Mapping of Knowledge in Natural History Museums

Author:
Karl-Heinz Lampe

CIDOC06
GOTHENBURG
SWEDEN

What kind of data do natural history museums have?

Natural History museums are archives of biodiversity. They house millions of specimens collected in space and time, providing first-hand information about geographical and historical presence of organisms, their morphological and other specific characters (anatomical, genetic etc.), and sometimes additional information about ecological environments, host-parasite relationships, etc.

The classification of these specimens/organisms is based on a taxonomic system that tries to mirror phylogenetic relationships between groups of organisms. The result is classes of organisms (or taxa at different taxonomic levels) such as species, subspecies or other infrasubspecific taxa. Furthermore, groups of species can form higher ranked taxa like genera, families, order etc., subdividing the entire world of plants, animals and other living beings into hierarchical biological units. The scientific naming procedure is bound to International Codes of Nomenclature (for animals, for plants and fungi, of bacteria, of viruses, for cultivated plants) and in the case of a species a new name is tied to a single specimen, the so called type specimen.

Therefore curators in a natural history museum are not only 'keepers of specimens'. They are 'keepers of names'. These scientific names are the only ones valid worldwide. They are internationally standardized. Species names are pivots in digital networks. They connect all kind of scientific (and trivial) information about the animated world ranging from the molecular level via the organism level up to the ecosystem level. The disadvantage is that they are based on a concept. Specimens, however, are pivots in nature's networks. They are the real existing actors in life. A name (a picture, a sound, a gen-sequence…) without a relation to the specimen is nothing more than a statement without a proof. The type specimen, on which the species description is based, is the reference for the species in the real world.

Knowledge associated with a well documented collection object

The knowledge associated with a well documented collection object is represented here as an ontology in terms of the CIDOC Conceptual Reference Model or CIDOC CRM[1]. In philosophy, ontology is the study of that what is (and not that what it is for). In informatics, the term is used for a formal specification of semantic concepts. The CIDOC CRM is an object-oriented semantic model. It consists of a class hierarchy of 81 named classes (so called entities; E1, E2, etc.) that are interlinked by 132 named properties (P1, P2, etc.). The CIDOC CRM is event-centric. People, objects, places and time-spans are interrelated by common events. As a knowledge representation tool it is therefore superior to class-centric entity-relation models.

People usually see what they want to see. Showing a pinned insect such as in the centre of *Fig.1* and asking: "What is it?" usually provokes the answer: "A bee!"
A pinned insect, however, is obviously a 'Man made Object'. It consists of a biological object, the bee, and an information carrier; a pin carrying labels as documents.

For ecologists the white label primarily documents a collecting event, for economists the same label documents an acquisition event and for lawyers it might document a transfer of custody (from nature to a collection). Thus the white label documents three CIDOC CRM specific events (multiple instantiation in E7, E8 and E10). It also shows the exact location of a particular person at a certain time. Similarly, the red label documents a type creation of a new species and a determination or type assignment. The type creation event has created a document, a species description, and a type that is a new class in a philosophical sense with the appellation *Colletes alini* Kuhlmann 1999. The bee specimen supported the type creation in the taxonomic role of a holotype. Type creations in other disciplines as for example archaeology are often based on the same or similar methodological approach; the terminology, however, is different.

Of course this ontology or knowledge representation can easily be extended (e.g. object identifier and place of storage within a collection in *(Fig. 1)* or refined to a more detailed description such

[1] The CIDOC CRM has been developed because of already existing different schemas to make them commonly understandable. It was primarily not developed to create new schemas (of course one can use it for that purpose as a best practise guide).

Fig. 1: Knowledge associated with a well documented collection object
in terms of the CIDOC CRM (E1,2.. = Entity No; P1,2... = Property No)

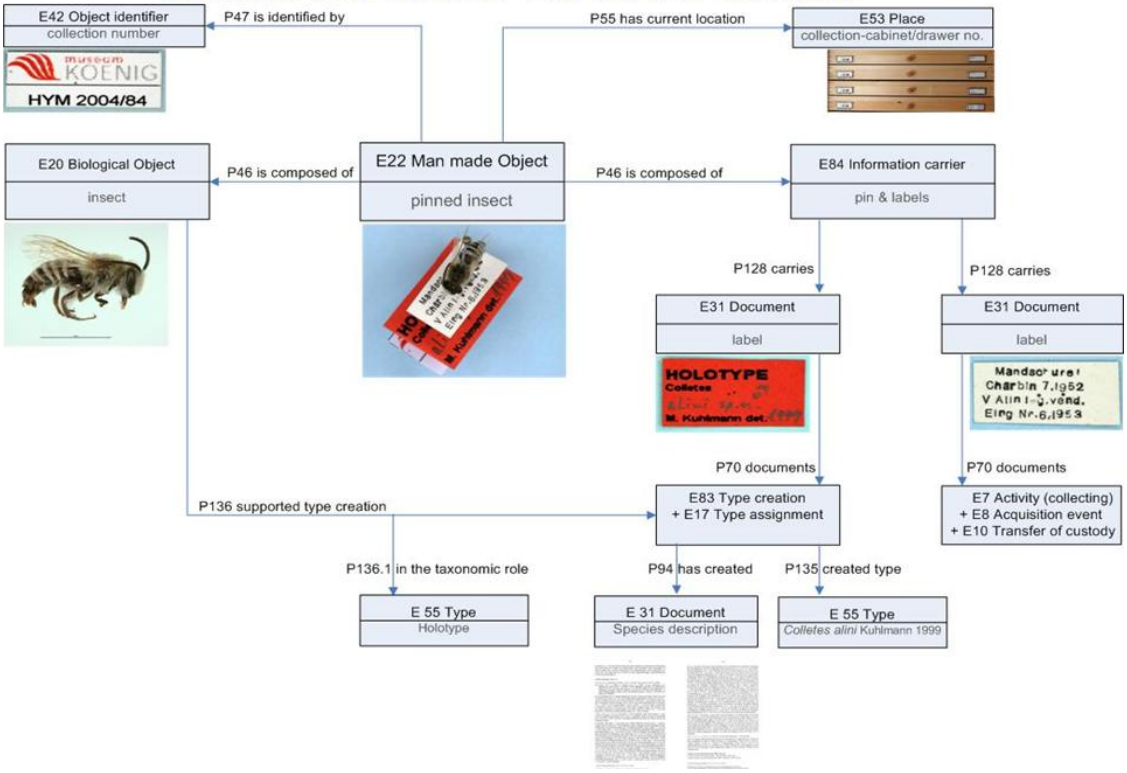
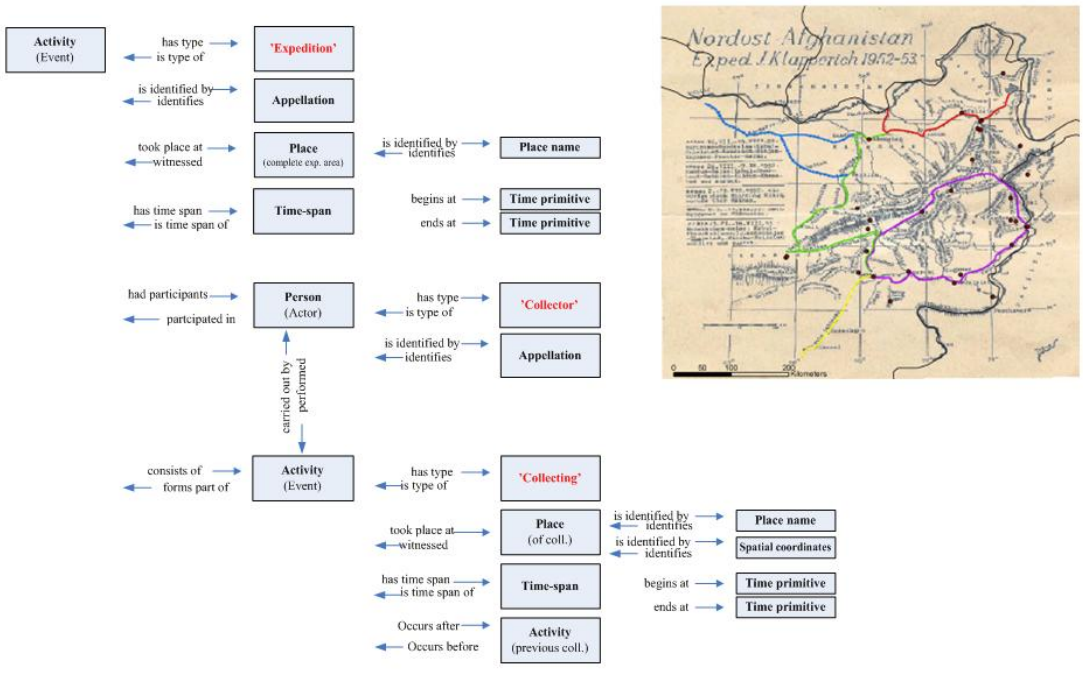


Fig. 2: Expedition- & collecting events in terms of CIDOC CRM



as a collecting event or even an expedition *(Fig. 2)*. The collecting event is an activity carried out by an actor (which is a person, the 'collector', identified by an appellation. It took place at a gathering site (identified either by a Place name and/or spatial coordinates) and lasted a certain time span. An expedition is an activity consisting of continuing collecting activities and participating persons which carry out these activities etc. *(Fig. 2)*. In a similar manner collector's itineraries can be mapped. A mapping of semantic relationships in this way makes domain specific information units understandable to everyone. Secondly, by replacing the term collector by observer/photographer or painter/drawer in *Fig. 2* the same model characterizes observation, photographing and painting or drawing events (instead of a collecting event). Cultural museums for example often house old paintings etc. depicting the animated world in times when no biological voucher specimens were collected, labelled and deposited in a museum of natural history, Digitizing these data in a proper way will complement our knowledge of the history of the animated world. What a fascinating 'new' information source for biologists!

Ontologies like these can be used to filter out one's own area of expertise and to identify additional or overlapping sources of information normally covered by other domains.
For example specimen based databases deal with real world objects. They are focused on primary biodiversity data documented on specimen labels; determination, gathering site, date collected, collector's name etc. Secondly, taxonomic authority files, such as 'Fish Base', the 'Hymenoptera Name Server' or the 'Orthoptera Species File', deal with taxonomic names that stand for human concepts, and are concerned with the validity of the respective names. They are primarily focused on published data. Thirdly, geographic authority files, such as the 'Getty Thesaurus' or the 'Alexandria Digital Library Gazetteer' (ADL), deal with geographic names in a geophysical and geopolitical context. Last but not least other authority files deal with the biography of persons and their identity.

Ontologies can be used as a best practise guide to improve already existing databases by means of dynamic linking, multiple verification, and semantic enrichment as well as developing new tools

via web services by using GRID technology. Furthermore it can help to identify other formerly hidden or even obscure information sources from various scientific and scholarly domains.

Improvement of data quality and generating scientific extra value through dynamic linking, multiple verification and semantic enrichment

Within the 'Digitised Orthoptera Specimen Access' (DORSA) project, initiated by Klaus Riede and myself, a virtual museum of Orthoptera type specimens (grasshoppers, locusts, katydids, crickets) housed in major German museums was created. Thereby the Orthoptera Species File (OSF), a global taxonomic authority file, served and still serves as a taxonomic backbone. The OSF was built and is maintained in Philadelphia/USA (http://osf2.orthoptera.org). DORSA contains full information about 16,000 specimens (incl. 2,300 primary types and 6,700 secondary types). 30,000 images and 11,000 sound recordings are linked to their respective specimens. Many of the specimens linked to the sound recordings belong to hitherto unknown species. They can be seen as 'types for tomorrow'. The DORSA virtual museum is available through the SYSTAX database infrastructure (http://www.biologie.uni-ulm.de/systax) and major knowledge portals such as BIOCASE (http://www.biocase.org) or GBIF (http://www.gbif.org). *Fig. 3* shows a screenshot of specimen based multimedia information units, incl. sound-files. The DORSA web-based repository provides sufficient detailed information, including diagnostic features (e.g. genitalia), to allow for example taxonomists to narrow down loan requests. As mentioned earlier the OSF is a global taxonomic authority file, mainly based on published information. It provides the user with information about the validity of a given taxon name, the source of the original description, the depository place of type specimens and so on. Reciprocal dynamic linking between DORSA and OSF allow direct access to the respective data, e.g. for validity checks of a given taxon *(Fig. 4)*. As you probably know the validity of a given taxon can change in time.

OSF authority files allow verification of published type data from our DORSA specimen data and vice versa *(Table 1)*. Our DORSA taxonomist, Sigfrid Ingrisch, detected about 25% of

Fig. 3: Screenshot of specimen based multimedia information units, incl. sound files

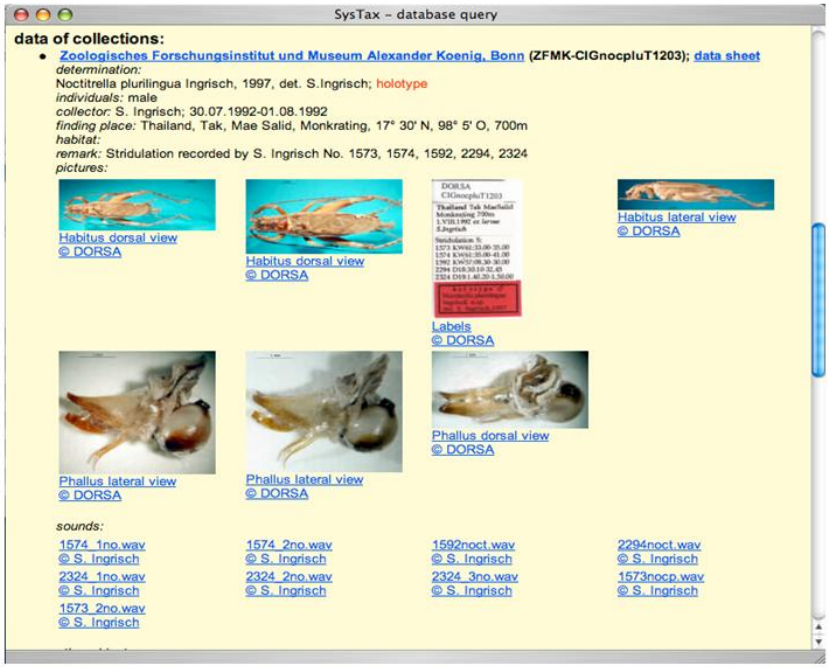


Fig. 4: Reciprocal linking between DORSA and OSF

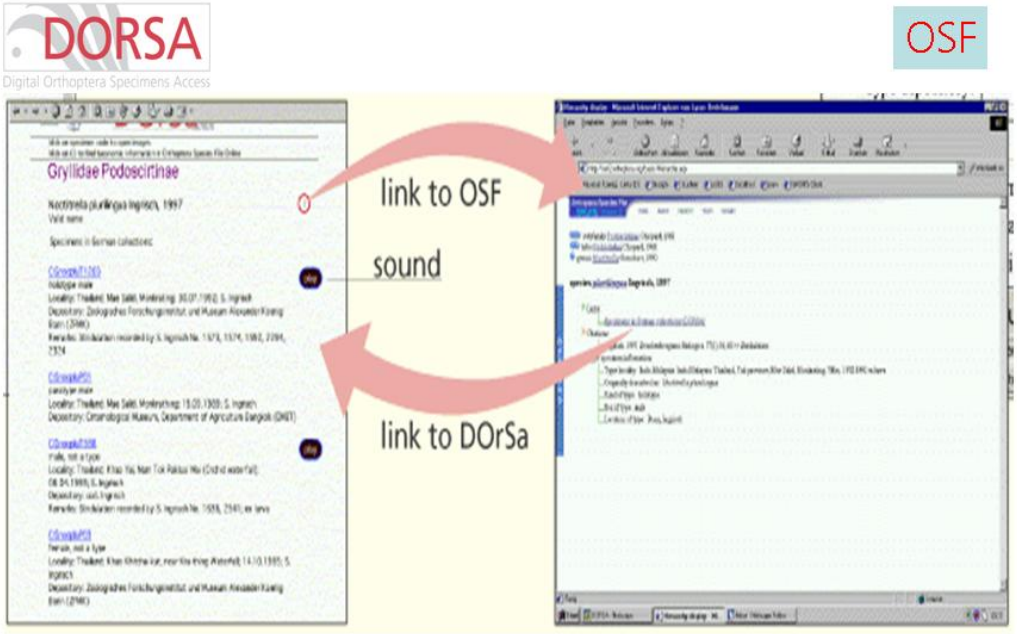


Table 1: Comparison of primary types in OSF and DORSA (reciprocal verification)

OSF DORSA Digital Orthoptera Specimens Access

	presumption	museum data	validity check				
museum	taxa with primary types in OSF *	taxa with primary types checked	type data in OSF not confirmed in museum	primary types lost: taxa	primary types not listed in OSF	no taxon entry found in OSF	unlabeled, newly recognized prim. types
Berlin	1093	*1272*	*15*		*221*	*68*	*12*
Eberswalde	45	79		1	34		
Dresden	66	117	6		55	10	7
Hamburg	142	135	5	55	43	1	
Halle	38	55	4		25	4	16
Bonn	6	44			12	13	
Frankfurt	82	151			69	5	1
Stuttgart	47	112			65	5	3
München	27	54		1	21	1	
Sum	1546	2019	30	57	545	107	39

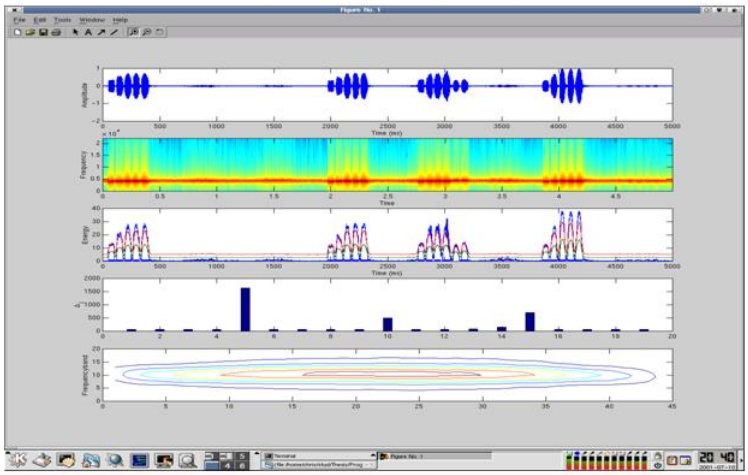
discrepancies between the two sources, such as 545 primary types, which had not been listed in OSF or 39 unlabeled, newly recognized primary type specimens within German museums. Beside that exact data about missing type specimens allow initiation of effective searches to fill these gaps, for example by designating neotypes.

The DORSA project was complemented by Christian Dietrich's thesis on automatic identification of cricket songs through neural networks, in cooperation with the informatics department of Ulm University (Prof. Palm). The tool he developed can be used in rapid assessment programs as a non-invasive technique to classify and map acoustic diversity in the field. 'MatLab' software was used for neural network programming, and the first steps consisted in extraction of basic song parameters, such as carrier frequency and pulse rate *(Fig. 5)*. This software module was applied to the entire DORSA sound repository and the calculated parameters were then annotated into new columns of the sound file table (semantic enrichment). As a result important sound parameters now complement the respective sound files. Sound parameters can be evaluated, or you can search for certain parameters by standard SQL queries,

Fig. 5: Semantic enrichment of specimen based databases

DORSA
Digital Orthoptera Specimens Access

Extraction of sound parameters by using MatLab Software



Carrier frequency

Pulse rate

Carrier frequency

In cooperation with: Dept of Neuroinformatics, Ulm University; PhD thesis C. Dietrich

and retrieve the respective sound files from the repository. Even complex data-mining is feasible, such as: "Are there any Orthoptera (in our database), from Ecuadorian lowland forests, and singing higher than 10 kHz, etc."

Because gathering sites of DORSA specimens became geo-referenced by adding latitude and longitude information, type localities can be visualised by linking to remote available map-servers such as the 'Canadian Biodiversity Information Facility' (CBIF) map server or Google Earth. Using 'Geographical Information Systems' (GIS) distribution maps can be created and intersected with layers such as political boundaries or vegetation zones or recent distribution of tropical rain forests etc. Extraction of special geographical information units such as 'county' or 'district' data allow a geographical semantic enrichment of the respective DORSA files; quite similar to the already mentioned enrichment of sound files. Beside that verification of geographical DORSA data against external geographical authority files can be done in a similar manner to the systematic verification of DORSA data against OSF data.

Another advantage of geo-referencing specimen data is that one can easily create a map illustrating the countries of origin of type specimens *(Fig. 6)*. As you can see most type specimens in German collections were collected in tropical countries reflecting Germany's scientific expeditions and colonial history. Publishing type specimen data and images through a virtual museum signifies repatriation of knowledge for countries of type origin.

Fig. 7 shows the present online status of our collection management and information system BIODAT, a specimen based database used in my home institution 'Zoologisches Forschungsmuseum Alexander Koenig' and in the "Museum für Naturkunde" of the Humboldt University in Berlin. BIODAT specimen data and related images are online available (http://www.biodat.de) via BIODAT's one field query tool and via the BIOCASE- and GBIF-portals. The one field query tool was developed by BIODAT's programmer, Dirk Striebing. It allows a 'Google-like' easy access to fine grained information about specimens housed in our collections. By selecting 'preferences' the user can decide what kind of information is represented as well as the number of results on display. The one field query can consider different kinds of information, such as collector and/or gathering site and systematic information (order, family, genus etc.). The user has direct access to various global authority files such as taxonomic and biographic name servers as well as geographical map servers by dynamic linking.

Perspectives

In view of the enormous ecological and economical problems of the world there is an urgent need to understand biodiversity in context of complex systems and to make knowledge directly available.

Today we have the technology to realize the vision of a vertical information transfer within biological science. That means an information transfer between the molecular, the organism and the ecosystem level and vice versa. A lot of working groups and international activities such as TDWG (Taxonomic Databases Working Group), BIOCASE and GBIF deal with this topic by developing protocols, schemas and other helpful tools. For a scientist today it is not sufficient to be an expert who can best talk to his nearest neighbour. There is a need to share knowledge and

expertise, to make it understandable. As a result the scientific & public demand becomes more and more focussed on transdisciplinarity instead of interdisciplinarity. Another challenge will therefore be a horizontal information transfer between biological and other domains such as libraries, archives etc. Buzz words like 'semantic web' underline the need of linking biodiversity informatics, geo-informatics and finally the wide area of cultural informatics.

Probably the next generation of data-basing will move from a class-centric to an event-centric approach. For example a virtual, worldwide catalogue of collecting events could serve as a backbone for an accelerated data capture of full collecting information by improving data quality (solving problems like identity of persons, ideas, places, objects, and concepts; allowing multiple verification of data, etc.). Thereby GRID technology should enable us to merge local, domain specific and global authority files (via web services).

Fig. 6: Countries of origin of Orthoptera types in German collections DORSA Digital Orthoptera Specimens Access

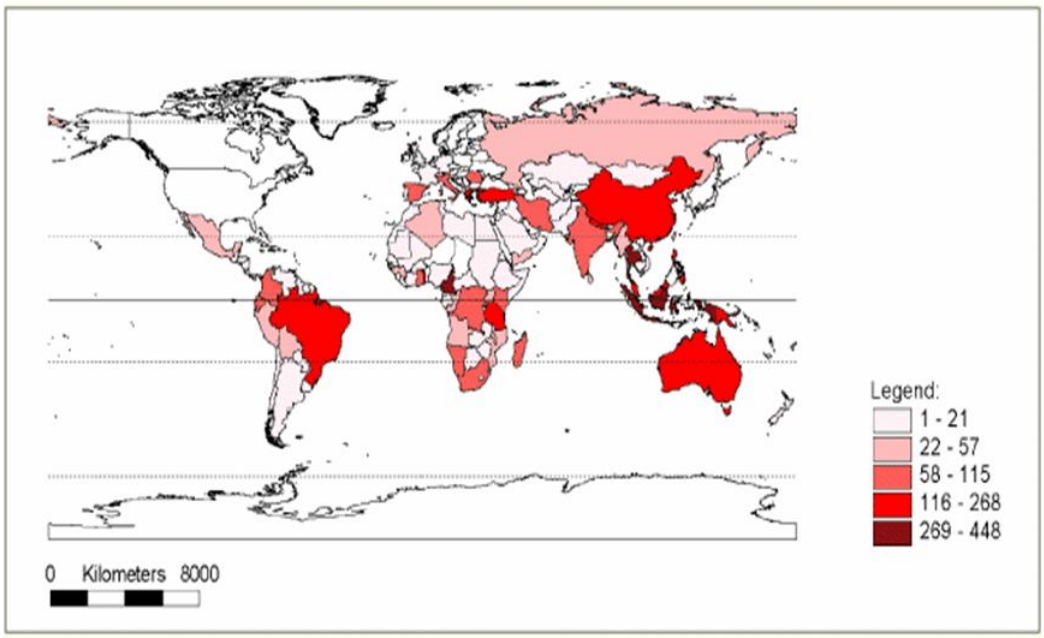
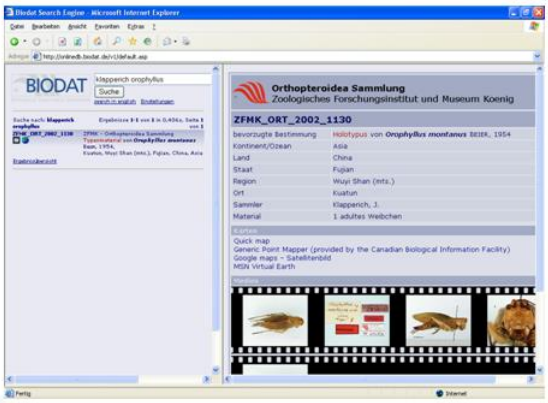
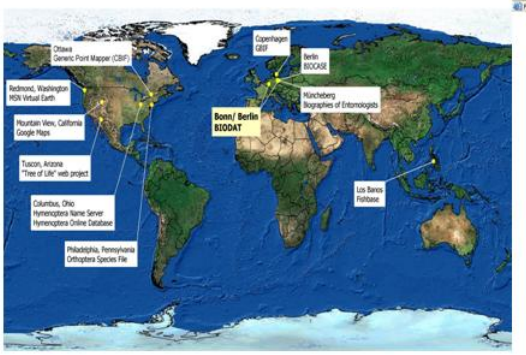


Fig. 7: BIODAT online

BIODAT

'Google-like' one-field query tool





Data capture in Bonn and Berlin

Data provider for BIOCASE and GBIF

Dynamic links to various global authority files

Acknowledgements

- Dr Martin Doerr, ICS Forth, Heraklion/Crete, Greece and Stephen Stead, Paveprime Ltd, London/UK for checking the Ontologies.
- Dr Klaus Riede & Dr Sigfrid Ingrisch, ZFMK, and the DORSA data capture team and curators responsible for Orthoptera in the museums in Berlin, Eberswalde, Dresden, Hamburg, Halle, Bonn, Frankfurt, Stuttgart and Munich.
- Dr Christian Dietrich, Dept. of Neuroinformatics, Ulm University
- Dr Brad Sinclair and the DIG-team, ZFMK
- Birgit Rach, Dirk Rohwedder and the BIODAT- & BIOTA E15- data capture teams, ZFMK
- The German Ministry of Education and Research (BMBF) for funding the DORSA-, DIG- & BIOTA East Africa E15 projects.