

International Networking of Large Amounts of Primary Biodiversity Data

J. Holetschek, P. Kelbert, A. Müller, P. Ciardelli, A. Güntsch, W.G. Berendsohn

Dept. of Biodiversity Informatics and Laboratories
Botanic Garden and Botanical Museum Berlin-Dahlem
Königin-Luise-Str. 6-8
D-14195 Berlin-Dahlem

{j.holetschek, p.kelbert, a.mueller, p.ciardelli, a.guentsch, w.berendsohn}@bgbm.org

Abstract: Primary Biodiversity Data is a general term for information documenting the planet's biodiversity, where each record represents the existence of a particular organism at a given location at a point in time. These data are scattered throughout numerous collections and databases worldwide, making it difficult to find all information available on, for instance, a certain species or a particular region. Several international networks and initiatives share a vision of free and open access to these resources and are working together to connect these heterogeneous data sources.

This article provides an overview of the basic system architecture of these networks and some of the efforts aimed at solving the issues arising from the explosion of the amount of biodiversity data in recent years. It discusses the concept of "special interest networks" aimed at facilitating the setup of thematic, user- or subject-specific biodiversity data portals; describes SYNTHESYS' data portal software and its query expansion mechanism; and finally, outlines the duplicate detection system for identifying possible duplicate records in biodiversity networks.

1 Introduction

Over the past ten years, international initiatives such as the Global Biodiversity Information Facility (GBIF¹) and the Biological Collection Access Service for Europe (BioCASE²) have set up an infrastructure for connecting primary biodiversity data from various distributed sources. The goal is to make the existing data on planetary biodiversity freely and universally available for everyone on the Internet, to be used for research and as an information resource for decision makers and nature conservationists [GBIFa].

¹ <http://www.gbif.org>

² <http://www.biocase.org>

The information in question is primary biodiversity data, i.e. direct observations of organisms in nature, or of specimens either preserved or cultivated. It embraces the billions of specimens from all organism groups in preserved and living collections worldwide as well as the billions of observations that have been recorded, e.g. from vegetation surveys or bird monitoring. Currently, GBIF provides access to nearly 175 million records in 7,445 datasets from 285 different data providers [GBIFb].

As the amount of providers connected and the number of records grow steadily, several problems arise for these networks. The SYNTHESYS project has responded to this trend with activities intended to overcome some of the negative effects of growth: development of the concept of thematic, user- or subject-specific networks and consequent development of generic software for specialised data portals, and duplicate detection tools used to improve the quality of the underlying data. After outlining the basic architecture of the BioCASE network, this article will provide an overview of these activities.

2 Basic Technical Architecture

GBIF and BioCASE use the same XML access protocols and XML data standards for a distributed information infrastructure with multiple provider and portal nodes (see figure 1, [GB07]).

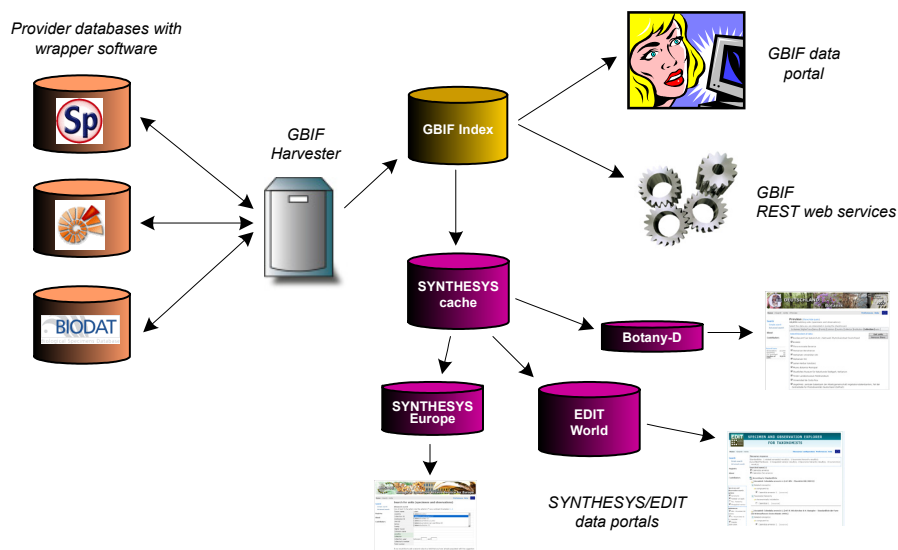


Fig. 1: Basic technical architecture of the BioCASE/GBIF network.

The original data are maintained in the provider's database, and remain so after being connected to the GBIF network, ensuring that the original providers retain full control over the information they publish. This is independent of the provider's database or information management system, of which the network includes a great variety: off-the-shelf collection management systems, custom management systems e.g. for species observation records, server databases or even single user database files or applications. Provider software installed on a web server establishes an XML-based interface to the network [DTG03], using an accepted data standard – in GBIF and BioCASE currently Darwin Core [TDWGa] or ABCD (Access to Biological Collection Data, [TDWGb]) – and one of the prescribed access protocols – DiGIR (Distributed Generic Information Retrieval, [An06]), BioCASE (Biological Collection Access Service, [Dö04]) or their improved successor TAPIR (TDWG Access Protocol for Information Retrieval, [Dö06], [DG05]). Several standards-compliant provider software packages have been implemented; some support only one protocol or data standard, with others support several.

The provider software accepts protocol-compliant queries, converts the requests into the native query language of the connected provider database, queries the database, transforms the resulting record set into a standard-compliant response document, and finally sends the document back to the querying node. Depending on the wrapper software used, the sort of possible requests range from simple capabilities information or inventory lists to complete and detailed data records.

GBIF uses this interface to harvest the information stored in the database, i.e. to build an inventory of all records connected to the network, along with the core information associated with these records: taxonomic tags, data on the geographic origin and available geo-referencing (particularly geographic coordinates), collector's name, gathering dates and existing multimedia documents of the record. This inventory is called the GBIF index, because it is a list of all records and data needed for quick searching rather than the full and detailed records stored in a central database.

This inventory can be accessed in two ways. Human users can use the GBIF data portal [GBIFa] to find data within the network and to view or download all the records of interest, while the same data is offered by GBIF's REST web services [GBIFb] and can be used directly by other projects or initiatives for a variety of biodiversity informatics applications.

The SYNTHESYS project is one of the initiatives implementing such technologies. The SYNTHESYS cache generator (see next chapter) extracts taxonomic or geographic-specific subsets of the GBIF index and stores them in local repositories, referred to as SYNTHESYS caches. The SYNTHESYS data portals (chapter 4) allow users to access these subsets, combining data from GBIF with data from different thematic thesauri in order to extend and improve search functionality. Additionally, the SYNTHESYS duplicate detection tool can be used to improve the quality of the data used by the network (chapter 5).

3 The SYNTHESYS cache generation system

The ever-growing amount of data available in the GBIF network makes it increasingly difficult for users to find information relevant to them. There are often thousands of records available, and for some species there are more than a million records in the GBIF index. Finding the records of interest can be like looking for a needle in a haystack, even with advanced search functionality [Ku08].

To relieve this situation, the SYNTHESYS project and the German GBIF node have been working together on a generic system for specialised search portals for special interest groups (e.g. for a certain taxonomic group or for a specific type of user) or regional organizations working on restricted geographic areas. In addition to filtering out large amounts of irrelevant data, this system offers these groups the opportunity to use their resources to enhance the value of the data by adding information drawn from additional sources such as regional or group-specific taxonomic thesauri, local geographic services or translation mechanisms [Ke09].

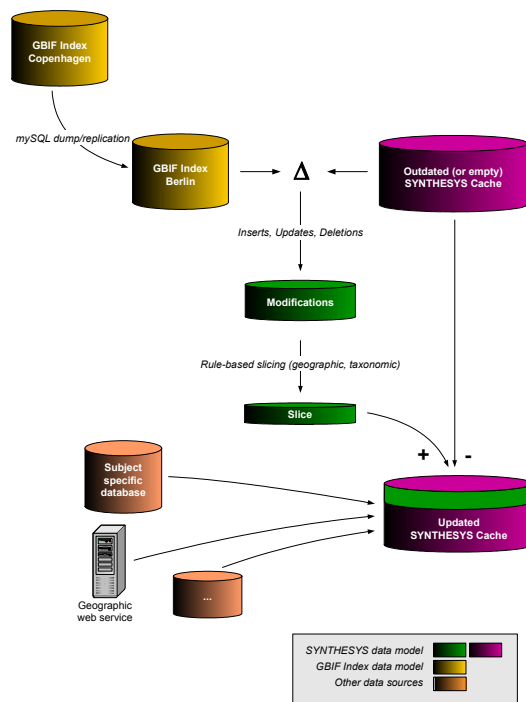


Fig. 2: The SYNTHESYS cache generation system.

The SYNTHESYS cache is based on the data in the GBIF index; the first prototype used the index at the Berlin mirror of the GBIF data portal [Ho06]. The process of stage generation and updates is shown in fig. 2, and comprises the following five steps:

1. Detection of modifications: In this first step, a list of changes (deletions, inserts and updates) is created by comparing the current GBIF index database with status information collected from the previous version.

When the cache generation system was conceived in 2005, the Berlin GBIF mirror was kept in sync with the Copenhagen index database using MySQL replication, which constantly updated the mirror as the original data sources were re-indexed by GBIF. The cache generator detects changes in the index database and restricts the subsequent processing steps to the data affected – either directly or via dependencies – rather than repeatedly processing the entire dataset. Otherwise, the load for the database server would have been too high for nightly updates..

The indexing procedure introduced by GBIF for the new data portal in 2007 is incompatible with the use of MySQL replication as a synchronizing approach. In its place, regular MySQL dumps are now used instead, meaning that at present the full database gets recognized as having been updated by the cache generator.

2. Conversion of data into the SYNTHESYS data model: The first version of the GBIF index database was used to directly store data from the indexing process; moreover it included the entire taxonomic tree used by GBIF. SYNTHESYS uses an alternate approach to link taxonomic information to occurrence data (see chapter 4), so that its data model can be optimised for query performance in the data portal. This step converts the data affected by the modifications between these different data models and creates a “delta cache”.

3. Slicing: At this stage, the index is cut down to the desired subject of the network. The filtering rules use different criteria (taxa, country codes, geographic coordinates, regional place or area names, collection metadata) or combinations and are used to create a “merge list” based on the delta version of the SYNTHESYS cache created in the previous step. If the cache generator is used to produce several caches on different subjects, a merge list is created for each one.

4. Merging: In this step, the merge list of each cache is applied to the delta cache, and the resulting data are merged into the respective SYNTHESYS cache. Deleted records are removed, and updated or inserted records are replaced or added.

5. Augmenting data: This optional step involves augmentation by adding data drawn from specialized information sources. Also, the expertise of the cache creators can be used to add missing data to certain datasets in the cache.

Currently, the SYNTHESYS cache system is used for five different specialised data portals. The EDIT Specimen and Observation explorer³ is tailored to taxonomist’s needs by using a taxonomic thesaurus system which extends user queries to synonyms, related taxonomic concepts such as misapplied names and related taxa in the taxonomic hierarchy [Ke08]. The SYNTHESYS European data portal⁴ offers access to European speci-

³ <http://search.biocase.org/edit>

⁴ <http://search.biocase.org/europe>

men and observation data; the search can be extended using several European checklists [Ho09]. The portal of the German GBIF node for botany⁵ offers information on German flora, linked to the standard lists of plants available for the German flora [Ki07]. The BGBM data portal⁶ provides access to specimen data of the Botanic Garden and Botanical Museum Berlin-Dahlem. Finally, the Central African Biodiversity Information Network portals⁷ (CABIN, see figure 3) of the Royal Museum for Central Africa, Tervuren, Belgium, and of the CEDESURK, Documentation and Library system of the University of Kinshasa, DR Congo, focus on African biodiversity.

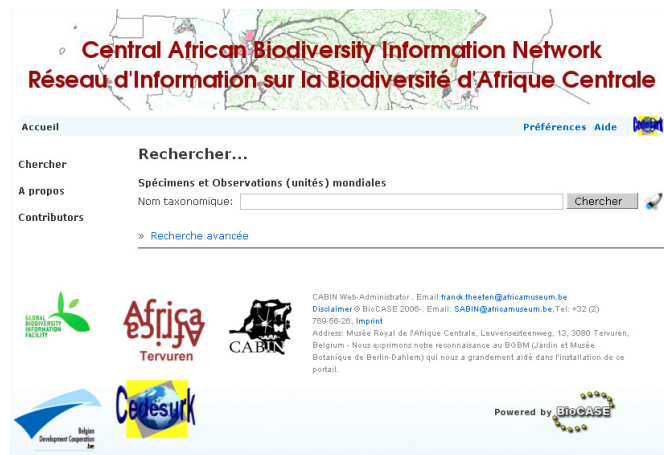


Fig. 3: The CABIN data portal

4 The SYNTHESYS data portals

In addition to a mechanism for deriving slices from the global index, a useful platform for building thematic networks requires a data portal that can be easily installed, configured and customized to the needs of the given network. For this purpose, the SYNTHESYS network has developed a Python-based portal application (see figure 4).

The portal can be used with a SYNTHESYS cache kept in a MySQL or SQL Server database, so it can be implemented in a completely open-source environment or, in case better performance is required based on projected amounts of data and users, in conjunction with a more efficient commercial database management system [Ho08]. Portal layout can be controlled with cascading style sheets. The application supports multiple languages and currently ships translations into eleven languages, including Chinese.

⁵ <http://search.biocase.de/botany>

⁶ <http://search.biocase.org/bgbm>

⁷ <http://cabin.ebale.cd>

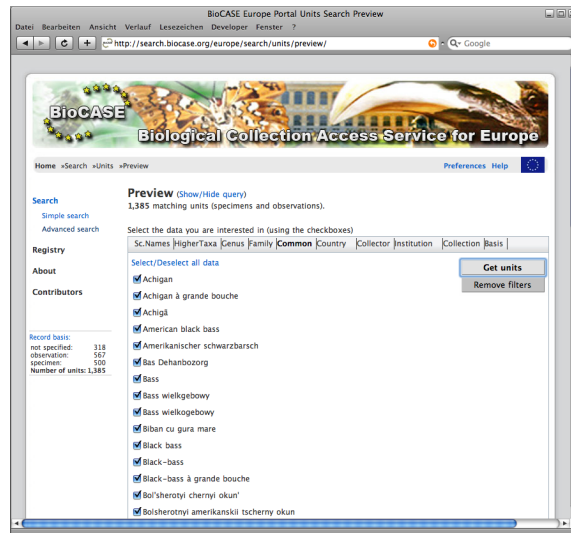


Fig. 4: The SYNTHESYS data portal for Europe.

In order to improve search results, user queries can be extended to include related concepts and synonyms. The query expansion process uses Thesaurus Optimized Query Expansion (TOQE) services developed in the SYNTHESYS project. TOQE provides a generic XML-based interface to thesaurus databases including taxonomic checklists, gazetteers, person names, and stratigraphic term lists [Hf07], such that the necessary domain knowledge for specific portal implementations can be drawn from specialized thesaurus databases and does not have to be imported into the SYNTHESYS cache.

TOQE-services are typically used in two steps. The first step queries the thesaurus dataset and retrieves the set of matching concepts for a given term in the original query (e.g. a taxonomic name or a country name). In a second step, the client retrieves the set of related concepts and their associated terms, which are then used to expand the original query.

A configurable Python-based TOQE-service implementation has been developed and deployed. TOQE-methods are called using simple REST-like get-requests. The TOQE schema defines the format of corresponding XML response documents⁸.

In order to ease the process of installing and setting up the portal, setup and configuration software has been developed which checks required packages and lists missing modules, installs the portal application, connects to the underlying SYNTHESYS cache, and potentially relevant thesauri. Moreover, several options for configuring the portal's behavior are available, including email notifications for the portal's administrator in case of errors.

⁸ <http://search.biocase.org/toqe/schema/>

5 Duplicate detection system

Currently, the GBIF network collates data from 7,445 collections worldwide. The detection of duplicate records becomes a matter of utmost importance; at best, duplicate records are an annoyance for a user of the data portals, but at worst, they can hinder the user from finding the information he or she is interested in.

SYNTHE-SYS has developed a tool for detecting duplicate records within a SYNTHE-SYS cache database. The main purpose of this tool is to enable researchers to complete and refine their information about data they work on, or in other cases, to enable collection holders to identify possible duplicate records on the network stemming from other sources, offering the opportunity to improve data quality by eliminating duplicates.

5.1 Specimen Duplicates

For specimen records two types of duplicates, digital and physical, must be considered:

- A digital duplicate refers to multiple digital records referencing the same physical specimen. Due to transformation processes on their way from the original field data into the GBIF index, records referring to the same object may differ in missing attributes, sharpness of values (e.g. location data), slightly varying attributes (due to typos or standardisation) and additional attributes (data added later).
- In contrast, physical duplicates are two or more specimens derived from the same original material. In this case, separate physical objects – usually kept in different collections – exist. Physical duplicates have more variation in their attributes than digital duplicates, not only because of the difference in the accession data, but also because additional data may be standardised and added differently at the different locations.
However, attributes such as the collector's name, the date of gathering or geographical coordinates should be similar for physical duplicates, since they refer to the same original data.

When searching for specimen duplicates, the duplicate type must thus be taken into account; for instance, different accession data in case of a digital duplicate are a strong indication that two specimen records are not duplicates, but should not be taken into account when computing the probability for an existing physical duplication.

5.2 Components

Identifying likely duplicate records is often referred to in the literature as record linkage. Record linkage generally involves comparing records and then deciding whether they are “link”, “non-link” or “possible-link” records. This process of linking records consists of several steps: standardisation, blocking, comparison and classification (see fig. 5).

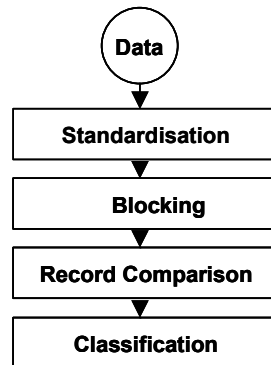


Fig. 5: Process of linking records

Standardisation: Information can be stored differently in different data sources. In order to make data comparable, it must first be standardised. For example, unless the name of the country is in the same language for all records, records with “Germany” and “Deutschland” in the country field will not be identified as duplicates.

Blocking: Measuring similarity commonly involves calculating a full similarity matrix for every pair of records, resulting in a time complexity of at least $O(n^2)$. Preselecting or “blocking” potential records likely to be similar has proven to be a successful way of reducing the complexity of this algorithm. Potentially matching records are grouped together, such that the huge number of potential comparisons is reduced by comparing records only with those in their group.

Blocking mechanisms include:

- *Multi-pass sorted neighbourhood:* records are inserted into sorted indexes with a high probability of storing duplicates close to each other. The multi-pass approach attempts to reduce the number of false negatives where duplicates are not found in close proximity by using different blocking keys.
- *Bigrams or q-grams:* blocking key values are converted into a list of bigrams/q-grams and lists of all possible permutations are built using a threshold. The resulting bigram/q-grams lists are sorted and inserted into an inverted index, which is then used to retrieve the corresponding record numbers in a block.
- *Canopy clustering:* this mechanism is intended to speed up clustering operations on large data sets. It consists of 2 stages: partitioning the data into overlapping subsets, called “canopies“, and then performing more expensive clustering solely within these canopies.

For a full comparison of blocking techniques see [BC03].

Comparison: Records are usually compared on a field by field basis. As fields differ in their data types and as some fields may be more distinctive than others, each attribute can have its own comparison function and the result may be weighted individually. For each pair of records, these field comparison functions return a basic distance for the relevant attributes. The results of all field comparisons are combined into a comparison vector.

Numerous comparison functions exist for strings, numbers, dates, ages and times. For approximate string matches, different methods can be used:

- *Jaro* is based on the number of common characters in two strings, the lengths of both strings and the number of transpositions.
- *Winkler* is based on the Jaro comparator, but takes into account the fact that typographical errors occur more often towards the end of words.
- *Edit distance* is based on the minimum number of character deletions, transpositions and insertions that must be made to transform one string into the other.
- *Bigram* is based on the number of common bigrams in the two strings.

The weighted comparison vectors are then passed to the decision model.

Classification: The decision model, also called the classifier, calculates a matching decision based on the comparison vector. The system sums up all weighted coordinates of the vector and uses thresholds to categorize the results as “link”, “non-link” or “possible link”.

5.3 Specimen Duplicate Detection Tool

The SYNTHESYS tool for specimen duplicate detection is based on FEBRL (Freely Extensible Biomedical Record Linkage, [CC04]; see also [DM07]), a Python software developed for record linkage investigations. It allows the selection of various algorithms at all stages and was developed by the Australian National University to test the performance of new algorithms. A number of changes were necessary to adapt the tool to the needs of specimen duplicate detection.

Standardisation: GBIF already does a great deal of normalization when indexing data sources; most attributes are standardized after harvesting. Nevertheless, attributes such as Latin names must undergo a second standardization process to reduce the non-match probability caused by misspelling and typographical errors.

Blocking: Without blocking, the duration of the process would be intolerable. In case of the GBIF index with more than 170 million records a complete comparison matrix would consist of more than 1016 independent similarity calculations. Assuming a cost of 0.1ms per calculation, it would take more than 70.000 years for calculations; one record alone would require more than 4 hours to compute all comparisons.

For blocking, a multi-channel sorted neighbourhood mechanism has been used that sorts all records multiple times, using newly generated strings concatenating several split or truncated record fields. The strings are stored in a fast-to-query index. To find duplicate candidates for a record, its n closest neighbours are taken from each of the indexes, with n being a predefined window size.

Creating the indexes is still extremely time consuming (in our case it took 4 days to compute 19 indexes for 171 million records), but need only be done once at the outset. In contrast, the selection of duplicate candidates from indexes is performed on the fly during the duplicate detection query and therefore extremely quick; SQL server's clustered indexes were used to retrieve the information with minimum cost [Ho08].

Comparison: For comparison, the more commonly populated attributes such as scientific names, gathering information (locality, geo-coordinates, altitude, collector) and metadata (information about the collection or collection holder) were chosen. We selected the Winkler method for scientific names, as it improves the probability of detecting duplicates arising from the misspelled names and typographical errors which frequently occur in these names. For the other attributes such as locality and collection name, we opted for the edit-distance comparison method.

For comparison, FEBRL takes into account the probabilities for equal attribute values in matching and in non-matching records to compute the comparison function for each attribute. In addition, a default missing weight for non-existing values can be defined.

Classification: FEBRL supports the decision model according to [FS69], which sums up all the weights in a weight-vector and uses the given-threshold to classify a record pair in one of three classes: link, possible-link, non-link.

In contrast to FEBRL and to classical record linkage systems, the specimen duplicate detection tool only distinguishes between two classification results: possible link and non-link; it is up to the user to decide if a possible link is a duplicate or not. Those records classified as possible links are returned as list of results ordered by duplicate probability.

5.4 Pilot implementation

As a pilot implementation, we developed a simple web interface⁹ to search for duplicates. Only attributes available in the SYNTHESYS cache [Ho06] can be submitted as parameters to the search. Fields such as scientific name, collector's number, locality, coordinates and gathering date can be populated for the search. The pilot implementation will be used for testing and adjusting the algorithm to the needs of users of primary occurrence data. As a next step, we will assess potential implementations of the algorithm into existing biodiversity informatics applications such as query portals, harvesting systems and web services.

⁹ <http://search.biocase.org/duplicate>

6 Conclusion and outlook

International biodiversity information infrastructures such as GBIF and BioCASE have developed a flexible service-based infrastructure for networking of biodiversity data with a focus on the provision of collection and observation data. The central goal of the first 10 years of network construction was data mobilization: open and free access to as much primary information as possible from as many biodiversity databases worldwide as possible. The EU project SYNTHESYS contributed to this endeavor by providing prototype tools for duplicate detection supporting data interpretation and quality measures, as well as a software suite for setting up specialized thematic portals and networks based on a subset of global information.

In its second phase, SYNTHESYS will shift its attention from data provision related developments to the implementation of mechanisms for information feedback from data users to the original provider. This will include continued development of the generic annotation system prototyped by SYNTHESYS, as well as the development of a “reverse wrapper” allowing data providers to automatically import accepted annotations into their databases.

References

- [An05] Anonymous: The Distributed Generic Information Retrieval Protocol (DiGIR). <http://digir.sourceforge.net/>, 2005.
- [BC03] Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: Proceedings of the Workshop on Data Cleaning, Record Linkage and Object Consolidation at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.
- [CC04] P. Christen, T. Churches, and M. Hegland.: A Parallel Open Source Data Linkage System. In Springer Lecture Notes in Artificial Intelligence, Sydney, Australia, May 2004.
- [DG05] Döring, M & de Giovanni, R.: Make the tapir work – Practical potential of the TDWG Access Protocol for Information Retrieval (TAPIR). 21st TDWG meeting, St. Petersburg, Russia, September 2005.
- [Dö04] Döring, M.; de Giovanni, R.; Hobern, D.; Vieglais, D.; Güntsch, A.; Blum, S.; Wieczorek, J. & de la Torre, J: The integration of DiGIR and BioCASE. 20th TDWG meeting, Christchurch, New Zealand, October 2004.
- [Dö06] Döring, M: Using TAPIR in biodiversity networks. 22nd TDWG meeting, St. Louis, USA, October 2006.
- [DM07] Döring, M & Müller, A: Analysis of existing software and methods to build a duplicate detection software for the GBIF cache database. In http://www.synthesys.info/NA_Documents/Deliverables/NA_D/D1_3.1_Analysis_of_duplicate_detection_software.doc Berlin, 2007.
- [DTG03] Döring, M.; de la Torre, J.; Güntsch, A.: Technical introduction to the BioCASE software modules. 19th TDWG meeting, Oeiras, Lisbon, Portugal, November 2003.

- [FS69] Fellegi, I.P., Sunter, A.B.: A theory for record linkage. *Journal of the American Statistical Association*. Issue 64, pp. 1183–1210, 1969.
- [GB07] Güntsch, A. & Berendsohn, W.G.: Networking distributed biodiversity data with GBIF and BioCASE. 6th Meeting on Vegetation Databases – Surveys of vegetation and floras – new prospects, joint ways. Bonn 2007.
- [GBIFa] The Global Biodiversity Information Facility: Memorandum of Understanding 2007-11. <http://www2.gbif.org/mou07-11.pdf>, Copenhagen, Denmark 2007.
- [GBIFb] The Global Biodiversity Information Facility: GBIF Data Portal. <http://data.gbif.org>, Copenhagen, Denmark, 2009 [accessed June 17].
- [GBIFc] The Global Biodiversity Information Facility: GBIF Web services. <http://data.gbif.org/ws>, Copenhagen, Denmark, 2009.
- [Hf07] Hoffmann, N.; Kelbert, P.; Ciardelli, P. & Güntsch, A.: TOQE - A Thesaurus Optimized Query Expander. Proceedings of the TDWG annual meeting, Bratislava, Slovakia, 2007.
- [Ho06] Holetschek, J.; Güntsch, A.; Oancea, C.; Döring, M.; Berendsohn, W. G.: Prototyping a Generic Slice Generation System for the GBIF Index. Pp. 51-52 in Belbin, L., Rissoné, A. and Weitzman, A. (eds.). Proceedings of TDWG, St Louis, MI, 2006.
- [Ho08] Holetschek, Jörg: How to Reduce Your Coffee Consumption – Performance Investigations on Large Occurrence Databases. Pp. 85-86 in Weitzman, A.L., and Belbin, L. (eds.). Proceedings of TDWG, Fremantle, Australia, 2008.
- [Ho09] Holetschek, J.; Kelbert, P.; Güntsch, A.; Kusber, W.-H.; Zippel, E. & Berendsohn, W.G.: The SYNTHESYS Specimen and Observation Portal. eBiosphere Conference, London 2009 (in print).
- [Ke08] Kelbert, P., Hoffmann, N., Holetschek, J., Güntsch, A., Berendsohn, W. G. 2008. The new EDIT Specimen and Observation Explorer for Taxonomists. In: Weitzman, A.L., and Belbin, L. (eds.). Proceedings of TDWG (2008), Fremantle, Australia, 2008.
- [Ke09] Kelbert, P.; Holetschek, J.; Güntsch, A.; Kusber, W.-H.; Zippel, E.; Müller, A. & Berendsohn, W. G.: Using the GBIF Infrastructure to set up Special Interest Networks. eBiosphere Conference, London 2009.
- [Ki07] Kirchhoff, A., Holetschek, J., Hahn, A., Kelbert, P., Jahn, R., Berendsohn W.G.: GBIF Germany and the Botanical Node. 1st central european diatom meeting, Berlin-Dahlem, Germany, 23-25 March 2007.
- [Ku08] Kusber, W.-H., Zippel, E., Kelbert, P., Holetschek, J., Hahn, A., Güntsch, A. & Berendsohn, W.G. 2008: From cleaning the valves to cleaning the data: European diatom biodiversity data on the Internet. P. 30 in Cantonati, M., Scalfi, A. & Bertuzzi, E. (ed.): 2nd Central European Diatom Meeting, Abstract Book, Trento (Italy), 2008.
- [TDWGa] The Taxonomic Databases Working Group: The Resource Directory for DarwinCore (DwC). <http://rs.tdwg.org/dwc/>, 2007.
- [TDWGb] The Taxonomic Databases Working Group: The Access to Biological Collection Data (ABCD) Standard 2.06. <http://rs.tdwg.org/abcd/>, 2007.