Proposal
Leibniz Collaborative Excellence

Digital Approaches for the Synthesis of Poorly Accessible Biodiversity Information

---

DiASPora

Applicant
Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen

Project leader
Prof. Dr. Jörg Overmann

A) Quality and innovativeness of the research project

Summary

The digitalization and integration of biodiversity information can generate substantial added value for existing data and yield novel scientific insights of relevance to bioeconomy, biotechnology, human health, and environmental protection. So far this potential has been exploited only rarely due the heterogeneity and fragmentation of data sources, and the little documentation, variable standards, and limited interoperability of data. For bacteria, research data are particularly diverse and broadly distributed; therefore these organisms will serve as the model group for the current project. The project *DiASPora* will establish an approach for synthesizing information for bacterial species by applying state-of-the-art data science methodology, genomics, and developing user-centric workflows. Extraction of phenotypic data from the microbiological literature will be achieved by large-scale text mining, applying artificial intelligence (AI) techniques that will be trained through the feedback of microbiologist curators. The data recovered will be hosted by the existing Bac*Dive* database and transformed into a machine readable and processable format using the Resource Description Framework (RDF). Subsequently, the transformed data will be used to establish a knowledge graph to generate innovative search options for the discovery of hidden data relationships. In parallel, phenotypic predictions will be derived from (meta)genomic data, through the application of metabolic models and comparison with the physiological and habitat data as obtained by data mining, and will be supported by an AI approach. The project is dedicated to an integral community engagement and an efficient dissemination of results. *DiASPora* builds upon the complementary expertise of three participating institutions, covering the fields of microbial databases and diversity research, bacterial genomics, text mining, artificial intelligence, and semantic technologies.

Introduction

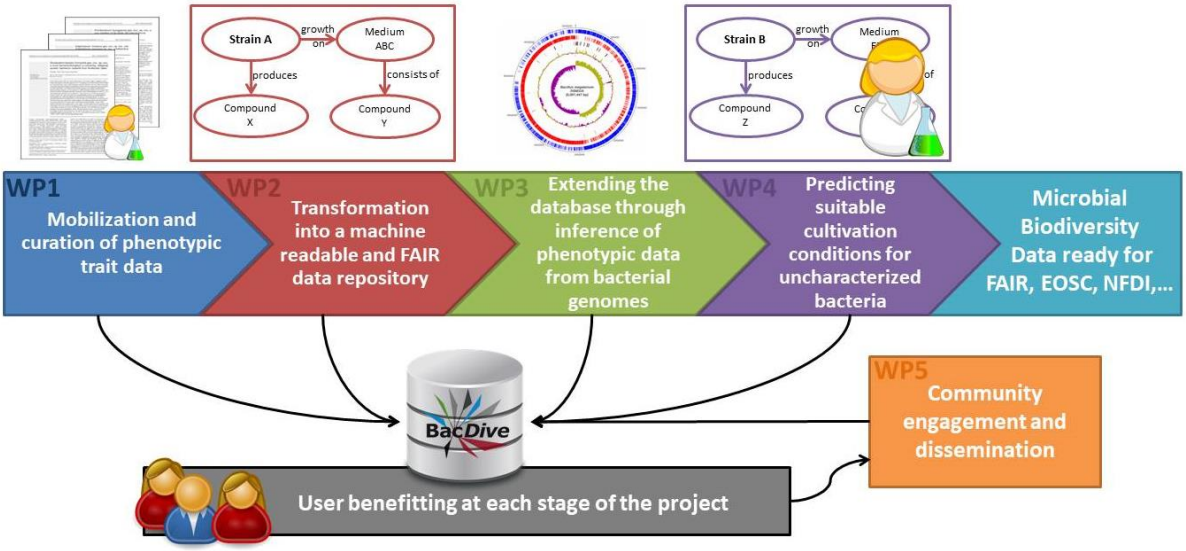
The transformation of scientific data and methods through digitalization is a key driver of future research and innovation [1]. Substantial value can be added through the systematic integration of information, but so far this potential has rarely been exploited due to the heterogeneity and fragmentation of data sources, little documentation, large variability or even lack of standards, and the limited interoperability of the available data [1]. In life sciences, data originate from observations, experiments or modelling, and range from molecular sequences, phenotypic data, over taxonomic frameworks, to environmental sensor readings. These results are captured in unstructured publications, or are distributed across different information systems and databases. Current data searches often rely on keywords, and a broad set of interlinked keywords connected to the topic of interest has to be queried. The logical connection of the different data silos is a precondition for integrating the existing information but so far is slow, cumbersome, and requires highly specialized knowledge. In addition, most data analyses are conducted in a discipline-specific way.

Novel approaches of data integration are particularly promising for microorganisms as these represent the largest reservoir of genetic information on Earth [2,3] and offer many novel solutions for bioeconomy, biotechnology, human health, and environmental protection [4,5,6,7]. So far, only about 15,000 species of *Bacteria* and *Archaea*, representing 0.1 - 0.001% of the estimated total have been described. Even for these species, phenotypic data were not digitally accessible until recently, due to their distribution across various heterogeneous sources and various formats. Over the past years, *The Bacterial Diversity Metadatabase* (Bac*Dive*; bacdive.dsmz.de) has been developed as the world-largest database of standardized prokaryotic phenotypic data [8,9]. For the majority of not-yet-cultured prokaryotes, recent advances in sequencing technology and capacity reveal novel ribosomal sequence types at a rate which is 100 times faster than the description of cultured species [10]. Meanwhile, even the metagenomic datasets of previously unknown prokaryotes

accumulate at a faster rate than species descriptions [11]. Still, large fractions of discovered genes (>75%) cannot be assigned a function [12].

A multidimensional analysis of molecular, phenotypic, and ecological data has the potential to identify previously unknown properties of the available cultured isolates and not-yet-cultured prokaryotes. Recent studies provide new insights into the functions of unknown genes and the specific adaptations of previously uncharacterized microorganisms: Enzyme kinetic data provided by the enzyme database BRENDA [13] can be standardized and employed to predict temperature optima of growth for the respective microorganism [14,15]. A combination of phenotypic information from Bac*Dive* with the phylogenetic affiliation of bacteria provided by the ribosomal RNA database SILVA [16] allowed predictions of the physiological potential of the intestinal microbiome of infants [17]. In order to generate novel scientific knowledge and innovative applications from the existing information, our project will use an integrative and interdisciplinary approach for the mobilization, standardization, and synthesis of microbiological data.

Work programme



WP 1 Mobilization and curation of phenotypic trait data (ZB MED, DSMZ)

Task 1.1 Extraction of phenotypic descriptions from the literature

The aim of this WP is to generate a comprehensive set of phenotypic trait data for all described prokaryotic species. Primary literature represents a highly diverse source for phenotypic data of prokaryotic strains, as this information is scattered over approximately 164 different scientific journals [18]. In addition, relevant information is hidden in the vast amount of unstructured scientific articles, available for instance in large bibliographic collections such as PubMed, as well as Ph.D. theses and reports to funding agencies (both centrally collected by TIB). The lack of standardization and of unique identifiers renders mobilization of these data challenging. Just the typical bacterial species descriptions alone, each in a single publication, contain over 150 different data types that require substantial time for manual annotation [8,9].

Data extraction will therefore be streamlined by establishing automated text mining, employing an artificial intelligence system with machine learning approaches with random forest trees and trained deep neural networks. This will increase the efficiency of data retrieval (i.e. >20 times faster) and thereby enable data extraction on a much larger scale and across multiple journals. ZB MED will provide the required infrastructure and expertise for large scale mining of the biomedical literature corpus. The solution will be built upon the ZB MED Knowledge Environment (ZB MED KE). Data will be hosted by Bac*Dive* that has

already gathered and published bacterial phenotypic metadata for 80,584 strains covering approx. 90% of the described species. So far, Bac*Dive* offers over 6000 datasets enriched with data manually extracted from species descriptions, and hence provides the necessary data infrastructure to conduct the present research project.

Task 1.2 Optimization of data retrieval through feedback loops involving curators

The accessibility of literature data to automated text mining is highly dependent on the specific data type. While some data types like the shape or the size of a bacterial species can already be extracted with high precision, other data types like those describing metabolic features are more challenging, since for instance multiple names for the same chemical compound exist. To attain high quality and correctness of the extracted data, a curation feedback loop will be integrated in the text mining pipeline. Data that were extracted from literature will be displayed together with a confidence score to a microbiologist curator who then has the option to accept, reject, or to correct the data. For this purpose a web interface to present the result and collect the curator's feedback will be developed. The information from curation will then be used to retrain the artificial intelligence iteratively (similar to Prodi.gy), resulting in an improved text mining workflow. It is expected, that the majority of the data fields can eventually be extracted without or with only little human interference reducing the workload of curators.

WP 2 Transformation into a machine readable and FAIR data repository (TIB)

Task 2.1 Semantification of prokaryotic data

The extended Bac*Dive* content will be standardized and transformed into a machine-readable and -processable format following the FAIR (findable, accessible, interoperable, reusable) and Linked Data (LD) principles. To achieve semantic interoperability, the research data in Bac*Dive* will be represented using semantic formalisms including the Resource Description Framework (RDF), ontologies, and R2RML mappings. Such a semantic representation is already widely supported (e.g. by OpenPhacts, Europeana, or search engines like Google) and enables a sustainable use of information and universal analyses. Mappings between existing data and RDF triples ("semantification") will be utilised by RMLMapper [19] and RDFizer [20] tools to achieve the transformation. Existing ontologies (e.g.[21]) will be reused wherever possible to ensure that Bac*Dive* data can be linked with other, already semantically enriched data and thereby integrated into the existing landscape of other semantic services. One example of such a linked data service is the *Chemical Entities of Biological Interest* dictionary [22]. Bac*Dive* data will be converted to RDF and will be offered as downloadable datasets, alongside import documentation for common RDF databases (triple store). In addition, R2RML mappings will be produced to support on-the-fly transformation. To enable data queries that take advantage of the logical relation of data in RDF triples, a SPARQL endpoint will be offered. The existing web-based advanced search of Bac*Dive* will be amended with finer-grained options to query all data fields individually. This will allow scientists to perform federated SPARQL queries that include Bac*Dive* data fields and will allow to detect correlations between microbiological information from heterogeneous sources.

Bac*Dive* currently also offers an application programming interface (API) as web service which will be extended in order to enable scientists to use the latest data in reproducible workflows (e.g. literal programming documents like Jupyter notebooks, or fully automated data processing pipelines).

Our combined approach will ultimately provide the scientific community with a facility for easy lookup, and systematic and focused download of species-associated data in two different, complementary ways. As an example, if a bacterium or a bacterial 16S rRNA sequence has been detected in a particular habitat with specific characteristics, this information can be used to query the SPARQL endpoint in the FAIR Bac*Dive* repository which will return additional potential habitats that feature the same specific characteristics. Using the tools mentioned in Task 2.3 below, scientists will also be able to load and compare all habitat characteristics with the physiological information about the bacterium.

Task 2.2 Expansion of search options through a knowledge graph

After transformation into a fully machine-readable and FAIR-compliant research data repository, the search options will be further developed by establishing the Bac*Dive* Knowledge Graph. This will be conducted in four steps: (1) Semantic integration of data, metadata, and schema. This includes the establishment of an agile, iterative, and community-driven method for ontology development by and with all stakeholders. In preparation, we already initialised the evaluation of the NCBI Taxon ontology [23]; (2) a community-driven definition of quality criteria and classification schemes for the microbiological sector; (3) describing the structure of typical data and metadata formats and creating mappings to the developed ontologies; and (4) realizing a machine-readable knowledge graph by interlinking all relevant information types in interoperable and reusable manners.

Task 2.3 Improving the graphical and programmatic access

Because the R programming language and code package format is popular among bio-scientists [24], a potential reference implementation of theBac*Dive* web service has already been established as an R package [25] in preparation of this proposal. Recognising that Python popularity is on the rise, we will monitor usage metrics of the web service and in case of growing user demand, will also offer a dedicated Python package or contribute to an existing one (e.g. BioServices). Establishing the monitoring by usage metrics will support the gathering of quantitative feedback in Task 5.1.

WP 3 Extending the database through inference of phenotypic data from bacterial genomes sequences (DSMZ, ZB MED)

Task 3.1 Mobilizing genome-based information on bacterial phenotypic traits

Nearly 200,000 bacterial genome sequences have become available [26], including also those of phenotypically characterized species. The bacterial genome sequences allow predictions of phenotypic traits like the utilization of carbon and nitrogen sources, biosynthetic capabilities, motility, sporulation, or secretion. So far this information has not been systematically mobilized and integrated with phenotypic information for the same bacterial species. Working closely together with Task 2.1, potential phenotypic characteristics will be extracted from available genome data. A comprehensive approach will be used starting with a standardized genome annotation based on NCBI RefSeq as well as the customizable pipeline Prokka [27] that builds on highly curated sequence databases such as SwissProt. Some phenotypic traits like motility and sporulation can be built upon simple keyword searches of the gene product names and/or EC numbers of encoded proteins. For the inference of metabolic and biosynthetic properties, however, metabolic models will be build using ModelSEED [28] with subsequent curation using GapSeq [29]. The results obtained will be fed into a novel, genome-based knowledge base which will be established within Bac*Dive*.

Task 3.2 Assessment of the genome-based predictions of phenotypic properties

In certain cases, results of the genome-based analyses of phenotypic properties are expected to be more complete than those derived directly from the literature, since genomic analyses potentially cover all phenotypic properties or metabolic pathways present. In order to evaluate the predictive power and plausibility of the genome-based approach, the derived phenotypic predictions will be tested against the actual phenotypic data retrieved by literature-mining (Task 2.1). This feedback loop will serve the iterative improvement of genome- and literature-based methods for retrieving phenotypic information. Genome-based data that reach or exceed a predefined confidence score will then be used to complement or correct phenotypic entries in the Bac*Dive* database. Genome-based predictions will also serve to extend the Bac*Dive* Knowledge Graph (Task 2.2) providing new semantic relationships as genomic RDF entries.

Task 3.3 Innovative prediction methods for genomes with low annotation level

Genomes of bacterial species that remain so-far-uncultured and/or are phylogenetically distant from known bacterial groups can only be incompletely analyzed since just a small fraction of genes (sometimes even <30 percent) of the open reading frames can actually be annotated [12]. So far, the high number of remaining, unassigned genes can only be described on the protein family level or just as hypothetical proteins. As an innovative approach, we will link the specific phenotypic properties and environmental preferences of so-far-uncharacterized bacteria (obtained from ecological data in WP1) with the occurrence patterns of particular non-annotated genes across their genomes. This will generate hypotheses on the specific functions of non-annotated genes. As a model case, we will use this approach for the analysis of a set of 22 complete genomes from the sparsely described phylum Acidobacteria. All genomes feature about 70 percent hypothetical proteins based on current Prokka annotation. For this particular set of bacteria, extensive habitat information is available, consisting of numerous environmental and soil parameters. Working together with WP4, AI-based methods will be developed and applied for these analyses.

WP 4 Predicting suitable cultivation conditions for uncharacterized bacteria (ZB MED, TIB, DSMZ)

Task 4.1 Prediction methods based on AI

The limited success in cultivating representatives of underrepresented bacterial phyla in the laboratory represents a major obstacle of current microbial diversity research [30,12]. In this WP, the extended phenotypic dataset (consisting of text-mined, genome-derived, semantically transformed data) will be exploited to infer optimal growth conditions (e.g., pH, temperature, salinity, carbon sources, growth factors requirements, trace elements) and corresponding culture media composition for selected target species that have so far not been cultured. Target species will be selected based on the amount of available data and their phylogenetic distance to the most closely related cultivated bacterial species. New AI-based methods will be developed and applied. Training and test sets will be generated from the Bac*Dive* database content and employed to design and benchmark different supervised machine learning methods for classification and regression tasks. Methods to fulfil these tasks and to find the optimal solutions include random forest trees and deep neural networks. We will select the most predictive features by manual as well as automated feature engineering approaches and will compare their results.

Task 4.2 Evaluation and validation of predicted cultivation conditions

The media and growth conditions inferred *in silico* by machine learning approaches will be quality-controlled and improved in subsequent wet lab experiments conducted at DSMZ, employing the identified incubation conditions. Those of the target bacteria for which cultivation conditions are predicted with highest confidence will be chosen for these proof-of-principle laboratory experiments. As inoculum for the cultivation assays, natural samples will be chosen that are readily accessible and contain a sufficient fractions of viable target bacteria, as assessed by Illumina high-throughput sequencing of 16S rRNA as well as 16S rRNA genes by the sequencing facility of the Leibniz Institute DSMZ. Results of wet-lab experiments will serve to retrain and improve the machine learning model.

WP 5 Community engagement and dissemination (DSMZ, TIB)

Task 5.1 Analysis of user feedback

Based on the experience gained by DSMZ since the launch of Bac*Dive* in 2012 [31], the established procedures for collecting quantitative and qualitative user feedback about datasets and services will be applied and continuously developed in the project *DiASPora*. Specifically, server requests and search engine referrers provide valuable quantitative feedback, whereas emails, surveys and personal feedback during conferences and symposia allow qualitative analyses. From the feedback data, and considering the specific comments

regarding graphical and programmatic access (received in Task 2.3), we will extract user personas and stories. This will allow us to identify additional requirements for the continued, agile development and operation of the FAIR Bac*Dive* repository in the future. For example, frequent search keywords could inform about the creation of currently lacking help pages, or about necessary improvements regarding the findability and accessibility of particular data. In addition measures to collect quantitative usage metrics of the Bac*Dive* services will be implemented, including clickstream analysis and A/B-testing [32] of the user responses to changes in structure and design of the FAIR Bac*Dive* website and services.

We will employ agile development practices to continuously improve the phenotypic data services of FAIR Bac*Dive*, i.e. we will address user needs by delivering desired functionality in an iterative manner, monitoring their impact, and thus validating their usefulness in short, fixed time intervals. During all monitoring activity, user privacy will be ensured by gathering only the minimally necessary and strongly anonymised data [33].

Task 5.2 Dissemination and follow-up development strategy

We will publish newly developed tools under OSI-compliant Open Source licenses, adhering to the DFG recommendations about research software and to best practices in the bioinformatics community [34]. We will also implement the best practices of software project management [35] by establishing a public issue tracker or forum for (potential) users. All data resources, analysis results, and tools produced in the *DiASPora* project will be made accessible by extending and reprogramming the existing Bac*Dive* webpage.

Two international workshops will be organized as an additional route to disseminate the *DiASPora* results. The first workshop will address user needs for specific datasets and analysis pipelines. A second workshop will be dedicated to the use of the graphical user interface and to discuss possible improvements; this workshop will be planned and coordinated using the de.NBI/ELIXIR network. In addition, presentations of the outcomes of the *DiASPora* network are scheduled at the project end in the format of a roadshow which will also be presented on three relevant conferences that are held in the year 2022, the International Symposium on Microbial Ecology (ISME), as well as the yearly conferences of the Vereinigung für Allgemeine und Angewandte Microbiology (VAAM), and of the Deutsche Gesellschaft für Hygiene und Mikrobiologie (DGHM). The roadshow will also allow us to receive feedback for subsequent developments of the FAIR Bac*Dive* repository.

		Months											
WP	Description	3	6	9	12	15	18	21	24	27	30	33	36
WP1.1	Extraction of phenotypic descriptions from literature				1								
WP1.2	Optimization of data retrieval through feedback loops involving curators												2
WP2.1a	Semantification of prokaryotic data				3								
WP2.1b											4		
WP2.2a	Expansion of search options through a knowledge graph				5								
WP2.2b													6
WP2.3a	Improving the graphical and programmatic access						7						
WP2.3b									8				
WP2.3c													9
WP3.1	Mobilizing genome-based information on bacterial phenotypic traits						10						
WP3.2	Assessment of the genome-based predictions of phenotypic properties										11		
WP3.3	Innovative prediction methods for genomes with low annotation level												12
WP4.1a	Prediction methods based on AI								13				
WP4.1b									14				
WP4.1c													15
WP4.2	Evaluation and validation of predicted cultivation conditions												16
WP5.1	Analysis of user feedback						17						
WP5.2a	Dissemination and follow-up development strategy												18
WP5.2b							19						
WP5.2c													20

Work programme with milestones (numbered 1 through 20). 1: First full run of the automatic literature analysis, extraction of known phenotypic representations, adaptation to the curation work. **2:** Continous extraction and optimization of literature data. **3:** Semantic data integration for Bac*Dive* initiated (use of ontological sources, RDF representation, full RDF schema). **4:** Concept for enhancement of Bac*Dive* content with complex novel information (e.g. regulatory elements, environmental distribution). **5:** Semantic data integration for Bac*Dive* initiated (use of ontological sources, RDF representation, full RDF schema). **6:** Concept for enhancement of Bac*Dive* content with complex novel information (e.g. regulatory elements, environmental distribution). **7:** Routine approach established to accept requests to Bac*Dive* and respond to the requests. **8:** User access to Bac*Dive* optimized (through web service). **9:** Full integration of Bac*Dive* with public data sources with semantic technologies, including data from automatic literature analysis. **10:** Genome-based information on bacterial phenotypic traits mobilized. **11:** Assessment of the genome-based predictions of phenotypic properties. **12:** Innovative prediction methods for genomes with low annotation level established. **13:** AI version for the prediction of functional annotations for genomes with lack of annotations established**. 14:** Prediction of requirements for the most suitable culture media for uncultured species based on genomic information. **15:** Computational prediction of different phenotypic properties like resistance characterization and industrial use established. **16:** Evaluation and validation of predicted cultivation conditions finalized. **17:** Analysis of user feedback. **18:** Dissemination and follow-up development strategy devised. **19:** First workshop with the user community for feedback to the user interface. **20:** Workshop for planning of the future setup of Bac*Dive* (i.e. extensions, further complex content, further types of use, roadshow).

Originality and innovativeness

Besides significantly facilitating the access to biological information and massively improving analysis options for database users, *DiASPora* will generate novel insights into bacterial enzymology, physiology, and functional genomics through the integration of cutting-edge data sciences and informatics.

Currently, functional analyses of microbial communities are either conducted by (1) standard direct metagenomic sequencing and annotation by database comparisons, (2) direct predictions for a limited number of well-known functional genes for almost complete genomes, or (3) indirect inference of functional genes after determining the phylogenetic position of unknown bacteria based on their 16S rRNA gene sequence, using functional information available for phylogenetically related bacterial genomes. Standard metagenomic sequencing often does not provide sufficient coverage of the entire sequence diversity in routine applications and subsequent database comparisons offer only limited information on functional genes, particularly for members of unsufficiently characterized bacterial groups. Secondly, the more advanced direct predictions of microbial phenotypes are based on comparative genomics and use machine learning to predict a limited number of general phenotypic traits (such as Gram-negative, free-living, anaerobic, or photosynthetic) from a genome that is compared to other, well-annotated genomes [36,37]. Thus far, these methods rely on traits that can be easily linked to the presence or absence of well understood proteins, like outer membrane porins that occur in Gram-negative cell walls. In the third, indirect approach, only 16S rRNA sequences are determined and functional genes are predicted employing annotated genome sequences that are available in the databases and that are phylogenetically related to the 16S rRNA sequence type of interest [38,39].

In contrast to the limitation of these existing methods, *DiASPora* will comprehensively and automatically access and analyze a much larger set of detailed enzymatic, physiological, and ecological information which is then systematically linked to individual bacterial species. Furthermore, *DiASPora* attempts to break new ground by choosing an innovative, data science-based approach to detect, mobilize, and logically link habitat parameters and types of biogeochemical reactions with the occurrence of uncharacterized bacteria in these habitats. This will allow to infer the phenotypic properties of previously uncharacterized types of bacteria from available ecological and ecophysiological information. Here, automatically recognizing correlations and patterns among semantically linked datasets provides the opportunity to discover entirely unexpected relationships. Thereby novel hypothesis can be developed which are less prone to confirmation bias than when testing data according to pre-existing hypotheses. Providing phenotypic information in unprecedented quantity and detail, and in an integrated manner, *DiASPora* will thus offer fundamentally novel insights into the functions of individual bacterial species.

Recently, several proofs of principle for our approach have been published from research conducted outside of our project consortium. Enzyme kinetic data from the BRENDA database were integrated with phenotypic data from Bac*Dive* to predict temperature optima of growth of specific microorganism with high confidence [14]. Also, the physiology of the intestinal microbial community of infants could reliable be predicted after integration of Bac*Dive* data with the phylogenetic affiliation taken from the SILVA database [17].

Academic, ecological and economic implications

The amount of data that is currently generated in the life sciences is rapidly increasing but typically these data are very heterogeneous and often not readily accessible. Therefore, the mobilization, standardization and integration of the disperse data have turned into major obstacles of future data analysis and knowledge generation. The logical interconnection of different data silos is still time-consuming, cumbersome, and requires specialized knowledge. So far, this work is mostly conducted locally and manually by data experts. *DiASPora* aims to systematically address these challenges and to facilitate the generation of substantial added value from the existing information. The project will change the disequilibrium in microbial diversity research between readily available sequence information on one hand and difficult-to-access phenotypic data hidden in literature on the other hand. By the semantic opening of microbial research data, this project will improve the access for non-bioinformatic researchers and enable new, so far impossible, large-scale investigations within the phenotypic data. It will enable linking genotypes more efficiently to phenotypes and thereby offer the opportunity to significantly improve genome annotation data in the rapidly increasing number of genome sequencing projects.

In particular, semantic processing, analysis of interrelations through knowledge graphs, and predictions of previously unknown bacterial phenotypes by AI will lead to an improved understanding of key functions of microbial communities. This novel knowledge in turn will open up new avenues, e.g., for the management of nutrient cycling in soils or the turnover of greenhouse gases. Particular functionalities that are of direct relevance to applications in biotechnology or infection medicine are the identification of previously uncharacterized bacteria that degrade plastics or xenobiotic compounds [40], the prediction of antibiotic resistance profiles for pathogenic bacteria, or the mining of microorganisms for their capabilities to synthesize novel types of natural compounds [41].

DiASPora will also significantly facilitate the integration of phenotypic data in the developing external digital knowledge base of life sciences and biomedical research (e.g., European open science cloud, EOSC-portal.eu). Through its highly standardized, curated, and user-centric approach, and by using an established public database as a starting point, the proposed project can also make a valuable contribution towards the sustainable and proactive research data management that is currently being implemented in the framework of the German National Research Data Infrastructure (NFDI; www.dfg.de/en/research_funding/programmes/nfdi/index.html). All three partners are active in the establishment of one or more of the currently forming NFDI consortia.

B) Quality of the network

DiASPora will bring together the expertise of the DSMZ for mobilizing, curating, and interpreting microbiological data [42] with the expertise of TIB in the fields of big research data, vocabularies, and ontologies [43] and the experiences of ZB MED of developing user-centric software infrastructure and services [44].

Prof. Dr. Jörg Overmann is Director of the Leibniz-Institute DSMZ-German Collection of Microorganisms and Cell Cultures, full Professor of Microbiology at the TU Braunschweig, and leads the DSMZ Department of Microbial Ecology and Diversity research (*DSMZ/MED*). His research focuses on molecular microbial diversity, bacterial physiology, and bacterial

interactions. Establishing innovative cultivation approaches, a large number of bacterial species could be isolated and characterized that had previously escaped cultivation. Jörg Overmann (co)authored over 194 peer-reviewed scientific publications as well as 58 book chapters and reviews, which resulted in an H-index of 55 (Google Scholar). He so far acquired more than 11 Mio € in funding. Amongst other position, he is member elect of the DFG review board 204 `Microbiology, Virology and Immunology´, member of the board `International Continental Drilling Program / Integrated Ocean Drilling Program´ of the DFG, appointed member of the Permanent Senate Commission on Fundamental Issues of Biological Diversity of the DFG, appointed member of the advisory board of the Staatliche Naturwissenschaftliche Sammlungen Bayerns, and member of the scientific advisory board of the Catalan Institute for Water Research (ICRA, Institut Català de Recerca de L´Aigua). Jörg Overmann received the doctoral award of the Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM), a DFG postdoctoral award, and the Inaugural Douglas Leigh Lecturer Award of the Waksman Foundation in 2013.

Dr. Lorenz Reimer is leading the database development team at the DSMZ and is responsible for the Bac*Dive* database. He is a member of the German bioinformatics network de.NBI and participates in the Leibniz workgroup `Forschungsdaten´ as well as in the newly built consortia NFDI4Earth, NFDI4Biodiversity and NFDI4Microbiome. His main research interests are research data hidden in culture collections and primary literature and to change the disequilibrium between readily available sequence data and hard to access phenotypic data in microbial research. He has published more than 10 research papers and conference contributions. Under his lead the content of the database Bac*Dive* has tripled and became a highly frequented database with over 12,000 users per month.

The **Leibniz-Institute DSMZ** is one of the largest biological resource centres worldwide and hosts more than 70,000 biological resources, including 30,000 different bacterial and 5,000 fungal strains. DSMZ has long-term expertise in bacterial genomics and is partner of the Genomic Encyclopedia of *Bacteria* and *Archaea* (GE*BA*) [45]. Since 2012 the DSMZ is developing *The Bacterial Diversity Metadatabase* (Bac*Dive*; bacdive.dsmz.de), the globally largest database for standardized bacterial phenotypic data [8,9]. Research data on the origin, taxonomy, morphology and physiology (e.g., enzymatic activity, metabolic profiles, antibiotic resistance) and cultivation conditions are mobilized, standardized, and assembled for all available bacterial strains. Currently, Bac*Dive* comprises 901,000 data points for 80,584 bacterial strains und receives > 500 users per day with the trend increasing. Because of this visibility and increasing use rate, an increasing number of international collections (CCUG, CIP, CABI) offer their data through Bac*Dive*, which results in a further increase in database contents. Bac*Dive* has been linked to external databases of molecular sequences and environmental data through the *German Federation for the Curation of Biological Data* (GFBio, funded by the German Research Foundation) and is part of the database node of the *German Network for Bioinformatics Infrastructure* (de.NBI, funded by the Federal Ministry of Education and Research). Through these networks, DSMZ has established tight links to the enzyme database BRENDA [46], the globally largest resource on enzyme information with 4.3 million manually annotated data on 84,000 enzymes and 100,000 users per month. Similarly tight links exist to the SILVA database [47] which represents the global database for quality-controlled ribosomal RNA sequences and currently comprises 6 million sequence entries. Being able to link the phenotypic data of microorganisms to the corresponding enzymatic and ribosomal sequence information offers substantial synergism, e.g. for the prediction of physiological traits based on enzyme kinetic data [14] or based on phylogenetic affiliation [17].

Prof. Dr. Sören Auer is a professor for Data Science and Digital Libraries at Leibniz University of Hannover and director of the German National Library of Science and Technology (TIB). He has made substantial contributions to semantic web technologies, knowledge engineering, software engineering, usability, as well as databases and information systems. He (co)authored over 150 peer-reviewed scientific publications, which resulted in an H-index of 50. The excellence of his research was acknowledged by numerous awards, including the Ten-Year Award of the Semantic Web Science Association, the ESWC

7-year best paper award or the OpenCourseware Innovation award. Since 2007 he acquired more than 20 Mio € in funding for the establishment and expansion of his research. He led several large-scale collaborative research projects such as the European Union's H2020 flagship projects BigDataEurope and LOD2. He is co-founder of several high-impact research and community projects such as the Wikipedia semantification project DBpedia, the OpenCourseWare authoring platform SlideWiki.org or the industrial data exchange initiative Industrial Data Space. The technology he develops with his team fuels many societal and industrial applications. He is organiser, programme or track co-chair of renowned conferences and workshops, including OKCON 2010, ESWC 2010, ICWE 2011, WWW 2012, European Data Forum. He also serves as an expert for industry, the European Commission, the W3C and board member of the Open Knowledge Foundation.

Dr. Angelina Kraft is a data scientist and head of the Research Data and Scientific Software team at TIB. The team develops quality standards and infrastructure services for the publication of research data and scientific software according to the FAIR principles. She co-facilitates the task force Open Science - Research Data management within the CESAER (Conference of European Schools for advanced engineering Education and Research) initiative and serves as reviewer for various conferences and journals, including the International Open Science Conference and MDPI. She has published more than 25 research papers at national and international conferences and more than 30 associated digital research data sets. She is member of the DataCite community engagement steering group (CESG) and responsible for the further development of services for digital data management at TIB such as the Leibniz Data Manager (datamanager.TIB.eu).

The German National Library of Science and Technology (TIB) represents and operates a national research infrastructure facility for the provision of scientific information covering engineering, information technology, mathematics, physics, chemistry and architecture. As the world's largest specialised information centre in its fields, TIB has outstanding expertise in developing, managing and preserving knowledge, particularly in key areas such as big research data, vocabularies and ontologies, as well as patents and standards. Examples include the TIB portal giving access to more that 100 million documents and research artefacts and the audio-visual portal AV-Portal comprising more than 15.000 scientific videos, Open Access and Digital Preservation offerings. In 2009 TIB founded the DataCite association with meanwhile more than 130 international member organizations and hosts the DataCite headquarter for providing DOI registration services. TIB performs worldwide recognized research aiming to advance information, data and knowledge sharing in the digital age for example with its Open Research Knowledge Graph and to facilitate the digitalization of science, industry and society at large.

Prof. Dr. Dietrich Rebholz-Schuhmann is professor for Biomedical Data Semantics and Analytics at the University of Cologne and the scientific director of ZB MED, Information Center for Life Sciences. He is a medical doctor and a computer scientist, holds a PhD in immunology and is Doctor of Science in computer science (Habilitation). During his career, he has published more than 150 peer reviewed publications with an h-index of 38. He has raised project funding above 10 Mio € and has contributed as PI to raising institutional funding in the range of 100 Mio €. During his career he has established innovative research-driven IT services for the integration of the scientific literature with life science databases during his employments at the LION Bioscience AG, Heidelberg (1998-2003) and at the European Bioinformatics Institute, Cambridge, UK (2003-2012). Before joining ZB MED he was the director of the Insight Center for Data Analytics in Galway (Ireland) heading 10 research teams in the scientific domains of semantic Web technologies, Internet of Things, semantic data analytics, text mining, knowledge graphs and related areas. Over his research career, Prof Rebholz-Schuhmann has been closely involved in the planning and development of the European Elixir network and has contributed to the roll-out during his work at the EMBL-EBI, but also during his leadership at the Insight Centre for Data Analytics. At ZB MED he is closely collaborating with the de.NBI network (coordinated by the Technical Faculty of the University of Bielefeld) for the development and roll-out of IT services of the life

science domain. Prof Rebholz-Schuhmann is the editor in chief of the Journal of Biomedical Semantics and has been part of the organising committee of the ECCB from 2005 to 2012.

Prof. Dr. Konrad Förstner holds a joint professorship for Information Literacy at the TH Köln – University of Applied Sciences and at ZB MED, where he also leads the unit `Information services´ and is responsible for the discovery service LIVIVO. His research activities include high-throughput sequence data analysis and systems biology in microbiology, text mining and knowledge management of biological literature but also further applications of data science methods in the life sciences. He has published more than 60 articles in these fields including several highly cited ones. Besides this, he is representative of the HRK (German Rectors' Conference) and speaker of the working group `Digital Tools: Software and Services´ in the initiative `Digital Information´ of the Alliance of Science Organizations in Germany and founding member of the working group `Open Science´ of the German Open Knowledge Foundation chapter. Furthermore, he leads the establishment of the NFDI4Microbiome consortium, that aims to create a national infrastructure research data infrastructure for Germany, and is involved in the RSE4NFDI consortium.

ZB MED - Information Centre for Life Sciences was founded in 1973 with offices in Cologne and Bonn. ZB MED holds a collection of 1.6 million books and journals. The mission of ZB MED is the provision of scientific literature and data for life sciences covering Medicine, Health, Nutrition, Environment, and Agriculture. ZB MED is also a driving force in related domains, such as Open Access resources, research data management, digital archiving, and the development of innovative data sources according to linked open data and FAIR principles, open access software solutions and data/information literacy training. Drawing on its digital resources, its positioning in the life science research community, and on the inhouse know-how in information sciences, ZB MED forms a central hub and offers key opportunities for researchers to access, explore, transform and deliver information and to drive progress based on research data in the life science domain. ZB MED is committed to the promotion of Open Access (OA) and forms a hub for the integration of life sciences data for the public domain.

The three partner institutions have collaborated previously in the framework of GFBio and de.NBI consortia which will form a solid basis for linking and synthesizing literature data and database contents as outlined in the current proposal. The partners have also worked jointly together during the conceptualization of the German NFDI. DSMZ and TIB have recently collaborated in a packaging R code to utilise the JSON-based Bac*Dive* web service in preparation of the current proposal [25].

C) Strategic efficacy of the proposed research project

DiASPora will be instrumental for the two participating Leibniz-Institutes, DSMZ and TIB, as well as for the ZB MED, to actively contribute towards the digital change in biodiversity and biomedical research within, but also outside, of the Leibniz Association. By establishing concepts and procedures to mobilize, link, and synthesize biodiversity data, the new project will support the current efforts of the *Leibniz Research Alliance Biodiversity* to translate results from state-of-the-art biodiversity research into sustainable solutions for the future use of biodiversity. Data synthesis also constitutes a key aspect of the work program of other Leibniz Research Alliances, in particular *Infections ´21*, *Sustainable Food Production and Healthy Nutrition*, and *Bioactive Compounds and Biotechnology* (the latter alliance hosts databases for natural compounds); therefore novel approaches to gain information on microbial functions and, in particular, novel natural compounds of microorganisms, will be of particular interest. A FAIR Bac*Dive* data repository will therefore be of strategic value for a considerable number of institutes in the Leibniz Association. Furthermore, the approaches developed for the annotation and interpretation of functional genes will be of direct and practical use for the *Leibniz Omics Network* (LiON; [48]). The LiON network coordinates the

enhancement of omics and bioinformatics capacities across the Leibniz Association.

The concepts and approaches developed for the *DiASPora* project are based on extensive discussions within the Leibniz project group *Digital Change* in which the three applying institutes participated. The project group identified the need for interlinking scientific data repositories across disciplines and the conversion of database content into machine-readable formats as key challenges of the digital transformation and a central goal for the digitalization strategy of the Leibniz Association as a whole.

Extending beyond the Leibniz Association, the national cooperative projects *Global Biodiversity Information Facility* (GBIF,[49]), GFBio [50] and de.NBI [51], have promoted the systematic mobilization and processing of biodiversity data and established the basis for the current project. The synthesis of the contents of the different databases and data sources linked in these previous projects will constitute an important next step in the development of the German database landscape.

Finally, *DiASPora* will support young academics by offering highly attractive, novel opportunities to work on a timely scientific topic at the intersection between life sciences, Omics technologies, and data sciences. All participating institutes have adopted the career guidelines on promoting junior academics and also will apply them for co-workers employed in the present project.

D) Finances

In order to successfully conduct the described work program, dedicated and experienced scientific personnel will be needed. We therefore apply for the position of one postdoctoral coworker for each of the participating institutes (TV-LE13/ Stufe 3) over the entire funding phase. Additional technical support will be needed for (1) wet-lab experiments to test culture conditions predicted by data analysis (DSMZ), (2) for dissemination and technical support for the evaluation of the knowledge graph (TIB) and (3) for technical assistance for the assessment of the literature mining results (ZB MED). Since wet-lab experiments (1) are based on the results of WP4.1, experiments will be started in the second year and the majority of the work will be conducted in the third year of the project. Therefore personnel funding for 50% technician is needed for the last 1.5 years of the project. For the technical support for the knowledge graph (2) and for the assessment of literature mining results (3) funding for 25% technician is needed over the whole project.

In addition, wet-lab experiments require funds for chemicals and consumables for producing culture media and to conduct molecular analysis (incl. sequencing). These include consumables for molecular biology (e.g. pipet tips, reaction vials, microtiter plates: 7.500€), chemicals for microbiology (e.g. culture media, agar: 4.500€), glassware (2.000€) and sequencing cost (8000 €). Corresponding to the technician, the funding for the wet-lab experiments is needed for the last 1.5 years of the project.

To retrieve feedback from users regarding the graphical user interface and the future needs for specific analyses, funding is needed to host two international workshops within the second and third year. Target audiences are scientists from all over Europe from the fields of biology, bioinformatics and computer science on a PhD and post-doc level. The 2-days workshops will be hosted at the DSMZ (or one of the partner institutes) and are limited to 15 participants. Therefore funding is needed for catering (30€ per person/day: 900€), traveling expenses (350€ per person: 5250€), training material (10€ per person: 150€) and rental fee for an external lecture room (300 € per day: 600€).

In order to conduct meetings with all three partners and to present results on international conferences (e.g. VAAM, DGHM, ISME) funding for travel is required over the entire project phase.

Finally, funding is needed in the last year to publish the results. Fees are estimated for two publications including funding for color graphs (1500€ per publication: 3000€).

Leibniz Competition

Bibliography

[1] RfII – Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen 2016. RfII – Rat für Informationsinfrastrukturen: Schritt für Schritt – oder: Was bringt wer mit? Ein Diskussionsimpuls zu Zielstellung und Voraussetzungen für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI), April 2017. www.rfii.de/de/category/dokumente/; RfII – Rat für Informationsinfrastrukturen: Zusammenarbeit als Chance. Zweiter Diskussionsimpuls zur Ausgestaltung einer Nationalen Forschungsdateninfrastruktur (NFDI) für die Wissenschaft in Deutschland, März 2018. www.rfii.de/de/category/dokumente/; RfII – Rat für Informationsinfrastrukturen: In der Breite und forschungsnah:Handlungsfähige Konsortien. Dritter Diskussionsimpuls zur Ausgestaltung einer Nationalen Forschungsdateninfrastruktur (NFDI) für die Wissenschaft in Deutschland, Göttingen 2018. www.rfii.de/de/category/dokumente/

[2] Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387-1390

[3] Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312, 1355–1359

[4] Bulgarelli D, Rott M, Schlaeppi K, van Themaat EVL, Ahmadinejad N, Assenza F et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488: 91-95

[5] Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013) Going back to the roots: the microbial ecology of the rhizosphere. Nature Rev Microbiol 11: 789 – 799

[6] Relman DA (2012) Microbiology: Learning about who we are. Nature 486: 194-195

[7] Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The Human Microbiome Project. Nature 449: 804 – 810

[8] **Reimer LC**, Söhngen C, Vetcininova A, **Overmann J** (2017). Mobilization and integration of bacterial phenotypic data - enabling next generation biodiversity analysis through the Bac*Dive* Metadatabase. J Biotechnol 261: 187-193 https://doi.org/10.1016/j.jbiotec.2017.05.004

[9] **Reimer LC**, Vetcininova A, Sardà Carbasse J, Söhngen C, Gleim D, Ebeling C, **Overmann J** (2019) Bac*Dive* in 2019: Bacterial phenotypic data for high-throughput biodiversity analysis. Nucl Acids Res 47(D1): D631-D636

[10] Yarza P, Yilmaz P, Prüße E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R (2014) Uniting the classification of cultured and uncultured Bacteria and Archaea by means of 16S rRNA gene sequences. Nat Rev Microbiol 12, 635–645.

[11] **Overmann J**, Huang S, Nübel U, Hahnke RL, Tindall BJ (2019) Relevance of phenotypic information for the taxonomy of not-yet-cultured microorganisms. Syst Appl Microbiol 42: 22-29

[12] **Overmann J**, Abt B, Sikorski J (2017) The significance and future of cultivation. Annu Rev Microbiol 71, 711–730

[13] Jeske, L., Placzek, S., Schomburg, I., Chang, A., & Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Research*, *47*(D1), D542–D549. https://doi.org/10.1093/nar/gky1048

[14] Engqvist MKM (2018) Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiol 18: 177
[15] Li G, Rabe KS, Nielsen J, Engqvist MKM (2019) Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. bioRxiv doi: http://dx.doi.org/10.1101/522342
[16] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies J. and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. https://doi.org/10.1093/nar/gks1219
[17] Guittar J, Shade A, Litchman E (2019) Trait-based community assembly and succession of the infant gut microbiome. Nature Comm 10: 1-11
[18] EZB (2019) http://ezb.uni-regensburg.de/ezeit/searchres.phtml?bibid=DSMZ&colors=7&lang=de&jq_type1=QS&jq_term1=microbiology
[19] RML-Mapper (2019) http://rml.io/ & https://github.com/RMLio/RML-Mapper
[20] RDFizer (2019) https://github.com/SDM-TIB/rdfizer-1
[21] Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, Siegele DA, Hu JC, Giglio MG (2014) An ontology for microbial phenotypes. BMC Microbiol 14: 294.
[22] Chebi (2019) https://www.ebi.ac.uk/chebi/
[23] NCBI (2019) https://www.ncbi.nlm.nih.gov/taxonomy
[24] Marwick B, Boettiger C, Mullen L (2018) Packaging Data Analytical Work Reproducibly Using R (and Friends). The American Statistician. 72: 80–88. doi.org/10.1080/00031305.2017.1375986
[25] **Leinweber K** (2018) TIBHannover/BacDiveR: a programmatic interface for the Bacterial Diversity Metadatabase. Zenodo. doi.org/10.5281/zenodo.1308060
[26] NCBI (2019b) https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/
[27] Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068-2069
[28] Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28: 977-982
[29] GapSeq (2019) https://github.com/jotech/gapseq
[30] **Overmann J** (2013) Principles of enrichment, isolation, cultivation, and preservation of bacteria. *In*: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds.) The Prokaryotes, 4th edition, Prokaryotic Biology and Symbiotic Associations. Springer, New York. pp 149-207 (DOI 10.1007/978-3-642-30194-0_7)
[31] Söhngen C, Bunk B, Podstawka A, Gleim D, **Overmann J** (2014) BacDive-the Bacterial Diversity Metadatabase. Nucl Acids Res 42 (D1): D592 - D599
[32] Kohavi R, Longbotham R (2017) Online Controlled Experiments and A/B Testing. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://link.springer.com/referenceworkentry/10.1007%2F978-1-4899-7687-1_891
[33] Ripp S, Falke S (2018) Analyzing user behavior with Matomo in the online information system Grammis. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/7706/file/Ripp_Falke_Analyzing_user_behavior_2018.pdf
[34] Jiménez RC, Kuzak M, Alhamdoosh M *et al.* (2017) Four simple recommendations to encourage best practices in research software. F1000Research 6: 876 (https://doi.org/10.12688/f1000research.11407.1)
[35] Open Source (2019) https://opensource.guide/
[36] Feldbauer R, Schulz F, Horn M, Rattei T (2015) Prediction of microbial phenotypes based on comparative genomics. BMC Bioinformatics 16: S1
[37] PhenDB (2019) https://phendb.csb.univie.ac.at
[38] Aßhauer KP, Wemneuer B, Daniel R, Meinicke P (2015) Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31: 2882-2884

[39] Tax4Fun (2019) http://tax4fun.gobics.de
[40] Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, Toyohara K, Miyamoto K, Kimura Y, Oda K (2016) A bacterium that degrades and assimilates poly(ethylene terephthalate). Science 351: 1196-1199
[41] Tracanna V, de Jong A, Medema MH, Kuipers OP (2017) Mining prokaryotes for antimicrobial compounds: from diversity to function. FEMS Micrbiol Rev 41: 417-429
[42] **Overmann J** (2018) Konzeption, Relevanz und Zukunftsperspektiven moderner mikrobiologischer Ressourcenzentren am Beispiel des Leibniz-Instituts DSMZ-*Deutsche Sammlung von Mikroorganismen und Zellkulturen*. In: Karafyllis, N.C. (ed.): Theorien der Lebendsammlung. Pflanzen, Mikroben und Tiere als Biofakte in Genbanken. (Lebenswissenschaften im Dialog, Vol. 25), Freiburg: Alber. pp. 229-249
[43] Saurbier F, Drees B, Garatzogianni A, **Kraft A**, Sohmen L (2018) Die TIB Labs: Eine Plattform für experimentelle digitale Dienstleistungen, Prototypen und Beta- Versionen der Technischen Informationsbibliothek. https://www.b-i-t-online.de/heft/2018-03/fachbeitrag-saurbier.pdf
[44] Arning U, Lindstädt B, Schmitz J (2016) PUBLISSO: an open access publication portal for life sciences. GMS Med Bibbl Inf 16: Doc 15. https://www.egms.de/static/en/journals/mbi/2016-16/mbi000370.shtml
[45] Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Avanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D´haeseleer P, Zemla A, Singer E, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP, Eisen JA (2009) A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. Nature 462: 1056-1060
[46] BRENDA (2019) www.brenda-enzymes.org
[47] SILVA (2019) https://www.arb-silva.de
[48] LiON (2019) (https://www.leibniz-gemeinschaft.de/en/infrastructures/leibniz-roadmap/lion/)
[49] GBIF (2019) www.gbif.de
[50] GFBio (2019) www.gfbio.org
[51] de.NBI (2019) www.denbi.de