Digital Herbarium Documentation

Release

Dmitry E. Kislov

Dec 11, 2017

Оглавление

Sections "Digital Herbarium: Basic Usage" and "Digital Herbarium's HTTP-API Description" are devoted to peculiarities of user interaction with the Digital Herbarium Database using web-interface and programming languages respectively.

 Printable version

Русскоязычная версия

DIGITAL HERBARIUM: BASIC USAGE

1.1 Features

Accessing to the Digital Herbarium's data is provided via the web-page on the official website of the Botanical Garden Institute or HTTP API service. The latter approach is used for making search queries automatically (from R, Python or other computational environment).

Main features of the search service:

- search in a given time interval either by date of collection or date of identification fields;

- accounting species synonyms when searching;

- search in a given rectangular region;

- search within additional species (only for multispecies herbarium records);

- search by record codes (e.g. field number, inventory number etc.);

- search by the country of origin;

- search by taxonomic name, e.g. family, genus or species epithet;

1.2 Search menu

General search possibilities are available via the search menu from the Digital Herbarium's web-page (Fig. 1)

When search conditions are given simultaneously, the service is trying to perform an AND-type query; it retrieves records satisfying all search conditions. Currently only AND-type search queries are available. To perform OR-type queries it is recommended to use the HTTP API service.

Values of Family, Genus and Country search fields can be selected via drop-down menu that rises when typing.

Start date of collection and end date of collection are filled out from pop-up calendar when the mouse is hovering these fields.

If only start date of collection is given, the service retrieves records having later dates in the corresponding field.

If only end date of collection is given, the service retrieves records having earlier dates in the corresponding field.

If start date of collection and end date of collection are given, the service retrieves records if its corresponding date interval intersects the given.

Рис. 1.1: Fig. 1. Basic search menu

Regarding the following text fields — Species epithet, Code, Collectors, Identifiers, Place of collection the condition satisfaction assumes containing the given value as a sub-string in the corresponding field (case insensitive comparison is performed).

If one performs search in the Collectors or Identifiers fields and fills these fields with Cyrillic letters, the service will automatically transliterate the given value into English (Latin letters) and return records satisfying both Cyrillic and transliterated values. If one provide the value only in Latin letters, no transliteration will be performed. Therefore, If you try, for example, to find records including "bakalin" as a sub-string in the field Collectors, the search will return the records which field Collectors (internally Collectedby field) includes the string "bakalin" (reverse transliteration (to Cyrillic letters) in this case wouldn't be performed); If you try to search "бакалин" (Cyrillic equivalent of 'bakalin') combined search results for both "bakalin" and "бакалин" queries will be returned.

Boolean fields Search within synonyms and Search within additional species indicate that, in the first case — the search engine will take into account known (to the system) table of species synonyms, and in the second — the search engine do searching within additional species if those are provided.

> Warning: When do searching within species synonyms, the search engine uses the table of species synonyms that, in turn, is being dynamically rebuilt each time records in the Table of known species are updated. The Table of known species can include errors. This could lead to surprising search results. These types of drawbacks (caused by incorrectness of species synonym relationships) tend to disappear in future, as the Table of known species will become more error-less.

Note: Search within synonyms works in cases when the exact names of the pair (genus, species epithet) are given.

1.2.1 Search by Code field

Herbarium records stored in Digital Herbarium of the BGI use triple coding system. Each record is provided with 1) inventory number (optional), used in the Herbarium's storage; 2) mandatory ID field (unique, digits only), assigned by the system automatically; 3) field number (code), assigned by the collector (it is optional and quite arbitrary);

Therefore, the table of search results includes the column Complex code, which accumulates codes of these three types.

Complex code has the following structure:

Note: Inventory number (if provided) or * symbol/ID code/Field code (if provided)

So, the Complex code values can look as follows:

- */27031/M.I.38 — denotes that the inventory number isn't provided, ID = 27031, and field code is M.I.38;

- 42/27029 — denotes that the inventory number is 47, ID = 27029, field code isn't provided;

- the following form of the code can take place as well: 132123/32032/F-3829-3k, where inventory number is 132123, ID is 32032 and field code is F-3829-3k (fake example);

When do searching by Code one should provide either an inventory number, ID or field code. For example, if the search field's value is "231" the search engine will return records including "231" as a sub-string in either the inventory number, ID or field code.

1.3 Filtering search results

Standard filtering interface allows to restrict results of searching by Herbarium's acronym, Herbarium's subdivision or select desired number of items showed per page Fig. 2.
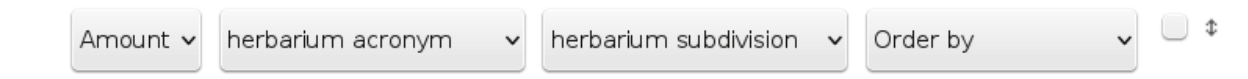

Рис. 1.2: Fig. 2. Search filtering menu

It has the following fields:

- Amount — the number of records showed per page;

- Herbarium acronym — filtering by Herbarium's acronym;

- Herbarium subdivision — filtering by Herbarium's subdivision;

- Order by — ordering rule (choose field you want to perform ordering the results);

Results of search request with applied filter (records only under VBGI-acronym are shown) are presented on the Fig. 3.

In the tab Common Info a table with the records satisfying current search and filtering conditions is shown (if no searching/filtering conditions were provided all published records are shown, by default the number-per-page is equal to 20).

The Details tab activates when a specific Herbarium's record is clicked. It shows minified version of the Personal web-page of the record.

The Map tab is a copy of Common Info tab excluding records with no coordinates (records with coordinates are rendered on the Google map as clickable markers).

One can click Previous or Next (switch page) to get another portion of search results.

The Automatization tools tab includes general information on performing queries using automatization possibilies provided by the web-application.

Working with the map, one can filter search results by user-defined rectangular area. To do that, just initialize a rectangular area by pressing [🔍], edit the appeared rectangular region, and press [🔍] again to activate the search (See Fig. 4, Fig. 5).

To clear particular search condition click small-trash icon near the corresponding search field.

To clear all search conditions press the [Clear] button.

Search within polygonal regions isn't supported by current database backend, but such behavior could be emulated programmatically with the help of the HTTP API Service.

Search results:

Page 1 of 606 Next   Total: 12116

Sheet's code (combined)	Species	Collection date	Collector(s)	Identifier(s)
*/27285	Camptosorus sibiricus Rupr.	2004-08-10	Kreshchenok I.A.	Kreshchenok I.A.
*/27284	Camptosorus sibiricus Rupr.	2004-08-15	Kreshchenok I.A.	Kreshchenok I.A.
*/27283	Camptosorus sibiricus Rupr.	2002-06-20	Vorob'eva A.N.	Vorob'eva A.N.
*/27282	Camptosorus sibiricus Rupr.	2015-06-24	Veklich T.N.	Veklich T.N.
*/27281	Camptosorus sibiricus Rupr.	2012-07-01	Veklich T.N., Darman G.F.	Veklich T.N., Darman G.F.
*/27280/16	Asplenium tenuicaule Hayata	2005-09-10	Koldaeva M.N.	Kreshchenok I.A.
*/27279	Asplenium ruta-muraria L.	2005-09-03	Kreshchenok I.A.	Kreshchenok I.A.
*/27278	Asplenium ruta-muraria L.	2007-07-31	Kreshchenok I.A., Rubtsova T.A., Fetisov D.M., Zaitseva N.V.	Kreshchenok I.A.
*/27277	Aleuritopteris kuhnii (Milde) Ching	2005-07-10	Kreshchenok I.A., Starchenko V.M., Darman G.F., Borisova I.G.	Kreshchenok I.A.
*/27276	Aleuritopteris argentea (S.G.Gmel.) Fee	2004-09-03	Kreshchenok I.A.	Kreshchenok I.A.

Рис. 1.3: Fig. 3. Search results tab

Page 1 of 0

Lat, max
65.089

Lon, min
112.81

Lon, max
132.89

Lat, min
50.133

Total: 0

On the map 0 record(s) showed*

Map Satellite

Google   Map data ©2017 INEGI Imagery ©2017 NASA, TerraMetrics | Terms of Use

*Only records with coordinates are listed
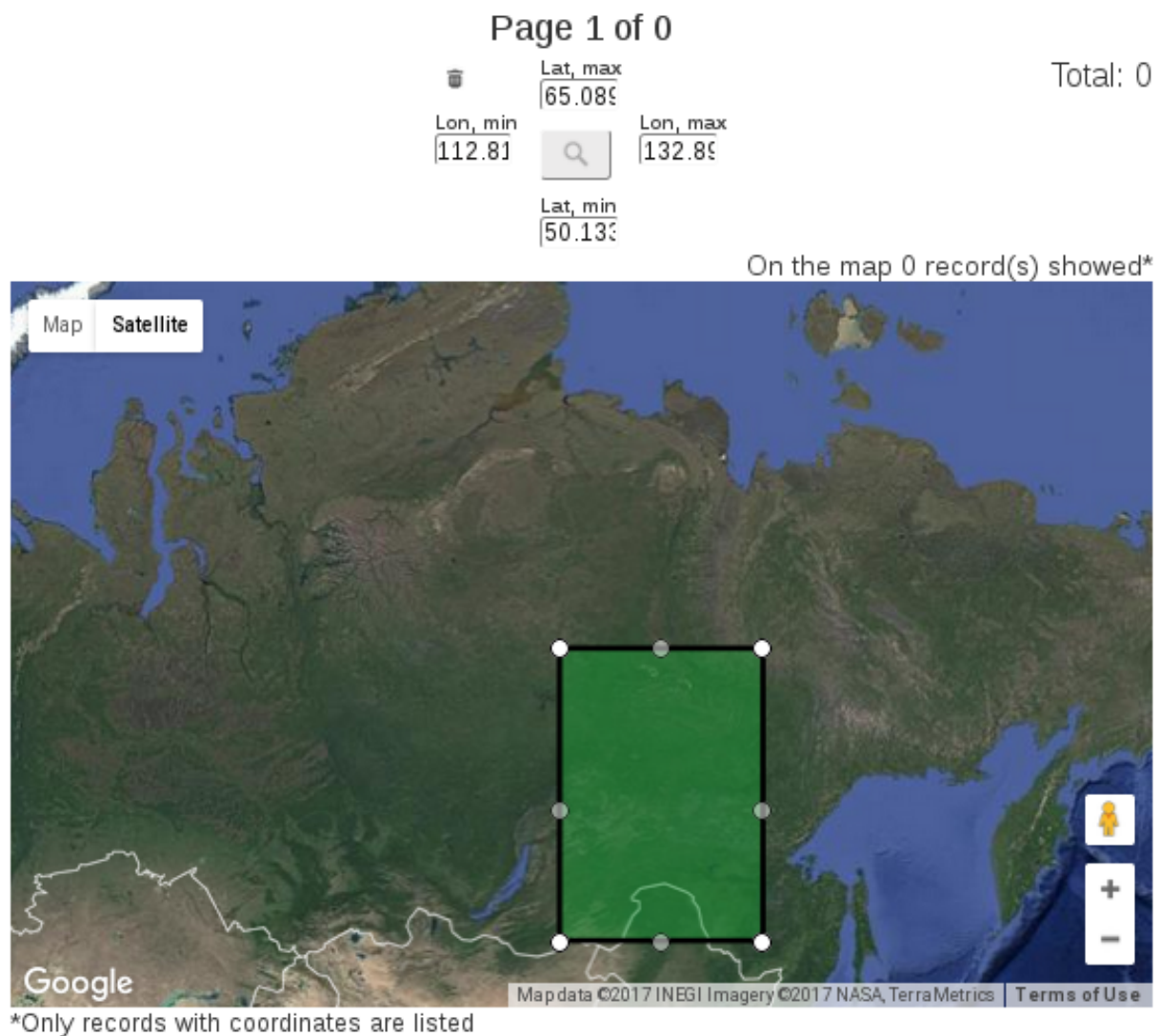
Рис. 1.4: Fig. 4. Initialize filtering region

Page 1 of 20 Next

Lat, max
65.089

Total: 3919

Lon, min
112.81

🔍

Lon, max
132.89

Lat, min
50.133

On the map 200 record(s) showed*

Map Satellite

Google

Map data ©2017 INEGI Imagery ©2017 NASA, TerraMetrics | Terms of Use

*Only records with coordinates are listed
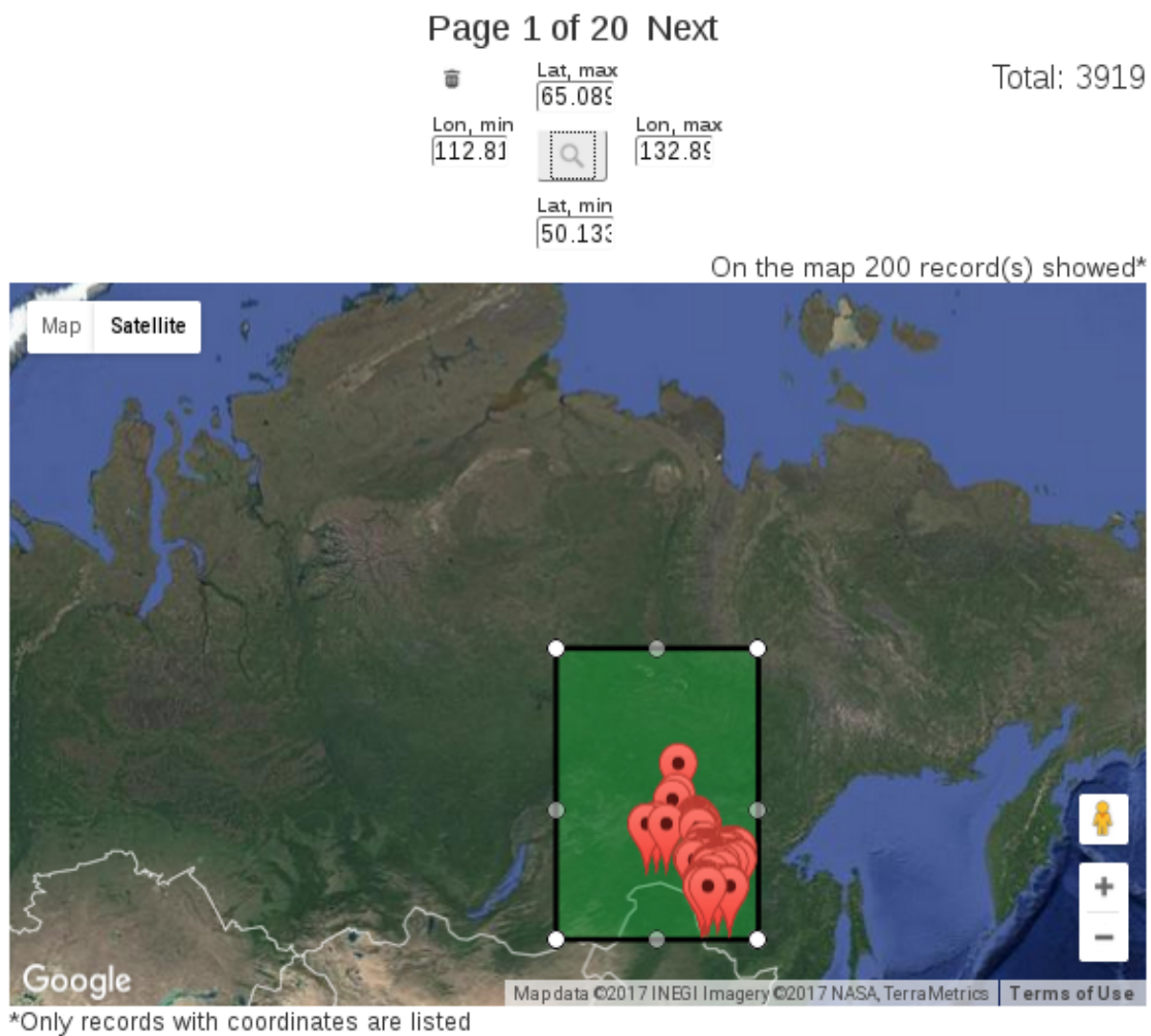
Рис. 1.5: Fig. 5. Getting results of geographical filtering/searching

DIGITAL HERBARIUM'S HTTP-API DESCRIPTION

2.1 Intro

This document describes HTTP-API (Application Programming Interface over HTTP protocol) which can be used to get access to Digital Herbarium Database of the BGI.

HTTP-API works in read-only mode. There is no way to make changes in the database using the API.

2.2 Description of HTTP request parameters

Only GET-requests are allowed when reffering to the HTTP API service. To establish connection with the service, one can use HTTP or HTTPS protocols.

Requests with multiple parameters, e.g. colstart=2016-01-01 and collectedby=bak, are treated as components of AND-type queries: in this example, all records collected after 2016-01-01 and including bak (case insensitive matching is performed) as a sub-string of Collectors field will be returned.

OR-type querying behavior can be emulated by a series of consequent queries to the database and isn't natively implemented in the current version of the HTTP API.

List of allowed GET-parameters:

- family — family name (matching condition: case insensitive, the same family name as provided);
- genus — genus name (matching condition: case insensitive, the same genus name as provided), note: if the value contradicts with the family name provided in the same request, an error will be returned;

- species_epithet — species epithet (matching condition: case insensitive, a sub-string of the record corresponding field);

- place — place of collection (matching condition: case insensitive, a sub-string occurring in one of the listed fields: Place, Region, District, Note;);

- collectedby — collectors (matching condition: case insensitive, a sub-string of the record corresponding field); if the field's value is given in Cyrillic, search will be performed (additionally) using its transliterated copy;

- identifiedby — identifiers; (matching condition: case insensitive, a sub-string of the record corresponding field); if the field's value is given in Cyrillic, search will be performed (additionally) using its transliterated copy;

- country — country's name (matching condition: case insensitive, a sub-string of the record corresponding field);

- colstart — date when herbarium sample collection was started (yyyy-mm-dd);

- colend — date when herbarium sample collection was finished (yyyy-mm-dd);

- acronym — acronym of the herbarium (matching condition: case insensitive, the same name as provided);

- subdivision — subdivision of the herbarium (matching condition: case insensitive, the same name as provided);

- latl — latitude lower bound, should be in (-90, 90);

- latu — latitude upper bound, should be in (-90, 90);

- lonl — longitude lower bound, should be in (-180, 180);

- lonu — longitude upper bound, should be in (-180, 180);

- synonyms — Boolean parameter, allowed values are false or true; absence of the parameter in GET-request is treated as its false value; true value (e.g. synonyms=true) tells the system to search records taking into account the table of species synonyms; Note: when performing search including known (known by the system) species synonyms one should provide both genus and species_epithet values, if only one of them is provided or both are leaved empty, a warning will be shown and the search condition will be ignored;

- additionals — Boolean parameter, allowed values are false or true; absence of the parameter in GET-request is treated as its false value; true value (e.g. additionals=true) tells the system to search within additional species (if such is provided); some herbarium records could include more than one species (such records are referred as multispecies records);

- id — record's ID (matching condition: the same value as provided); if this parameter is provided in GET-request, all other search parameters are ignored and the only one record with the requested ID is returned (if it exists and is published);

- fieldid — field code/number; (matching condition: case insensitive, a sub-string of the record corresponding field);

- itemcode — storage number (matching condition: case insensitive, a sub-string of the record corresponding field);

- authorship — authorship of the main species (matching condition: case insensitive, a sub-string of the record corresponding field);

- imonly — allowed values are false or true; absence of the parameter in GET-request is treateda as its false value; when filtering with imonly=true records having images will be shown only.

Note: The search engine performs only one-way transliteration of collectedby and identifiedby fields into English language. So, if you try to search, e.g. collectedby=боб (that corresponds to bob in English), the system will find records including (in the collectedby field) both боб and bob sub-strings. On the contrary, If you try to send collectedby=bob search query, only records that include bob will be found (regardless the text case).

Warning: Transliteration from Cyrillic (Russian) to Latin (English) is fully automatic and could be quite straightforward, e.g. Джон will be transliterated into something like Dzhon, instead of John, as it would expected.

2.3 Description of server response

The server response is a JSON-formatted text transferred via HTTP-protocol and having the following attributes:

- errors — array of errors (each error is a string) occurred during search request processing;

- warnings — array of warnings (each warning is a string) occurred during search request processing;

- data — array of structured data, i.e. result of the search query.

Note: Warnings are informative messages that are intended to tell the user what went in an unexpected way during interaction with the database: e.g. which search parameters contradict each other, which parameters were ignored, which parameters weren't recognized by the system etc.

2.3.1 Format of the data attributes

The data attribute is a JSON-formatted array. Each item of this array describes a herbarium record and has the following attributes:

- family — family name (Latin uppercase letters);

- family_authorship — self explanatory parameter;

- genus — genus name;

- genus_authorship — self explanatory parameter;

- species_epithet — self explanatory parameter;

- species_id — ID of the species-level taxon (unique integer value); don't mix with ID of the herbarium record. ID of the herbarium record is unique among all herbarium records, ID of the species-level taxon is unique among all species-level taxa;

- infraspecific_rank — allowed values: subsp., subvar., f., subf., var. or null (i.e. left blank);

- infraspecific_epithet — self explanatory parameter;

- infraspecific_authorship — self explanatory parameter;

- short_note — used in multispecies herbarium records; the field provides important information about the main species of the herbarium record (it could be empty);

- species_authorship — self explanatory parameter;

- species_status — current species status; the term "species status" is related to species-level taxon not herbarium record; it describes a degree of acceptance of species by scientific community (current state); possible values of species_status are 'Recently added' — the species was recently included to the database and wasn't checked by an expert, 'Approved' — the species was approved by an expert (a user having some privileges), 'Deleted' — the species name is probably obsolete and should be avoided, 'From plantlist' — the species was imported from the http://theplantlist.org;

- type_status — type status of the collection;

- species_fullname — full species name, e.g. Genus + species epithet + species authorship;

- significance — measure of ambiguity regarding the main species (possible values: "", aff., cf.);

- id — integer identifier of a herbarium record, it is unique;

- gpsbased — Boolean parameter, its true value means that a herbarium record position is obtained via the GNSS (GPS/GLONASS);

- latitude — latitude, degrees (WGS84);

- longitude — longitude, degrees (WGS84);

- fieldid — field number; an arbitrary string assigned by a collector;

- itemcode — inventory (storage) number, a string assigned by the herbarium's curator; it is used to identify the place of the record in the herbarium storage;

- acronym — herbarium acronym (e.g. VBGI);

- branch — herbarium branch/subdivision (e.g. "Herbarium of Fungi", "Bryophyte Herbarium" etc.);

- collectors — collectors;

- identifiers — identifiers;

- devstage — development stage; available values: Development stage partly, Life form or empty string;

- updated — the date the record was saved/updated;

- created — the date the record was created;

- identification_started — the date the species identification was started;

- identification_finished — the date the species identification was finished;

- collection_started — the date the herbarium item was collected (first day or null if no information provided);

- collection_finished — the date the herbarium item was collected (last day or null);

- country — country name;

- country_id — unique (integer) id of the country internally assigned by the system;

- altitude — altitude (sea level is treated as zero), this parameter is a string, therefore its form of altitude's representation might be quite fuzzy: '100-300', '100-300 m', '100', '100 m' etc.; it is assumed that altitude value is given in meters;

- region — administrative region of collection;

- district — administrative district of collection;

- details — environmental conditions of collection, additional info;

- note — everything that wasn't yet included in the previous fields (this field could include information about the place of collection, details on environmental conditions etc.);

- dethistory — an array; history of species identifications for this herbarium record;

- additionals — some herbarium records could include more than one species, this array describes them;

- images — a list of images related to the herbarium record ([] – an empty list, means that no images attached to the herbarium record were found);

Note: Images from the images array are provided in several resolutions. Currently, the system stores images of different resolutions in directories named ss – small size (30% of original size); ms – medium size (60% original size); fs – full size (original size).

Therefore, each image url includes one of the following components ' /ts/ , /ss/ ', ' /ms/ ' или ' /fs/ '. These components point to resolution of the image available from this url.

Note: All images are saved as jpeg via ImageMagick image processing utilities with the following parameters:

```
'-strip', '-interlace', 'Plane',
'-sampling-factor', r'4:2:0',
'-quality',
r'90%'
```

It comes from practice that such compression don't significantly impact on images. In the save time, compression is very important and allows to save a lot of storage space.

List of images attached to the herbarium record (example):

```
['http://botsad.ru/herbarium/view/snapshots/VBGI/ss/VBGI32618_1.jpg',
 'http://botsad.ru/herbarium/view/snapshots/VBGI/ts/VBGI32618_1.jpg',
 'http://botsad.ru/herbarium/view/snapshots/VBGI/ms/VBGI32618_1.jpg',
 'http://botsad.ru/herbarium/view/snapshots/VBGI/fs/VBGI32618_1.jpg'
...
]
```

Note: Attributes region, district, details, note, altitude could be filled in bilingual mode: English first, than – Russian (or vice versa), with special symbol "|" separating two spellings (for instance, region's value"Russian Far East|Дальний Восток России"). Removing unnecessary sub-strings from the left or the right side of the "|" symbol couldn't be done in the current implementation of the API service, it should be performed by the user.

Note: Unpublished records are excluded from the search results.

Structure of dethistory and additionals arrays are described below.

History of species identifications and additional species

History of species identifications

Each item of the array "History of species identifications" (dethistory) describes an attempt of identification/confirmation of the main species related to the herbarium record.

History of species identifications (dethistory) is an array having the following fields:

- valid_from — start date of assignment validity to particular species name;
- valid_to — end date of assignment validity to particular species name; empty field means that species' name assignment is actual since the valid_from date;
- family — family name;
- family_authorship — self explanatory parameter;
- genus — genus name;
- genus_authorship — self explanatory parameter;
- species_epithet — self explanatory parameter;
- species_id — ID of the species-level taxon;
- species_authorship — self explanatory parameter;
- species_status — status of the species-level taxon;
- species_fullname — full species name (Genus name + species epithet + species authorship);
- infraspecific_rank — allowed values: subsp., subvar., f., subf., var. or null (i.e. left blank);
- infraspecific_epithet — self explanatory parameter;
- infraspecific_authorship — self explanatory parameter;
- significance — measure of ambiguity regarding the current species (possible values: "", aff., cf.);

Note: If herbarium record/sheet include more than one species, than "history of species identifications" is related to the main species of the record only.

Additional species

"Additional species" (additionals) is an array describing all the species (except the main species) attached to the current herbarium record/sheet. It is non-empty only for multispecies herbarium records. Each element of the additionals array has the following fields (fields have almost the same meaning as for dethistory array):

- valid_from — beginning date of validity of identification;
- valid_to — ending date of validity of identification; empty field means that species' name assignment to the herbarium record is actual since valid_from date;
- family — family name;
- family_authorship — self explanatory parameter;
- genus — genus name;
- genus_authorship — self explanatory parameter;
- species_epithet — self explanatory parameter;
- species_id — ID of the species-level taxon;
- species_authorship — self explanatory parameter;
- species_status — status of the species-level taxon;
- species_fullname — full species name;
- significance — measure of ambiguity regard the current species (possible values: "", aff., cf.);
- infraspecific_rank — allowed values: subsp., subvar., f., subf., var. or null (i.e. left blank);

- infraspecific_epithet — self explanatory parameter;

- infraspecific_authorship — self explanatory parameter;

- note — additional information about the current species;

Note: The note field could be filled out bilingually (e.g. using the "|" symbol); So, it behaves like described early.

Example

Let us consider an example of additionals array (not all fields are shown for short):

```
[
{'genus': 'Quercus', 'species_epithet': 'mongolica', ... ,'valid_from': '2015-05-05', 'valid_to': '2016-01-01
↪'},
{'genus': 'Quercus', 'species_epithet': 'dentata', ... ,'valid_from': '2016-01-01', 'valid_to': ''},
{'genus': 'Betula', 'species_epithet': 'manshurica', ... ,'valid_from': '2015-05-05', 'valid_to': ''},
{'genus': 'Betula', 'species_epithet': 'davurica', ... ,'valid_from': '2015-05-05', 'valid_to': ''},
]
```

Interpretation:

So, if today is 2015, 1 Sept, than the array includes Quercus mongolica, Betula manshurica and Betula davurica, but Quercus dentata should be treated as out-of-date for this date.

If today is 2017, 1 Jan, than out-of-date status should be assigned to Quercus mongolica, and, therefore, actual set of species includes Quercus dentata, Betula manshurica и Betula davurica.

2.4 Service usage limitations

Due to the long processing time needed to handle each HTTP-request, there are some restrictions on creating such (long running) keep-alive HTTP-connections (when using the HTTP API Service).

The number of allowed simultaneous connections to the service is determined by JSON_API_SIMULTANEOUS_CONN value.

When the number of simultaneous connections is exceeded, the server doesn't process search requests, but an error message is returned.

This behavior isn't related to search-by-id queries. Search-by-id queries are evaluated quickly and have no special limitations.

Attempt to get data for unpublished record by its ID leads to an error message.

2.5 Examples

To test the service, one can build a search request using web-browser (just follow the links below):

http://botsad.ru/hitem/json/?genus=riccardia&collectedby=bakalin

Following the link will lead to json-response that includes all known (and published) herbarium records of genus Riccardia collected by bakalin.

Searching by ID (colstart will be ignored):

http://botsad.ru/hitem/json?id=500&colstart=2016-01-01

---

http://botsad.ru/hitem/json?id=44

http://botsad.ru/hitem/json?id=5

See also:

Accessing Digital Herbarium using Python

Accessing Digital Herbarium using R

THREE

HERBARIUM RECORD'S CITING

Not yet completed.

C
citing, 18

M
map, 6

R
rectangular area, 6

S
search by region, 6
search form, 3
search in a region, 6
search results filtering, 6