of the thumbnails that belong to the same organism. If you are letting the software use the default "org-"+ your agent code for the namespace and deciding for itself what to use for the local identifier, you can just hit the "Group selected thumbnails as an individual organism" button. If you already are using a different organism namespace (e.g. "ind-kirchoff" rather than "org-kirchoff"), you can change the default namespace to something else using the "Scheme" button. If you want to use an organism namespace other than the default or a previously assigned namespace, please discuss this with Steve. To set the local identifier to a specific value instead of letting the software decide what to use, click on the "Manually Set" button. This will be the case any time you are associating an image with an organism that already has a record.

Here's what the entry screen looks like:

Here's the organism page that results from those data:

Here's the image page that results from those data:

6. Geolocation is a critical feature of Bioimages. All images must be geolocated (associated with a decimal latitude and longitude, a.k.a. geocoordinates). However, there is no required level of precision, so images with uncertain location can be given a large uncertainty. Geolocating images is typically accomplished in one of two ways:
- by making use of known geocoordinates for the organism.
- by making use of GPS information embedded in the EXIF data for the images.
Technically these are not the same thing, since an organism may be photographed from a considerable distance from its actual location. There is also the problem that animals can move around and therefore won't necessarily have a single location. However, for the sake of simplicity, it is convenient to assume that images of an organism are taken at the location of the organism itself. The software accomplishes these two ways of geolocating as follows:

- **Make use of the geocoordinates of the organism and then assign those coordinates to all images that depict it.** The organism geocoordinates may have been collected using a handheld GPS unit, looked up on Google Maps, assigned based on the centroid of the political subdivision the organism was known to have been located in, taken from a database of known locations (e.g. the Vanderbilt Arboretum tree database), etc. In this method, the dropdown value for "Org. Georeference Remarks" is set to an appropriate value that is NOT "Location calculated as average of its images' coordinates", then the geocoordinates are pasted into the Org. Lat, Long box. (Note: the selection must be made first. If it is set on "Location calculated as average of its images' coordinates." you can't enter or paste a value.) While all of the images that depict the organism are selected, select the value "Location inferred from organism coordinates." from the "Image georeference remarks" dropdown. This will fill assign the organism's value to all of the images. While the images are still selected, you should also set the Uncertainty in meters to an appropriate value if it differs from the default.

- **Make use of GPS information embedded in the EXIF of the images**. These geocoordinates might have been recorded with the images by the camera when they were taken, or added to the EXIF later using software. In this method, the values of "Latitude, Longitude" already exist for the images and should appear in the box when a single image is selected. While the thumbnails associated with a particular organism are highlighted, set the value of uncertainty for all of the images in the dropdown below the image "Latitude, Longitude" values. Then in the organism section, dropdown the "Org. Georeference Remarks" list and select "Location calculated as average of its images' coordinates." The value shown in the "Org. Lat, Long" box should reflect this average.

- **Make use of separate geocoordinates for both the organism and the images.** It is possible that both forms of data are available (e.g. a tree's location is accurately determined in a GIS database but the image geocoordinates were also recorded automatically when the images were taken). In this case, be sure that the Georeference Remarks for neither the organism nor the image are set to derive the geocoordinates from the other.

7. The remaining values for organism properties can be set as long as any thumbnail depicting the organism is selected. The Cameo is a characteristic image for the organism that will be shown in large size on the organism's web page. If you don't like the default chosen by the

software, click on a different thumbnail and click the "Set" button. All organisms should have values for Establishment Means and Organism Scope; other fields are optional. Collection Code should only be set if the organism is managed as part of a collection (i.e. is an accessioned part of a permanent collection). In that case, the code for the agent representing the organization managing the collection should be selected from the list, and the Catalog Number assigned by the organization should be entered. Optional HTML Notes are displayed in a "Notes:" section on the organism's web page. They should consist of valid HTML tags enclosed in a <div> element. Because of the way the data are transferred for the generation of the website, all of the HTML should be on a single line without linefeeds.

8. At this point, all images associated with the organism you are working on should have geocoordinates. With all of those images selected, you can use reverse geocoding to fill in several of the location-related fields by clicking on the Reverse Geocode button. (Additional notes on this at the end.) Examine the Locality description that was given to the location. Change it if you don't like it. All images should have a value for "Latitude, Longitude", uncertainty, county (i.e. second-order administrative division), state (i.e. first-order administrative division), country code, continent code, locality, and "Image georeference remarks".

"Reason for withheld information" and "How data was generalized" are optional and should be used when the geocoordinates were made less precise to protect organisms or habitat. **Note** that the software does NOT reduce the precision automatically. If the geocoordinates are embedded in the EXIF, they will remain there throughout the image ingestion process and will probably persist in the images that are posted on the Web. So geocoordinates automatically embedded in the EXIF should be changed using software before the images are submitted to Bioimages.

9. There are a number of additional properties that that are associated with the image and are found on the lower right part of the screen. Most of these are autofilled by the software and won't need to be changed unless you have one of the less typical situations, like different photographers or rights owners.

Note: if the image is in the public domain, you should select "none" as the Copyright Owner. The copyright year will be ignored, so the value is unimportant. There may be no value. The copyright statement should be set to "Public Domain". The Usage Terms should be set to "CC0 1.0". There will still probably be a credit line indicating the acknowledgement of the source of the image.

A critical value that must be set manually is the timezone offset. For whatever reason, camera manufacturers almost never provide this information when they embed the time that the image was photographed. The value is the number of hours and minutes to be added or subtracted from GMT. The value should begin with a + or - sign and the hours and minutes in the form hh:mm. Typical values are:

timezone (U.S./Canada)	standard time	daylight savings
Eastern	-05:00	-04:00
Central	-06:00	-05:00

Mountain	-07:00	-06:00
Pacific	-08:00	-07:00
Alaska	-09:00	-08:00
Hawaii/Aleutian	-10:00	-09:00 (Aleutian only)

The "Image title" and "Image description" values are set automatically by the software any time that a Specimen View is changed or a Taxonomic Determination is added. These fields can be changed manually, but they will be overwritten if the view or determination is changed. So custom values should only be entered after all other information has been entered.

The Geonames Admin is automatically set during the Reverse Geocoding process to the Geonames ID for the lowest level political subdivision. The optional Geonames Other value can be used if the location is within some other kind of boundary, such as a park, university, nature reserve. These values can be looked up on the Geonames website and added to the "favorites" list maintained by the software. It is possible to add Geonames features if they aren't already present.

The Copyright Year defaults to the current year, but should be changed if the image was previously published.

Usage terms can be changed from the default value to any value you prefer that is no more restrictive than CC BY-NC-SA.

The URL to High Res image will usually be blank until the original image has been archived somewhere like Morphbank.

10. In order for an organism's images to be published, at least one determination must be assigned to the organism. This is done by clicking on the New determination button. The value of "Identified by" will automatically be populated with the agent code set in the opening screen. If someone else made the identification, the dropdown should be used to change the agent code. The determination date defaults to today's date and should be replaced by the actual date the organism was identified (if known). The taxonomic hierarchy is automatically filled in when you select a name using one of the search boxes. **Note: once the "Done" button is clicked, the determination cannot be deleted nor can any of the values entered so far be changed.** So get it right or cancel. It is possible that the name you want isn't in the database. If so, click on the "Manually enter" button and enter the information. As the instructions say, the ITIS Taxonomic Serial Number (tsnID) is particularly important, so you'll probably need to go to the ITIS website (http://www.itis.gov/) to get it and that's the easiest place to get the rest of the higher taxonomy needed to fill in the box. Copy and paste is good to avoid typos. For the Species entry, enter only the specific epithet, not the whole species name (e.g. "sapiens", not "Homo sapiens"). For the Scientific Name Authorship of animal species, put the author name followed by the year with no comma between. **NOTE: the ITIS website capitalizes the taxon rank (e.g. "Species"). It should be entered as lower case (e.g. "species").** To get the ubioID, go to ubio.org and search. [Note on 2017-01-11: uBio appears to be dead. However, http://resolver.globalnames.org/ can be used to find the name, then look at the XML of the results to get the uBio identifier.]

Identification remarks are optional.

The Source of Name should be included whenever possible. This is the "sensu" or "secundum" reference that links to a publication that describes how the taxon with which the name is associated was circumscribed. Preferably the value for "Source of Name" will be some sort of taxonomic treatment (e.g. Flora of North America, a regional flora, a published taxonomic treatment, etc.). At a minimum it should be the guidebook that was referenced when the identification was made. If a (human) taxonomist made the determination for your, press them to tell you the taxon authority they go by when they make determinations. Hopefully they will know what you are talking about. If a determination has no value, it will be listed as "nominal" (i.e. we have no idea what the name "means" in terms of the taxon it identifies). Some name sources are included in the dropdown. If you want to add a source, click on the "+" button and add the information requested. The TCS signature is formed based on rules described in the TDWG Taxon Concept Schema standard. For one or two authors, the author last names are given with no "and", e.g. "Jones" and "Smith Jones". For three or more authors, the names of the first two authors are given, followed by "et al.", e.g. "Radford Ahles et al.". For Internet sources, the domain name is given, e.g. "fna.org". This is followed by the date of publication, e.g. "Jones 1879", "Smith Jones 2008", and "fna.org 1993". The Unique Identifier value should be chosen to be locally different from others in Bioimages - usually it is the last name of the first author (all lower case) followed by the year with no space, e.g. "jones1879", "smith2008", "fna1993". If the publication has a DOI, it should be used as the IRI in its HTTP form, e.g. "http://dx.doi.org/10.1016/j.ympev.2013.06.010". If the publication is a book with an ISBN, form a URN by appending the ISBN with no spaces or dashes to "urn:isbn:", e.g. "urn:isbn:0691121443".

11. The last step is to set the view for each image. The Group, Part, and View values can be set in bulk for all images to which they apply. For example, if all of the images are woody angiosperms, you can select all of the thumbnails and select the "woody angiosperms" value from the "Specimen Group" dropdown. Then select only thumbnails of bark and set the value of "Specimen Part" to "bark", etc. If you are working with a lot of images, it is easy to forget which ones you've finished assigning views to. There is a thumbnail screen dropdown above the images that you can use to show only the images that don't yet have Specimen views assigned.

12. The optional "Image Caption" field is used to enter text that should be displayed below the image (i.e. with the image). In contrast, the Image description is usually displayed instead of the image (e.g. if the image doesn't load).

Occurrence remarks are not common. They are used to indicate specific information about the occurrence of the organism at that particular place and time (e.g. "The animal was released after being photographed."). As such, it isn't really a property of a particular image. Bioimages considers occurrences to be on the scale of a day, so the value of Occurrence remarks should be assigned to all of the images of a particular organism that were taken on a particular day.

13. **Exporting data**. Within the image import tool software, the image database is managed using a SQL implementation. However, all of the information about images and organisms can

be encoded in several simple CSV text files. After images are fully processed, the files can be exported by going to the Manage CSVs option and selecting "Export database to CSV files". All changes made since the last publication of Bioimages will be exported.   [Note to Steve: when exporting the entire changed database, uncheck the "Only export records changed locally" box.] You will be prompted for a location to save the files. The names of the exported files are "images.csv", "organisms.csv", and "determinations.csv". All of the generated files should be sent to Steve along with the original highres versions of the image that are described in the CSV files. You can discuss with him whether it's better to use a cloud-based system (Dropbox, Box, Google Drive, etc.) or mail a flashdrive with the files.

Additional notes on Geocoding from Ken:

When you click the [Reverse Geocode] button:

First, it takes coordinates from the [Latitude, Longitude] box and sees if those exact coordinates have been reverse geocoded before. If so, it uses the existing location names for those coordinates.

If the coordinates haven't been reverse geocoded before it sends a query to a web service for the location names. It saves the location names to the appropriate image records as well as to a local cache of reverse geocoded coordinates (this cache is what's checked in the previous step). If something went wrong with the web service it is possible the location information may be saved as blank fields in the cache. This is one reason why you would check the "Bypass cache" box, to get updated values for the location.

If for some reason the reverse geocoding web service is temporarily down (or your own internet connection), this is one of the reasons why it is a good habit to use the thumbnail filter "Thumbnails not reverse geocoded" before submitting to you (or you publishing to the web). It will show all the thumbnails that haven't been reverse geocoded (e.g. the ones that returned blank location names), even if you thought you did.