Taxonomic Data in BHL-E

The whence, why and whither of taxonomic data in the BHL-Europe architecture and data flow

Andreas Kohlbecker, Botanic Garden and Botanical Museum Berlin-Dahlem, March 1st, 2011; last updated March 3rd, 2011

# Term definitions

First of all before going into deeper discussion on this topic it is important to get an idea about the terms which will be used. This might be well understood but to assure a common sense on these terms I am giving again a definition. I’ll use a butterfly as example which is well known:



It has a **common name** (= colloquial name, =vernacular name) in many languages: en= European Peacock, de= Tagpfauenauge, fr= Paon du jour, etc.

**Scientific** Taxon **Name**: The name of a species, genus, etc. as it appears in a publication. This name is formed according to the rules defined in the [Nomenclature Codes](http://en.wikipedia.org/wiki/Nomenclature_Codes). A name consists of several parts:

Genus name: In our example this is *Inachis*. You can think ofthe genus name as something like a person’s surname or familyname. A genus groups multiple species.  
The specific epithet: This is name part is given to species, in our example this is *io*. The specific epithet is comparable to a person’s given name.

A scientific name is not bound to a specific taxonomic rank although it is possible to infer according ranks from the name, e.g.

*Inachis L.* refers to a taxon with a rank above the species level  
*Inachis io L.* refers to a taxon with a rank at the species level or below

A scientific taxon name might occur in publications in an abbreviated form: “*I. io*”, “*io*” or in other **orthographic variants**: “*Inachis io L.”, “Inachis io* (Linneaus, 1758)”, “*Inachis io* (Linneaus), “*Inachis io* (L.)”**.** An orthographic variant is just another way of writing this name, which may be invalid in terms of the Nomenclature Codes. However, these variants have absolutely nothing to do with synonyms, which are related to the “real” taxa.

A “real” **Taxon** is given a scientific name and a rank by an author who is expressing his opinion on a group of living (or conserved) beings by this act. Lower taxa like species are structures in a classification tree. In this tree any taxon might have child taxa. For example the species *Inachis* io L. is a child of the genus *Inachis* L. Different scientists, authors, may have different opinions on where a taxon should be sorted in in this classification tree. For example the *Inachis io L*. once was located in the genus *Vanessa* together but KARSHOLT & RAZOWSKI published their opinion in 1996 that this butterfly best fits into the genus *Inachis.*

Excuse me, but the information on the most recent classification of this butterfly has changed again in 2003 and now this opinion is now commonly accepted. WAHLBERG & NYLIN (2003) have put it into the genus *Aglais.* So the accepted taxon name now is *Aglais io* (Linneaus, 1758). By the way the author name in the brackets honors the author of the original publication, who first described this butterfly. Now we know the accepted name and some synonyms of it:

***Aglais io* (Linneaus, 1758)** *Vanessa io  
 Inachis io*

But under which name has Linné actually published it? I was *Nymphalis io,* so we now know another synonym of this name. The full synonymy now looks like:

***Aglais io* (Linneaus, 1758)** *Nymphalis io   
 Vanessa io   
 Inachis io*

At the same time multiple concurrent opinions on the “right” classification might exist in parallel. In this quit usual situation different classification trees and synonym lists exist at the same time.

By the way if you [search BHL for the scientific name *Aglais io*](http://biodiversitylibrary.org/search.aspx?SearchTerm=Aglais+io&SearchCat=M)you get not a single result!

The page in the original publication is this one: <http://biodiversitylibrary.org/page/727383> due to the bad OCR quality TaxonFinder is not able to recognize a single one of the butterflies described in this page. The only taxon found is the forage crops of *Aglais* *io.*

# Whence? Where is BHL-E getting taxonomic information from?

## Scientific Names & orthographic variants

1. There is one content provider which will actually deliver scientific names together with the metadata. 2. The OCR’ed text will be analyzed by TaxonFinder which will also deliver scientific names. In both cases we get scientific names and their position in the text, whereas the actual occurrence in text might be an orthographic variant of the correct scientific name. Other orthographic variants can be generated from the scientific name.

Scientific Names & orthographic variants is data which actually belongs to the published text.

It is very hard to infer the “real” taxon from the OCR’ed text without any further information. Well, the author might have had an opinion on the classification of the taxon for which the published name has been used. This means that the author was using the name in the sense (*sensu*) of a certain publication but this information is not published and can only be retrieved in rare cases.

So storing any taxonomic information in the BHL-E metadata is semantically not correct since the text in almost all cases contains only names. It is questionable if the BHL-E metadata scheme should provide the means to store this information at all. There are further reasons against taxonomic data in this schema. A) Taxonomic concepts change, so any real taxonomic information might be outdated after rather short time. From this time on BHL-E would publish incorrect metadata via the OAIS-Access component. B) BHL-US and others are not interested into this data, so there is no need to make this data available to the global BHL infrastructure. So there is no need to store it in the metadata.

## “real” taxonomic information

“real” taxonomic information is best retrieved from data providers which are an authority for this kind of information. The first address for BHL-E is for sure the Species 2000 & ITIS Catalogue of Life. ITIS Catalogue of Life is being published annually on cd-rom. In addition there is the dynamic checklist which is a virtual catalogue operated on the Internet and available both for users and as an electronic web-service at <http://www.catalogueoflife.org/dynamic-checklist>.

So there are two kinds of sources to deal with. Databases published periodically on a medium and web-services. The web -services may be offline, have a varying response performance and so on. Therefore it would be a good idea to provide some caching mechanism for this information inside the BHL-E architecture.

In future also other providers of taxonomic data might be used besides of the Catalogue of Life.

## Common names

Well, where to retrieve them from … TODO…

# Whither? Where to put taxonomic information in the BHL-E Architecture?

In order to answer this question it is important to take a look at the requirements which are listed in the “Feature\_list\_BHL-Europe” (docs.google.com). The taxonomic information is actually needed to provide an advanced search which will have some “taxonomic intelligence”. The following section summarizes the features regarding the taxonomic search but may be incomplete:

## Taxonomic intelligence requirements

* Search by taxon name, common name (multilingual), included taxa (classification)
* Expand taxon name by synonymy, vernacular name (multilingual), misapplied names
* Search for content by common name (=common names are related to taxa not to names !!!)
* Search by geography (Point, Geographic name, ISO country, TDWG area, …)
* Search for protologues (=original publications) => allow to mark nomenclatural relevant literature in the BHL-E schema.

Here with geography another kind of information occurs which also is related to the “real” taxa. The geography of a taxon encompasses its distribution, observations and the location from where specimens have been taken. A prerequisite to be able to handle this information is the ability to handle the taxonomic data correctly. Thus the geography is better discussed after a clear concept on the handling of taxonomic data has been achieved.

All requirements listed above are related to the search functionality provided either to humans via the BHL-E portal user interface or to other applications via web-services. The search functionalities will allow to search based on name information (scientific names, orthographic variants) and on “real” taxonomic information (classification, ranks, synonyms, common names, etc.). Regarding the classification it might be sufficient to always regard to the most recent classification and to ignore historic once.

The name information will be stored in the BHL-E metadata during the metadata enrichment process. The taxonomic information must be treated differently.

## 

## Metadata Enrichment

Metadata enrichment takes place during or prior to the Pre-Ingest. The current data workflow distinguishes a manual enrichment step done by the content providers and an automatic enrichment with is actually much more important. The automatic enrichment is based on the OCR’ed text which will be analyzed by TaxonFinder which will provide scientific names to the actual enrichment process. Before actually storing the newly retrieved names into the metadata records, other services could be to generate further orthographic variants. 

## Search index enrichment and query expansion

According to the information which has been circulated during the past months the search functionalities will be implemented using the Solr/Lucene framework. Since the BHL-E metadata contains the scientific names and orthographic variants Solr will create an index over these. In order to also allow searching the index based on taxonomic information and relations the query terms must be expanded using the taxonomic information (=query expansion). With the analyzer classes Solr provides an elegant means to implement this expansion process. Going deeper into this topic would however exceed the scope of this document; however it is important to note that it is possible to use different Analyzers during the creation of the index and also during the processing of search queries. This means that the query expansion can take place when a search query is being processed or at the time of the index creation which would provide a much better performance. It has to be determined which information actually is suitable to be used for the query expansion during the index creation.

As pointed out earlier the source of taxonomic information will either be quasi static data (long term updates) distributed on media like CD-ROMs or web-services providing life data. Web-services may be inavailable etc. To overcome unreliability the concept foresees to let Sorl retrieve all data through a caching proxy (e.g. Apache2), so all responses on any http requests send will be cached in this proxy. If a service is unavailable, the proxy will return the last response of the service to this specific request. This caching proxy it not meant to be a mirror, so it only caches those responses for which a request has been sent before. 