Opening Up the Natural History Heritage for Europeana

D18 - DELIVERABLE 6.6.0

Project Acronym: **OpenUp!**

Grant Agreement No: **270890**

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

Report on multilingual data for natural history objects

D18 - Deliverable 6.6.0

Revision: 3a

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
Draft	2013-02-20	Wolfgang Koller	NHMW	Initial setup of document
1	2013-02-21	Wolfgang Koller	NHMW	Finalizing first version of document
2	2013-02-22	Wolfgang Koller	NHMW	Adding common names properties diagram
3	2013-02-05	Heimo Rainer	NHMW	Proof reading & comments
3a	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This deliverable contains original unpublished work except where clearly indicated otherwise.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
Work Package Leader (Heimo Rainer, WP6)	2013-02-25	3	YES
Project Coordinator (W. Berendsohn, BGBM)	2013-02-25	3a	YES

Table of Contents

1 DESCRIPTION OF WORK

This document provides an overview of the multilingual content available for natural history objects. In addition content relevant for Europeana is highlighted in order to provide a list of sources already available through the common names webservice implemented by WP6 and described in Deliverable 17 (D 6.2.4).

As a result of the gathering effort of common names for the service, several languages (including non-European languages) were discovered to be useful for enriching content within Europeana.

2 ASPECTS OF NATURAL HISTORY OBJECTS

Natural history objects are mostly handled through their scientific classification (i.e. latin name). Other properties (like the description) are today mostly provided directly in English or in a language neutral way (e.g. sizes, years, weights, etc.).

Therefore, the main challenge for multilingual content is the enhancement of scientific specific vocabularies towards a description recognizable for the public. As mentioned above most natural history objects are handled through their scientific classification. Through services, like the common names webservice developed in WP6, this scientific classification can be enhanced to be searchable by the general public.

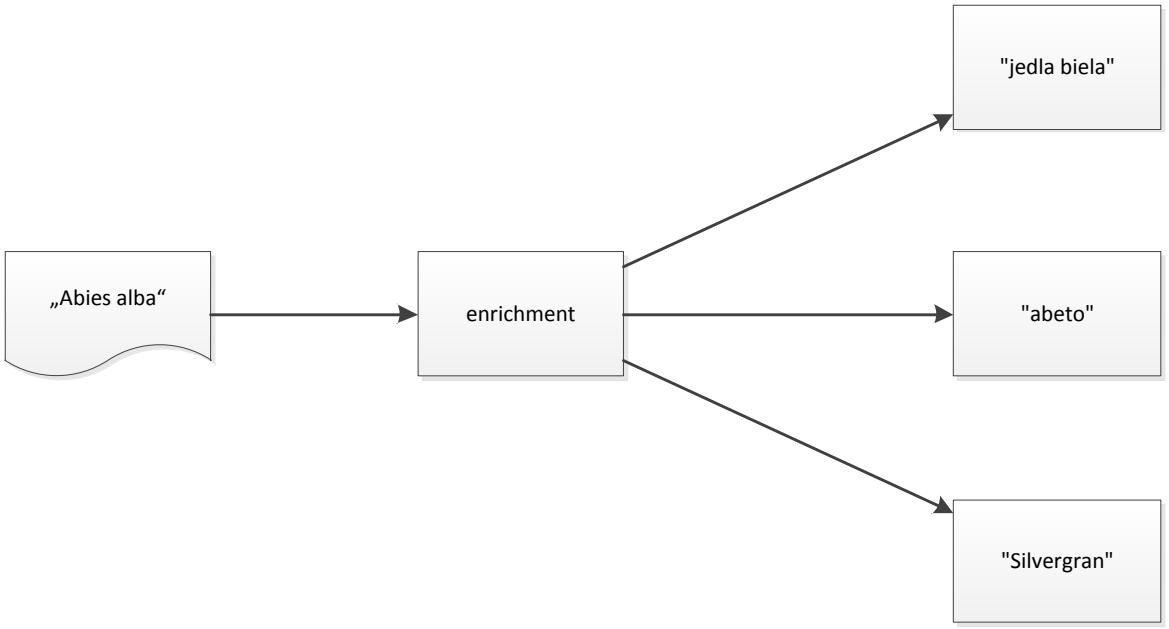

Figure 1 Improving Europeana significance through multilingual enhancement

3 CONTROLLED VOCABULARIES

Within the domain of multilingual content, controlled vocabularies can be applied for some of the metadata. Several classifying properties of a common name were identified as part of the specification of the common names webservice. Table 1 lists the identified properties including a description and an example.

Table 1 Properties of multilingual content

Property	Description	example
Name	The actual common name	Boasslbee
Geography	Geographic region it is used in	Aut
Language	The language	Bar
Period	Period of time it was used	1934-1950
Ethnicities	Ethnicities which use it	Austrian
Entity	The part / object / etc. this name applies to	Berberis vulgaris

The main challenge for a controlled vocabulary is to classify the given properties.

3.1 Geography

Geography related terms are widely used and therefore solutions already exist. As part of the OpenUp! project it was recommended to use the geonames webservice[1] infrastructure, which offers a very detailed set of geography related terms, together with stable identifiers.

3.2 Language

Particularly challenging are language terms. There exist a wide range of classifications for expressing them (e.g. ISO639-3, ISO639-6, etc.). In addition some providers are using simple written expressions (like “English”, “Deutsch”, etc.). The cross mapping between all of those expressions is a challenge and an ongoing activity. By utilizing several services and incorporating different standards it should be possible to provide at least a robust cross mapping of the most common language expressions. The WP6 common names webservice uses the ISO639-6 standard, as it allows expressing the most accurate specification of a language (including spoken variants). A database of the current ISO639-6 language terms is provided by geolang[2].

[1] http://www.geonames.org/export/ws-overview.html

[2] http://www.geolang.com/iso639-6/

3.3 Period

Period terms are often expressed using very different vocabularies. Possible expressions are: written terms (e.g. “recent”), pure numbers (e.g. “1987”), number ranges (e.g. “1847-1940”), incomplete dates (e.g. “19XX”), etc. These very different expressions make it hard to classify the terms. As part of the WP6 common names webservice it was decided to treat the period as a simple string expression.

3.4 Ethnicities

This property was identified but is not used within the WP6 common names webservice, as there is not much source data available. However as the specification tries to be as complete as possible it was still included.

3.5 Entity

Entities are very individual and are specific to the context of the common name. In the context of natural history objects most of the time the entity is the scientific name of the object. This allows easy mapping of names to the object, as the scientific name already provides a common understanding.

3.6 Properties relations

As a direct result of the specification of the common names webservice, a model for handling common names for scientific content was developed. This model takes into account all properties as listed above. In addition it allows to even specify parts of entities the name applies to (this is relevant for e.g. parts of the plants where the fruit has an own name, the stem might have another, etc.). The model was expressed using an entity relationship (ER) diagram as shown in Figure 2.

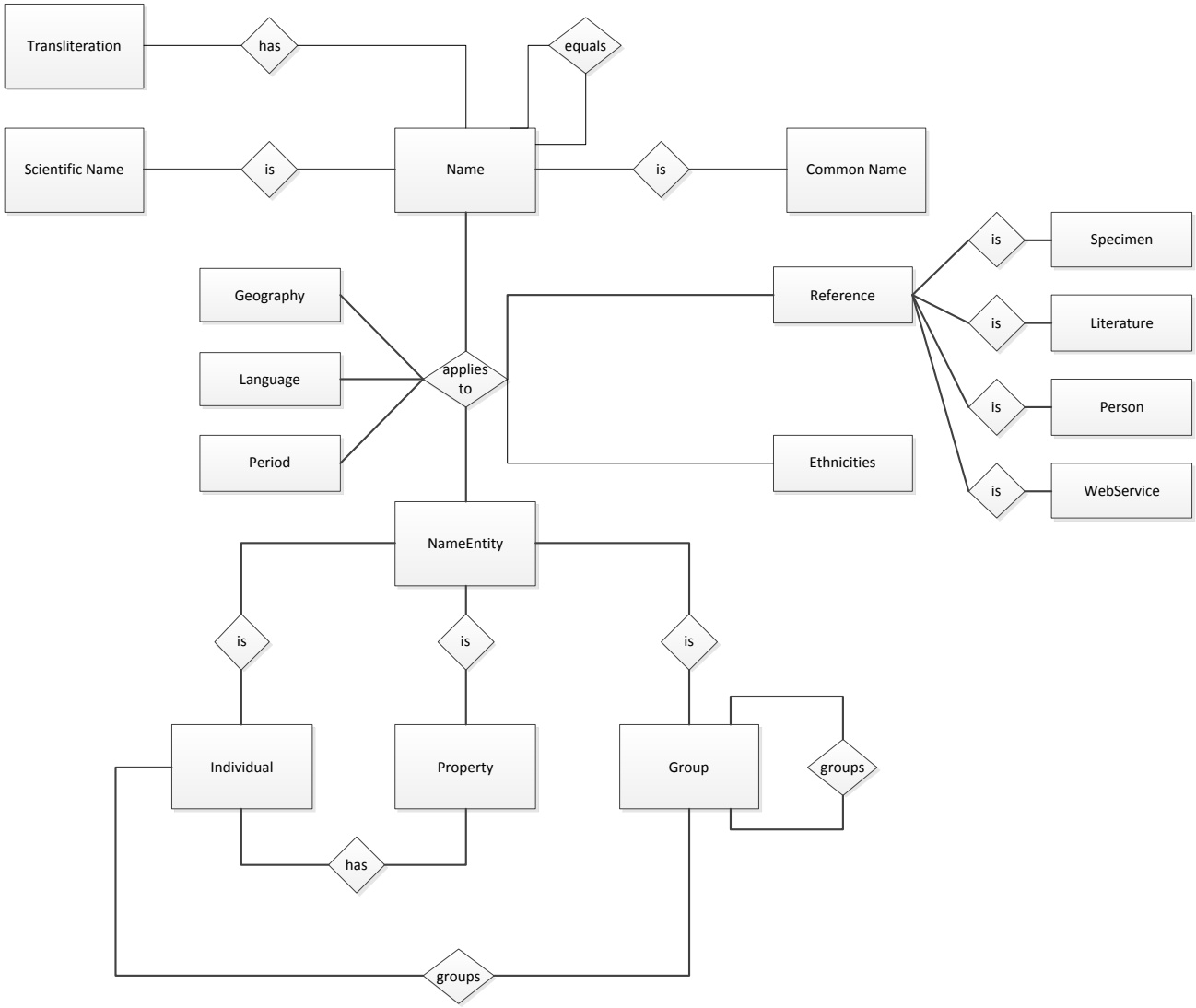

Figure 2 ER-diagram for common name properties

4 LIST OF SOURCES

Table 2 provides a list of all sources identified so far. Not all of them are integrated into the productive implementation yet, but will be towards the end of the project.

Table 2 List of sources identified so far

name	language(s)	description
AllearterDk	Danish	Danish names from various references (cited inside source)
Artsdatenbanken	Norwegian	
Austrian academy of sciences	Austrian dialects	Fungi and plant names in various Austrian dialects
Bratislava	Slovak	Fungi and plant names
Catalogue of Life	multiple	REST service provided by the Catalogue of Life
CzechPrague	Czech	Czech names for plants
Dyntaxa	Swedish	
Linda	Hebrew	
New Zealand landcare	English, Maori	Fungi and plant names
NHMW	multiple	Multiple languages and common names taken from the virtual herbaria
PESI	multiple	SOAP service provided by the PESI project

5 LIST OF FIGURES

6 LIST OF TABLES