# Exporting Data

FungiDB has various ways to export data: from already built-in bulk files, tools to generate simple fasta files, and option to download sophisticated data mining results that contain the selected user's IDs with the specific information the user chooses. This powerful ability can be initiated from within the website (eg.: in a gene page or in a search results page) or programmatically using web services. In the near future we will be adding the exporting of these data to EuPathDB Galaxy as part of the user workspace.

## Available bulk files for download

With any new website release (every two/three months) we generate various text files with genome and proteome information in FASTA, GFF and other formats for each organism and species, ready to be downloaded. To access them the user selects the option "Data files" under the Download" tab in the grey Menubar. These files contain for each organism the full genome/proteome.

## Sequence Retrieval Tool

The Sequence Retrieval Tool (SRT) allows the user to specify coordinates on the genome or proteome to be downloaded in FASTA format. SRT is accessible from the home page or from the Tools tab in the grey Menubar under the Header. It works independently from the data mining infrastructure used when running searches and building search strategies. SRT allows to export the specified sequence information in a FASTA file where the sequences correspond to various "record types" defined in a genome or proteome: genes, proteins, chromosomes etc. The User Interface (UI) is a simple form that allows the user to enter one or several IDs (such as gene IDs) and specify the requested genome coordinates relative to the ID genome location. The result can be shown in the browser or downloaded to a text file.

This service can be requested programmatically using a URL such as:
http://FungiDB.org/cgi-bin/geneSrt?project_id=FungiDB&ids=NCU06658&type=genomic&upstreamAnchor=Start&upstreamSign=minus&upstreamOffset=10&downstreamAnchor=End&downstreamSign=plus&downstreamOffset=2000

## Download a gene page

A gene page contains all the available information for a given gene ID. At the top of the page the user may click to download selected information shown in the page. Two options are offered: (1) download a FASTA file with the gene sequence (Genomic, CDS, Transcript or Protein) -or a relative sequence segment, or (2) download a text file that will include the information selected by the user. The exported data will be returned in the browser or in a text file. The text file will include first the univalue attributes such as the chromosome where the gene is located followed by multivalue attributes such as user comments, or EC numbers, these organized in a tabular format.

## Download a search/strategy result

When the user runs a search, the result table will include the IDs returned by the search and a subset of the information FungiDB has about each ID (its “attributes”). A Download link at the right top corner will send the user to select a format and the specific information (attributes) to download for each ID in the result. The exported data will be returned in the browser or in a text file. Three formats are currently offered: tabular which may be opened as an excel sheet, FASTA or GFF (when appropriate).

## Web services

Until recently the only way to download data programmatically was using the web services explained in http://fungidb.org/fungidb/serviceList.jsp. These make use of a Struts action to request the data from the server. These are always a GET operation, which means they can be expressed in a URL and therefore run on a browser immediately or from within a program. The result is a text file in xml or json format (this is specified in the URL) containing the IDs returned by a search, and the attributes requested (o-fields parameter in the URL below). For example:
http://FungiDB.org/webservices/GeneQuestions/GenesByMolecularWeight.xml?
min_molecular_weight=10000&
max_molecular_weight=50000&
reference_strains_only=Yes&
organism=Aspergillus clavatus&
o-fields=gene_type,organism

A current problem with these services for genes is the inability to download multi-value information for a gene, such as the EC numbers associated (the reason is technical and beyond the scope of this paper). Furthermore, our backend is transitioning to a more modern architecture based on a REST service backend, thus abandoning Struts. This means that we will stop supporting the previous services and instead will be offering a new set, a more powerful set of REST web services that our websites will use. These will enable the user not only download search results but also run full search strategies by combining results, analyzing results, and finally downloading results programmatically.

The new REST services are implemented as a POST to http://fungidb.org/fungidb/service/answer and the required parameters must be provided in a json object (see examples below).

Both old and new services offer currently only search results. The WADL for old web services informs for each search (1) how to construct the URL, (2) parameter names and values and (3) possible information for each gene. The new POST service informs of the parameter and attributes names but not the parameter values; this is work in progress.

While the old services offered 2 formats: xml and json, the new services can return the result in three more formats : (1) tabular,  which generates a tab delimited file to be shown in an excel sheet, (2) an adhoc format that contains first the univalue attributes, such as the chromosome where the gene is located, followed by multivalue attributes such as user comments, or EC numbers, these organized in a tabular format, and (3) fasta.

| | | | **POST to**<br>**http://fungidb.org/fungidb/service/answer** | **GET** |
|---|---|---|---|---|
| **Information to generate the request** | | | Search parameter names:<br>`http://fungidb.org/fungidb/service/question/GeneQuestions.GenesByMolecularWeight`<br>Gene information:<br>`http://fungidb.org/fungidb/service/record/GeneRecordClasses.GeneRecordClass` | WADL( by search):<br>`http://fungidb.org/fungidb/webservices/GeneQuestions/GenesByMolecularWeight.wadl` |
| **download 1 search result (set of genes or genomic sequences or any other data type returned in searches in our websites)** | INPUT | request | **`Case 1a: tabular`**<br><br>`{`<br>` "answerSpec": {`<br>`   "`**`questionName`**`": "GeneQuestions.GenesByMolecularWeight",`<br>`    "`**`parameters`**`": {`<br>`    "organism": "Aspergillus clavatus",`<br>`    "min_molecular_weight":"10000",`<br>`    "max_molecular_weight":"50000"`<br>`    },`<br>`    "filters": []`<br>` },`<br>` "formatting": {`<br>`    "formatConfig": {`<br>`      "includeHeader": true,`<br>`      "`**`attributes`**`": [`<br>`        "organism",`<br>`        "gene_type"`<br>`       ],`<br>`      "attachmentType": "plain"`<br>`    },`<br>`    "`**`format`**`": "`**`attributesTabular`**`"`<br>`  }`<br>`}`<br><br>**`Case 1b: tabular`**<br><br>`{`<br>`  "answerSpec": {` | `http://FungiDB.org/webservices/GeneQuestions/GenesByMolecularWeight.xml?`<br>`min_molecular_weight=10000&`<br>`max_molecular_weight=50000&`<br>`reference_strains_only=Yes&`<br>`organism=Aspergillus clavatus&o-fields=gene_type,organism` |

```
       "questionName":
"GeneQuestions.GenesByTaxonGene",
   "parameters": {"organism": "Candida
albicans SC5314"},
       "viewFilters": [],
       "filters": [],
       "wdk_weight": 10
  },
  "formatting": {
       "formatConfig": {
       "tables": ["GOTerms"],
       "includeEmptyTables": true,
       "attachmentType": "plain"
       },
       "format": "tableTabular"
  }
}
```

**Case 2: ad hoc**

```
{
  "answerSpec": {
       "questionName":
"GeneQuestions.GenesByTaxonGene",
   "parameters": {"organism": "Candida
albicans SC5314,Albugo candida 2VRR"},
       "viewFilters": [],
       "filters": [],
       "wdk_weight": 10
  },
  "formatting": {
       "formatConfig": {
       "tables":
["GOTerms","Sequences"],
       "attributes": ["primary_key"],
       "includeEmptyTables": true,
       "attachmentType": "plain"
       },
       "format": "fullRecord"
  }
}
```

**Case 3: FASTA**

```
{
  "answerSpec": {
       "wdk_weight": 10,
       "viewFilters": [],
       "questionName":
"GeneQuestions.GenesByExonCount",
       "filters": [{
       "name":
"matched_transcript_filter_array",
       "disabled": false,
       "value": {"values": ["Y"]}
       }],
```

| | | | | |
|---|---|---|---|---|
| | | | **"parameters"**: {<br>"num_exons_lte": "20",<br>"num_exons_gte": "2",<br>"organism": "Albugo candida,Albugo candida 2VRR",<br>"scope": "Transcript"<br>}<br>},<br>"formatting": {<br>**"format"**: **"srt"**,<br>"formatConfig": {<br>"endOffset3": 0,<br>"upstreamAnchor": "Start",<br>"attachmentType": "plain",<br>"endAnchor3": "End",<br>"upstreamOffset": 0,<br>"downstreamSign": "plus",<br>"type": "genomic",<br>//protein or processed_transcript or CDS<br>"downstreamOffset": 0,<br>"startAnchor3": "Start",<br>"startOffset3": 0,<br>"downstreamAnchor": "End",<br>"sourceIdFilter": "genesOnly",<br>"upstreamSign": "plus"<br>}<br>}<br>} | |
| | OUTPUT | tabular | "format": "attributesTabular", "tableTabular" | no |
| | | FASTA | "format": "srt" | no |
| | | xml | "format": "xml" instead of “fullRecord” | .xml in url |
| | | json | if missing ”format”, it will return json | .json in url |
| **1 gene** | INPUT | request | {<br>"answerSpec": {<br>"questionName": "__GeneRecordClasses.GeneRecordClass__singleRecordQuestion__",<br>"parameters": {<br>"primaryKeys": "NCU06658,FungiDB"<br>}<br>},<br>"formatting": {<br>"formatConfig": {<br>"attributes": "__ALL_ATTRIBUTES__", | **(limited to attributes)**<br>http://FungiDB.org/webservices/GeneQuestions/GeneBySingleLocusTag.json?single_gene_id=NCU06658&o-fields=gene_type,organism |

| | | | | |
|---|---|---|---|---|
| | | | `"tables": "__ALL_TABLES__"`<br>`}`<br>`}`<br>`}` | |
| | OUTPUT | tabular | no | no |
| | | xml | no | yes |
| | | json | yes | yes |

Caveat: because the internals of genes and transcripts, the only search that will return tables is: `GeneQuestions.GenesByTaxonGene`

Todos:

1. Provide service returning allowed parameter values for a given search.
2. Post to singleRecord Gene returns only Gene attributes and Tables.
    a. Do these include transcript information?
    b. Cleanup, internal information returned.
3. Post to a search: all parameters are required, should we allow empty for default values?
4. Offer “all” for attributes and tables
5. Why SingleLocus not available?

## Future Work

There are several areas of current development that will enable users to export the same data in different ways or to export data not currently available for download. The analysis of search results is an area of active development where downloads are only offered depending on the tool result. Soon we will offer the ability to send results to our EuPathDB Galaxy  website. Furthermore we are developing a User Workspace that will enable the user to import data from a Galaxy result into FungiDB for further analysis. Finally the development of the new REST web services will continue towards our goal to abandon Struts for a REST service backend. Search results are currently available but in the future a user will be able to run strategies programmatically.