



# D03 – DELIVERABLE 4.1.3

Project Acronym: OpenUp!  
Grant Agreement No: 270890  
Project Title: Opening up the Natural History Heritage for Europeana

---

## Data quality service for zoological names production version

D03 – Deliverable 4.1.3

Revision: 7a [Final]

---

### Authors:

Tommi Koskinen      University of Helsinki (UH)  
Hanna Koivula      University of Helsinki (UH)  
Jörgen Palmgren      University of Helsinki (UH)

|  |  |   |
|--|--|---|
| Project co-funded by the European Commission within the ICT Policy Support Programme |  |   |
| Dissemination Level  |  |   |
| P  | Public   | P |
| C  | Confidential, only for members of the consortium and the Commission Services |   |



## 0 REVISION AND DISTRIBUTION HISTORY AND STATEMENT OF ORIGINALITY

### Revision History

| Revision | Date       | Author            | Organisation | Description   |
|----------|------------|-------------------|--------------|---|
| 1        | 2011-07-15 | T. Koskinen       | UH           | The software is adjusted to use an SQL Server version of the Catalogue of Life (CoL) database   |
| 2        | 2011-08-04 | T. Koskinen       | UH           | Implementation of combined result sets from two source portals (CoL and eu.nomen.eu)  |
| 3        | 2011-08-15 | T. Koskinen       | UH           | Implementation of a concept check against the CoL   |
| 4        | 2011-09-09 | T. Koskinen       | UH           | After considering different alternative approaches to matching scientific names, the members of the TMG settled on parsing the names locally rather than only matching against the full scientific name strings that include authors and years. |
| 5        | 2011-09-20 | T. Koskinen       | UH           | Implementation of name string parsing. This allowed the use of canonical names in the name check and provided speed optimization in fuzzy matching.   |
| 6        | 2011-10-03 | T. Koskinen       | UH           | The system of assigning the score to the result set items was agreed and harmonized, so that both name services from London and from Helsinki use a similar logic.  |
| 7        | 2011-10-05 | T. Koskinen       | UH           | Installation of a local database of Fauna Europaea and software development for a name check against this data set  |
| 7a       | 2011-11-17 | Coordination Team | BGBM         | Minor editing.  |

### Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

### Distribution

| Recipient                                    | Date       | Version | Accepted YES/NO |
|--|------------|---------|-----------------|
| Technical Management Group                   | various    | 1-6     | N.A.            |
| Anton Güntsch                                | 2011-11-11 | 7       | YES             |
| Work Package 4 Leader: Christoph Häuser, MFN | 2011-11-20 | 7a      | YES             |
| Project Coordinator: Walter Berendsohn, BGBM | 2011-11-17 | 7a      | YES             |



## TABLE OF CONTENTS

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>REVISION AND DISTRIBUTION HISTORY AND STATEMENT OF ORIGINALITY .....</b> | <b>2</b> |
| <b>2</b> | <b>PURPOSE .....</b>  | <b>4</b> |
| <b>3</b> | <b>FUNCTIONS AVAILABLE IN THE WEB SERVICE .....</b>                         | <b>4</b> |
| <b>4</b> | <b>LEVELS OF RESPONSE PROVIDED BY THE WEB SERVICE .....</b>                 | <b>4</b> |
| 4.1      | Level 1 response .....  | 4        |
| 4.2      | Level 2 response .....  | 4        |
| 4.3      | Level 3 response .....  | 4        |
| <b>5</b> | <b>URL STRUCTURE .....</b>  | <b>5</b> |
| <b>6</b> | <b>EXAMPLES .....</b>   | <b>5</b> |
| 6.1      | Name check .....  | 5        |
| 6.2      | Concept check .....   | 5        |
| 6.3      | Error handling with malformed search parameters: .....                      | 6        |
| 6.4      | Combined result set .....   | 6        |
| 6.5      | Metadata service:.....  | 6        |
| 6.6      | SOAP search .....   | 6        |
| 6.7      | POST requests .....   | 6        |



## **1 PURPOSE**

The purpose of the zoological data quality service is to provide a machine readable back end data service. This web service is to be used by a software called Data Quality Toolkit that is being separately developed by the OpenUp! project.

## **2 FUNCTIONS AVAILABLE IN THE WEB SERVICE**

The prototype provides an application programming interface (API) for searching and matching zoological scientific names. Locally installed versions of the Catalogue of Life and Fauna Europaea databases running on an SQL Server are used in the background as the data source. The service provides query parsing with error handling and offers two search modes: a single search mode and a multiple search mode. A brief description of the service is also provided, this is called the Metadata service.

The name check provides a result set of exact and near matches from the databases used. A score between 0 and 100 is assigned to each result set record.

The concept check result set includes synonyms and accepted names as well as the relevant status information. The JSON data format is used in the requests and in the responses.

## **3 LEVELS OF RESPONSE PROVIDED BY THE WEB SERVICE**

The name checks support all the three response levels and the concept check supports response levels 1 and 2.

### *3.1 Level 1 response*

The service provides exact matching of the full name including author and year. The score value in the results is set to 100.

### *3.2 Level 2 response*

The service provides exact matching using the canonical name obtained via parsing the full query string. The score value is set to 90.

### *3.3 Level 3 response*

The web service does fuzzy matching against the canonical names in the database and the result set includes both exact matches and near matches. Here, for the sake of speed optimization, the genus value in the query string is presumed correct. Near matches are assigned a score value below 90 as per the edit distance from the query name.



## 4 URL STRUCTURE

The URLs are set up using the following pattern:

- names/col.php (name check using of the Catalogue of Life database)
- names/fauna.php (name check using the Fauna Europaea database)
- names/all.php php (name check using both the Catalogue of Life and the Fauna Europaea databases)
- names/nomen.php (name check using eu.nomen via the SOAP interface)
- concepts/col.php (concept check using the Catalogue of Life database)

## 5 EXAMPLES

A list of links is provided below, demonstrating the current functionality of the zoological name services. Please note that the result pages of these links are best viewed by installing the **JSONView plugin** in the browser. This plugin is available both for the Firefox and Chrome browsers. Internet Explorer offers the response for downloading and saving. The saved JSON-file can be viewed with an editor.

### 5.1 Name check

- Single query from the Catalogue of Life database:  
[http://kirjotikka.fmnh.helsinki.fi/names/col.php?query={%22query%22%3A%22Tribolium%20castaneum%20\(Herbst,%201797\)%22}](http://kirjotikka.fmnh.helsinki.fi/names/col.php?query={%22query%22%3A%22Tribolium%20castaneum%20(Herbst,%201797)%22})
- Single query from the Fauna Europaea database:  
[http://kirjotikka.fmnh.helsinki.fi/names/fauna.php?query={%22query%22%3A%22Cygnus%20cygnus%20\(Linnaeus,%201758\)%22}](http://kirjotikka.fmnh.helsinki.fi/names/fauna.php?query={%22query%22%3A%22Cygnus%20cygnus%20(Linnaeus,%201758)%22})
- Multiple query mode:  
[http://kirjotikka.fmnh.helsinki.fi/names/col.php?queries={%22q1%22:%22query%22:%22Tribolium%20castaneum%20\(Herbst,%201797\)%22,%22q2%22:%22query%22:%22Branta%20leucopsis%22}](http://kirjotikka.fmnh.helsinki.fi/names/col.php?queries={%22q1%22:%22query%22:%22Tribolium%20castaneum%20(Herbst,%201797)%22,%22q2%22:%22query%22:%22Branta%20leucopsis%22})
- Near matches found with fuzzy searching - the search terms are slightly misspelled:  
[http://kirjotikka.fmnh.helsinki.fi/names/col.php?queries={%22q1%22:%22query%22:%22Tribolium%20castaneux%20\(Herbst,%201797\)%22,%22q2%22:%22query%22:%22Branta%20leucopsix%22}](http://kirjotikka.fmnh.helsinki.fi/names/col.php?queries={%22q1%22:%22query%22:%22Tribolium%20castaneux%20(Herbst,%201797)%22,%22q2%22:%22query%22:%22Branta%20leucopsix%22})

### 5.2 Concept check

The concept search returns both synonyms and accepted names of a search term. Status information is included.

<http://kirjotikka.fmnh.helsinki.fi/concepts/col.php?query={%22query%22:%22Orellia%20falcata%22}>



### 5.3 Error handling with malformed search parameters:

<http://kirjotikka.fmnh.helsinki.fi/ci/index.php/names/col?queries=q1%22:%22query%22:%22Tribolium%20castaneum>

### 5.4 Combined result set

Here result sets from both source portals are combined.

<http://kirjotikka.fmnh.helsinki.fi/names/all.php?queries=%7b%22q1%22:%7b%22query%22:%22Tribolium%20castaneum%22%7d,%22q2%22:%7b%22query%22:%22Branta%20leucopsis%22%7d%7d>

### 5.5 Metadata service:

A brief description is returned when the base URL is called with no parameters.

<http://kirjotikka.fmnh.helsinki.fi/names/col.php>

### 5.6 SOAP search

This is an example of a SOAP response from eu.nomen.eu.

[http://kirjotikka.fmnh.helsinki.fi/names/nomen.php?query={%22query%22:%22Cygnus%20cygnus%20\(Linnaeus,%201758\)%22}](http://kirjotikka.fmnh.helsinki.fi/names/nomen.php?query={%22query%22:%22Cygnus%20cygnus%20(Linnaeus,%201758)%22})

### 5.7 POST requests

Both GET and POST requests are supported.

The above examples use GET parameters via the URL string. This is appropriate for demonstration purposes. For production use, it is recommended to use POST parameters and the multiple request mode. This has speed advantage and places less limits to the length of the parameters.

HTML forms demonstrating the use of POST parameters are available in these URLs:

[http://kirjotikka.fmnh.helsinki.fi/testing/multiple\\_request.html](http://kirjotikka.fmnh.helsinki.fi/testing/multiple_request.html)

[http://kirjotikka.fmnh.helsinki.fi/testing/multiple\\_request\\_concepts.html](http://kirjotikka.fmnh.helsinki.fi/testing/multiple_request_concepts.html)