Opening Up the Natural History Heritage for Europeana

D17 – DELIVERABLE 6.2.4

Project Acronym: **OpenUp!**

Grant Agreement No: **270890**

Project Title: **Opening up the Natural History Heritage for Europeana**

Productive system for harvesting and parsing reference information

D17 – Deliverable 6.2.4

Revision: 4a

Authors:

Wolfgang Koller NHMW

Heimo Rainer NHMW

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
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
-	2012-12	TMG	-	Discussion of productive system within TMG
-	2012-12-12	Participants of workshop	-	Discussion of productive implementation as part of BioVel Taxonomy services workshop
1	2013-02-18	Wolfgang Koller	NHMW	Initial setup of document
2	2013-02-19	Wolfgang Koller	NHMW	Improving overview, adding source components description
3	2013-02-19	Heimo Rainer	NHMW	Remarks regarding data sources
4	2013-02-20	Wolfgang Koller	NHMW	Applying correct template; updating revision history to reflect discussions
4a	2013-02-25	A. Michel, P. Böttinger & W. Berendsohn (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This deliverable contains original unpublished work except where clearly indicated otherwise.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
TMG (AIT, BGBM, GBIF, IBSAS, MFN, MRAC, NHM, NHMW, RBGK, UH)	2013-02-19	3	NO
TMG (AIT, BGBM, GBIF, IBSAS, MFN, MRAC, NHM, NHMW, RBGK, UH)	2013-02-20	4	YES
Work Package Leader (Heimo Rainer, WP6)	2013-02-25	4	YES
Project Coordinator (W. Berendsohn, BGBM)	2013-02-25	4a	YES

Table of Contents

1 DESCRIPTION OF WORK

The purpose of this document is to provide a description of the productive system for harvesting and parsing reference information. The main implementation for this component was done using the common names webservice. As shown in Figure 1, its main purpose is to be utilized by the OpenUp! Metadata Enrichment Service (documented in C3.4.0) in order to provide enriched content to Europeana.

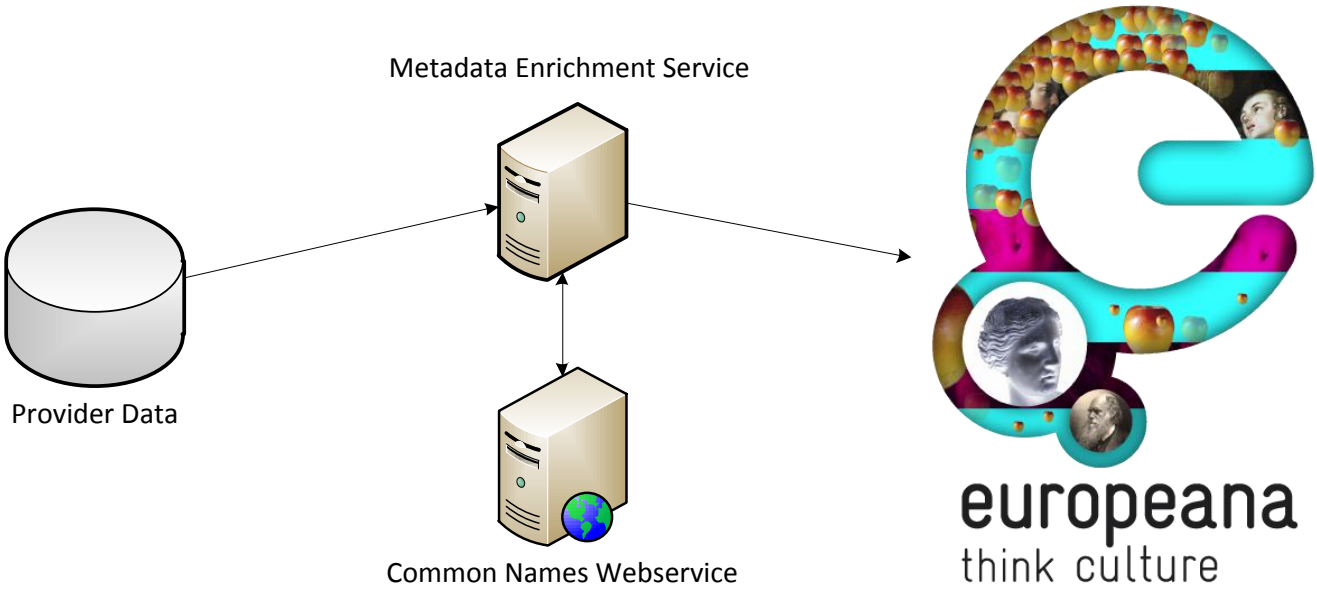

Figure 1 Common names webservice utilization

The service itself is hosted at the Natural History Museum in Vienna. The request and response format are described in detail later on in this document.

Yii[1] was chosen as framework for implementing the service, as it is open source and it provides a large collection of verified components. The actual implementation of the common names webservice is published on github[2] as an open source project.

In order to provide a diverse range of languages, various sources for common names were taken into account. More details on this can be found in Deliverable 18 (D6.6.0) “Report on multilingual data for natural history objects”.

[1] http://www.yiiframework.com/

[2] https://github.com/wkollernhm/openup

2 ARCHITECTURE

The overall architecture of the productive implementation is similar to the prototype. Figure 2 shows the slightly adapted architecture.

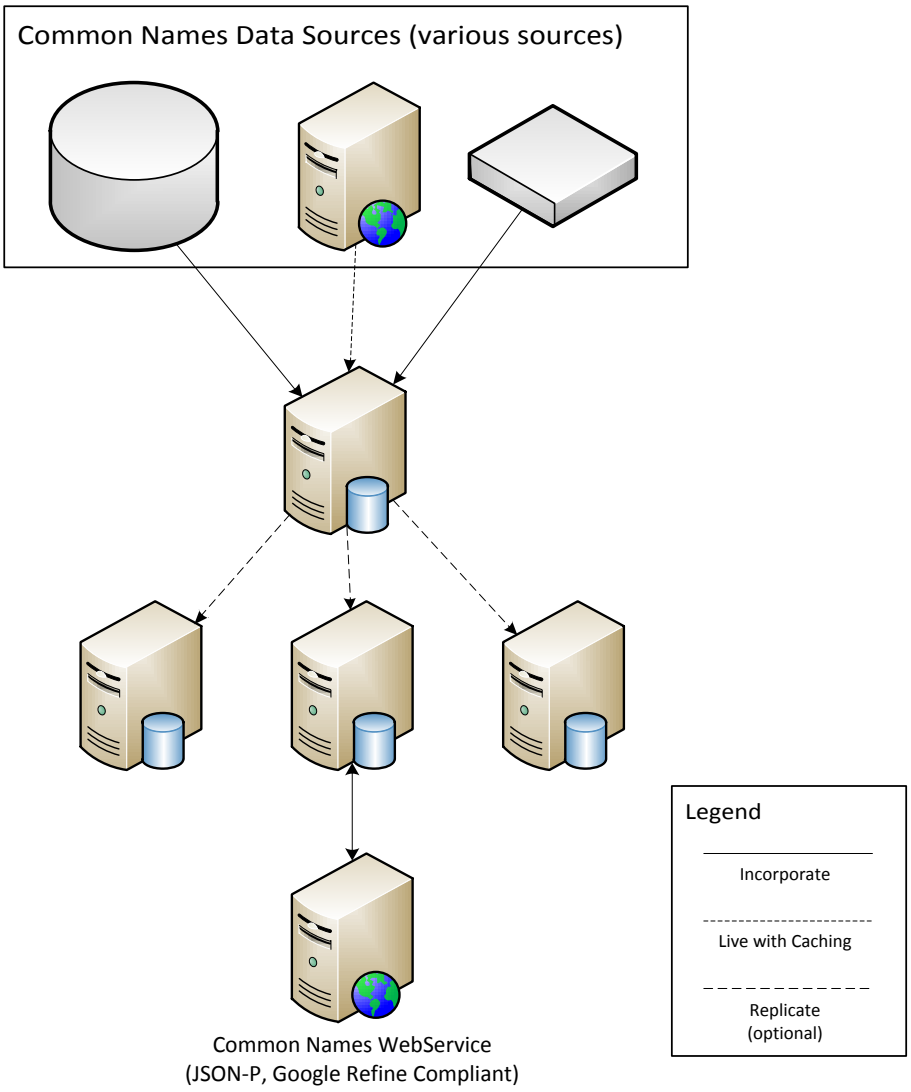

Figure 2 Overall architecture of common names webservice

The references to the botanical and zoological services have been removed as they were replaced by the Global Names Architecture (GNA) nameParser service[3]. This is now an internal component of the common names webservice and is therefore not displayed in the architecture above but instead documented below.

[3] https://rubygems.org/gems/biodiversity19

2.1 Internal dataflow

The internal flow of the service was adapted to the new architecture. In order to provide a more stable response the GNA nameParser service was installed to pre-parse all requests and use only the parsed canonical name returned from the service. In order to speed up response times the nameParser is running locally on the same machine as the common names webservice.

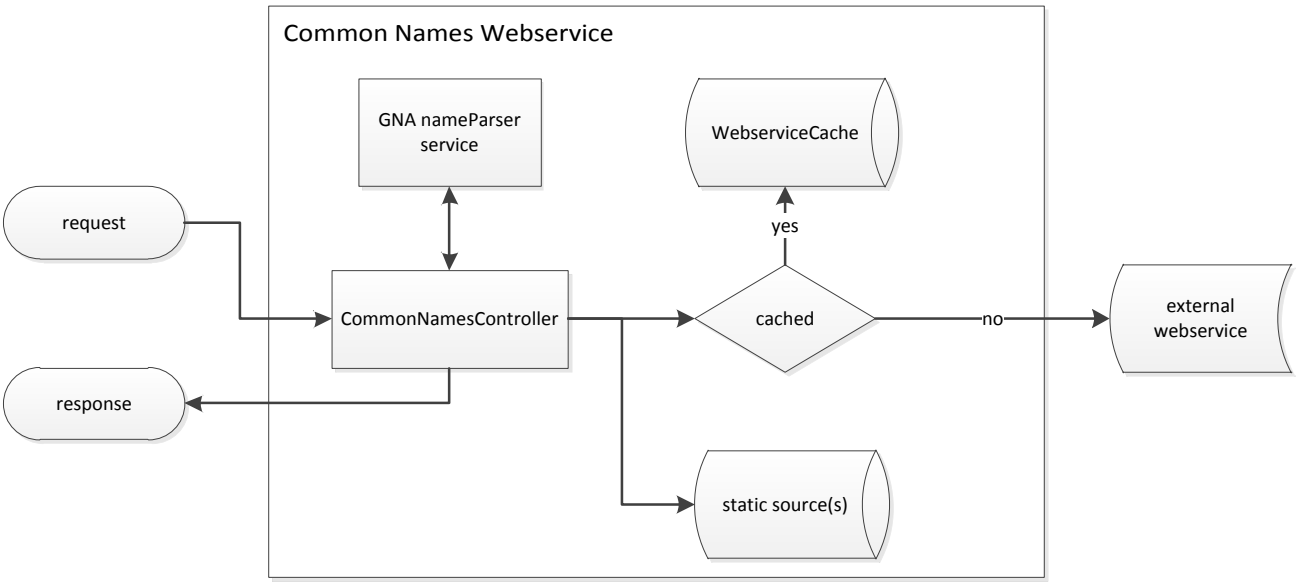

Figure 3 Structured flowchart outlining the internal information flow

In Figure 3 the information flow inside the service is outlined using a structured flowchart diagram. As illustrated the GNA nameParser service is used to pre-parse all requests to the common names webservice. Only the returned canonical scientific name is used to process all further requests.

Any registered source inside the service is then queried using the canonical scientific name. For external sources (i.e. webservices) there is an internal cache for preventing the common names webservice from spamming external sources for the same request. In addition this cache is used to gain a performance increase. Internal (static) sources are queried directly using a MySQL database for storage.

The productive system for the caching environment (due in M30 – August 2013) will contain an additional de-duplication layer for all sources. In combination with the GNA nameParser service this will then return a more compact response.

2.2 Source components

Sources for common names are not only available as webservices but rather also in static formats (like Excel, CSV, Text, etc.). Therefore it is necessary to prepare those sources to be used by the common names webservice.

All static sources are imported into their own MySQL table in order to be able to query for properties (like the scientific name). As the format of those tables highly depends on the provided source, it is necessary to build a wrapper for those tables to work together with the common names service.

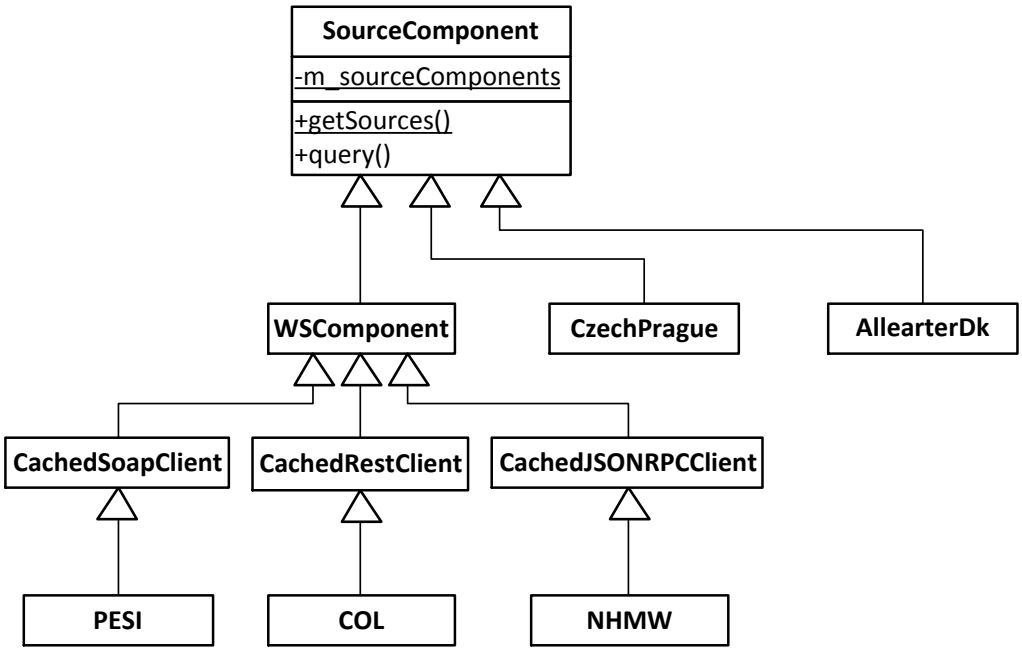

Figure 4 UML class diagram for source components

Those wrappers are implemented inside the service using the component architecture of Yii. As shown in Figure 4 every wrapper is its own component preparing the source data (including external webservices) for a common interface towards the common names webservice.

Common functionalities (e.g. access protocol, caching functionality, etc.) are handled in base classes. Those classes can be re-used for new sources.

The SourceComponent specification also includes an automatic registration service for new sources, which means no changes to the actual service are required in order to integrate and activate that new source. As soon as a new component implementation is instantiated, it is registered and included as a source in the common names webservice.

This mechanism ensures that new common names can be integrated with least effort. In addition changes to existing source (e.g. leveraging a static source to a webservice) can be done by deactivating the static, and integrating the dynamic source. This yields a dynamic service infrastructure without the requirement to directly incorporate external content, thus allowing to always serve the latest content provided by all sources.

3 REQUEST AND RESPONSE

The service is located at the Natural History Museum Vienna and can be reached using the following URL:

http://openup.nhm-wien.ac.at/commonNames/

Requests and responses follow the Google Reconciliation Service API definition[4]. However as the definition has limited response fields, it was extended to include additional information (a more detailed explanation will follow below).

```
{
  "name":"OpenUp! Common Names Service",
  "identifierSpace":"http:\/\/open-up.eu\/commonNames\/",
  "schemaSpace":"http:\/\/open-up.eu\/commonNames\/"
}
```

Figure 5 Metadata for common names service

Figure 5 shows the default response when querying the service without any parameters. This response is equal to the "Service Metadata" as defined by the API specifications.

3.1 Requests

Requests have to follow the specification of the Google Reconciliation Service API, both single- and multiple query-mode are supported. An example query is expressed by the following URL:

http://openup.nhm-wien.ac.at/commonNames/?query={%22type%22:%22/name/common%22,%22query%22:%22Platalea+leucorodia%22}

This URL queries the service for all common names which can be found for the scientific name "Platalea leucorodia". The query parameter is an URL-encoded JSON object (as shown in Figure 6) which contains parameters according to the Google Reconciliation API specification.

[4] https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-Api

```
{
  "query":"Platalea leucorodia",
  "type":"/name/common"
}
```

Figure 6 Sample query for "Platalea leucorodia"

The response for this query contains the common names from all sources together in a result as specified by the Google Reconciliation API.

The possible request parameters are summarized in Table 1.

Table 1 Request parameters

field	description	example
query	Taxon Name to query for	Platalea leucorodia
limit	Limit number of results	3
type	should always be "/name/common"	/name/common

3.2 Response

Figure 7 shows the response (shortened for better readability) for the sample given above. As mentioned earlier there are additional fields returned since the Google Reconciliation API definition has only limited response fields.


```
{
  "result":[
    {
      "id":"urn:lsid:marinespecies.org:taxname:416678",
      "name":"כף-אבי כפן",
      "type":"\/name\/common",
      "score":100,
      "match":true,
      "language":"he",
      "reference":"pesi",
      "references":[
        "pesi"
      ],
      "taxon":"Platalea leucorodia",
      "taxon_id":"urn:lsid:marinespecies.org:taxname:416678"
    },
    {
      "id":"Eurasian Spoonbill",
      "name":"Eurasian Spoonbill",
      "type":"\/name\/common",
      "score":100,
      "match":true,
      "language":"English",
      "reference":"Bisby F., Roskov Y., Culham A., Orrell T., Nicolson D., Paglinawan L., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Hernandez F., De Wever A., Kunze T., eds (2013). Species 2000 & ITIS Catalogue of Life, 8th February 2013. Digital resource at www.catalogueoflife.org\/col\/. Species 2000: Reading, UK.; 2011  IOC World Bird List, Master List v2.9;Gill, Frank, and Minturn Wright 2006  Birds of the World: Recommended English Names; 2011  AOU Check-List (08-2011)",
      "references":[
        "Bisby F., Roskov Y., Culham A., Orrell T., Nicolson D., Paglinawan L., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Hernandez F., De Wever A., Kunze T., eds (2013). Species 2000 & ITIS Catalogue of Life, 8th February 2013. Digital resource at www.catalogueoflife.org\/col\/. Species 2000: Reading, UK.",
        " 2011  IOC World Bird List, Master List v2.9",
        "Gill, Frank, and Minturn Wright 2006  Birds of the World: Recommended English Names",
        " 2011  AOU Check-List (08-2011)"
      ],
      "taxon":"Platalea leucorodia",
      "taxon_id":11909258
    }
  ]
}
```


Figure 7 Sample response for "Platalae leucorodia"

Table 2 shows a list of all returned fields including their description and providing an example.

Table 2 Response parameters

field	description	example
id	identifier for common name	62162
name	the actual common name	Boasslbee
type	will always be "/name/common"	/name/common
score	accuracy of found taxon	87.3
match	true if this is an exact match	false
language	ISO 639-6 code for language of the common name	bar
geography	geographical region which this common name is used in	aut
period	time-Period in which the common name was/is used	1934-1950
taxon	name of taxon the common name applies too (required due to fuzzy matching)	Berberis vulgaris
taxon_id	id of taxon	523674
reference	(deprecated) reference (source) for this common name	[Catalogue of Life] 2011 AOU Check-List (08-2011)
references	array of references for this common name	[“2011 AOU Check-List (08-2011)”,”PESI”]

The “reference” response field is deprecated and should not be used anymore. Instead the “references” field contains an array of all used references, each entry being a unique one. For compatibility reasons the “reference” field is still returned containing a concatenated list of references.

4 LIST OF FIGURES

5 LIST OF TABLES