Archives of Microbiology (2019) 201:143–145
https://doi.org/10.1007/s00203-018-1609-9

EDITORIAL

Paradigm shift in species description: the need to move towards a tabular format

Erko Stackebrandt[1] · **David Smith**[2]

Published online: 11 December 2018

As published by the European Commission (https://ec.europe.eu/digital-single-market/eu/big.data) on their policy vision of Big Data, four steps are outlined for the shift of analog data to digital data storage that should leverage their full potential. These range from the investigation into ideas (e.g. research in Horizon 2020) over investigations in infrastructures to the development of building blocks, e.g. multifunctional single portals to open data or guidelines for datasets and finally to trusted building measures (e.g. data protection, ownership issues).

Although datasets from microorganisms currently play a subordinate role as compared to those generated from geographical, weather of health data, they greatly assist, foster and drive innovation in human welfare, agriculture and the pharmaceutical and food industries (Tocchetti et al. 2018). Molecular data have been stored and made accessible shortly after the recognition of their relevance to science at the beginning of the 1980s (Hanson 2000) for studies and predictions, especially in health science (Zentner 2017), food and beverage industries and clinical pharmacy (Ma et al. 2015) and in microbiology such as, to name a few, prediction of metabolic pathways (Langille et al. 2013), analysis of ecosystems (Huson and Mitra 2012) or microbial 'dark matter' (Hedlund et al. 2014).

In contrast, phenotypic data describing and characterizing the function of microorganisms have been gathered for more than 140 years but their access remains fragmented and scattered in the scientific literature. Users searching for properties, especially those to be used in health, industrial application or for any comparative analyses, either have to look into the strain catalogues of public collections of microorganisms or they have to individually screen the literature. Both approaches are tedious as the documentation laid down in catalogues differs from collection to collection and a priori there is no information whether or not a given reference includes the required information. There is an obvious need to bring the comprehensibility of these data on an equal level to the one already reached by genomic data.

The main problem for doing so is first the lack of a unified format in which phenotypic data are recorded and second the inability to mobilize these data into a centralized place, connecting them with genomic data to allow global searches. This change will here be demonstrated using bacterial species descriptions with their associated phenotypic data of type strains as an example. The range of data encompasses culture properties, generated between the 1870s and today. Before 1960, before the advent of Numerical Taxonomy (Sokal and Sneath 1963), the number of phenotypic tests focused on cultural and a few physiological properties. The introduction of commercial test kits and chemotaxonomic analyses for strain characterization, introduced in the 1970s, increased the number of generated traits. Since 1980, with the publication of the 'Approved List' (Skerman et al. 1980) and the advent of molecular data about 16,000 types with about 250 individual phenotypic traits per type strain have been described (http://www.bacterio.net/).

It is one of the stabilizing features of bacterial taxonomy that rarely only characters of proven strain-specific reliability are deleted from the spectrum of properties while, on the other hand, the demands for identifying new features are constantly increasing to discriminate between strains and species. Many of these traits, which are the results of enzymic activities against commercial substrate panels (API, BIOLOG) can today be verified by in silico analysis of genome sequences such as utilization of carbohydrates while others are not accessible to genomic analysis such as cultural properties or chemotaxonomic moieties. The issues of single-strain descriptions (Felis and Dellaglio 2007) or the more recent initiative for the naming of species with

✉ Erko Stackebrandt
erko@dsmz.de

David Smith
d.smith@cabi.org

[1] Kirchbergstr. 9, 38170 Ampleben, Germany

[2] CABI, Egham, UK

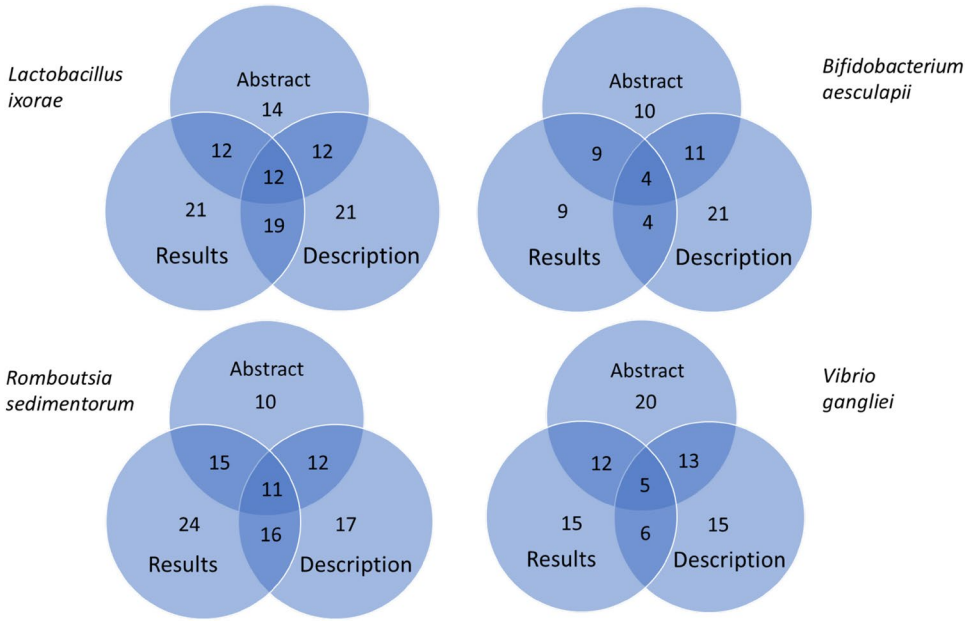

Fig. 1 Number of main results repeated in different sections of a manuscript covering the description of four bacterial species. Individual metabolic traits which are only recorded in the species description part and detailed chemotaxonomic constituents (e.g. of polar lipids, peptidoglycan, polyamines or fatty acids) are not considered. *L. ixorae* (Techo et al. 2016), *B. aesculapii* (Modesto et al. 2014), *R. sedimentorum* (Wang et al. 2015), and *Vibrio gangliei* (Meng et al. 2018)

emphasis on genomic data (Chun et al. 2018) will not be covered in this essay.

Recently, attempts have been started to facilitate access to a broad range of collection catalogues (http://abc.wfcc.info/) or even to generate cumulative databases to facilitate the search for properties, such as BacDive (https://bacdive.dsmz.de/) for bacteria and Mycobank (http://www.mycobank.org/) for fungi but in both of these databases data need to be transferred manually from the original publication. In any case, be it the manual transfer of data into individual catalogues or into centralized databases the transfer itself is prone to human errors and highly time consuming, and additionally, expensive, particularly if data transfer is done according to a quality management regime. A more advanced solution has recently been offered by the generation of the Digital Protologue Database (Rosselló-Móra et al. 2017) which is generated by the authors of a species description and stored centrally at the same time as the description itself. Though in principle machine readable, mechanisms for the transfer of such data into a centralized database have not yet been established.

However, there is a second issue for consideration. Recently, publishers of scientific journals began to screen manuscripts for plagiarism by comparing the text of new submissions to those already published. We will only refer to results of some Springer Nature journals which may be representative of the entire publication scene. Among all manuscripts screened, those covering species descriptions contain the highest degree of overlap with previously published sources which may be up to 60%, often, in species-rich genera, with 30% from a single source. This is due to a very standardized format and the conformity to the requirements of 'Minimal Standards' (e.g., Bernardet et al. 2002) and the majority of species descriptions differ only in the listing of taxon-specific properties. More disturbing for the reader is the repetition of results, provided in the Abstract, Results and Description sections as well as often in Tables of comparative properties. In the majority of such publications, the ecological significance of an organism's occurrence in the environment or its pathogenic/beneficial importance for humans, plants or animals is missing. In effect, most species descriptions are spiritless, using the same template for mainly changing phenotypic +/− reactions, but often missing vital information such as confirmation of properties derived from genomic in silico-derived information.

To demonstrate the degree of overlap between individual manuscript sections four different species descriptions were randomly selected as representatives, ranging from those based upon a traditional polyphasic approach (Cowell 1970) to a more genome sequence-centered description scheme (De Vos et al. 2017). The degree of overlap is visualized in Venn diagrams (Fig. 1). The overlap between the three sections is mainly seen in the repetition of characterizing phenotypic properties while those related to genome/gene sequence similarities, phylogenetic tree topologies and delineation of species threshold values are only indicated in Abstract and Results sections as this information contributes to the justification of the taxon description but is not part thereof. This is obvious in the description of the *Bifidobacterium* and the *Vibrio* species which include a higher proportion of sequence data. It must also be noted that in several instances the original description sees a lack of a (though small) number of phenotypic traits which are either

indicated in the Abstract or Result section but missing in the Description part.

We conclude that the traditional taxon description that served microbiology for so many decades is outdated as it contains too much redundant information and its information cannot be directly transferred to public databases. It is our recommendation to initiate a paradigm shift by replacing the current species description format in which the traditional protocol with its repeated information is replaced by a highly standardized tabular format. This would include a set of mandatory data fields to be decided upon by an international committee of taxonomists (based upon work already done), the option to include fields for referenced material, references and methods as well as the option to include new or modified methods. The new format would allow adaptation to any concept change of species descriptions.

Today we have reached a point which is almost comparable to the situation in molecular biology 30 years ago, when the publication of mere short sequences as full papers (such as those reporting ribosomal RNA species) was replaced by deposition of sequences into sequence data bases. As the validation of a bacterial name requires publication in the scientific literature an exclusive electronic publication is excluded (Lapage et al. 1990). We do not want to give the impression that the transition into a tabular species description will be straight forward as the ground has to be prepared first which would include

- the consent of the majority of bacterial taxonomists and the International Committee on Systematics of Prokaryotes
- the publisher's agreement to have the new format accepted and the data transferred,
- a decision upon the format of the data fields as well as the most appropriate electronic exchange format of data, such as the Microbiological Common Language (Verslyppe et al. 2010).

We would like to initiate the discussion with interested readers on this topic to develop such a structure of high importance for linking phenotypic and molecular data of microorganisms.

References

Bernardet JF, Nakagawa Y, Holmes B (2002) Subcommittee on the taxonomy of *Flavobacterium* and *Cytophaga*-like bacteria of the International Committee on Systematics of Prokaryotes. Proposed minimal standards for describing new taxa of the family *Flavobacteriaceae* and emended description of the family. Int J Syst Evol Microbiol 52:1049–1070

Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol 68:461–466

Colwell RR (1970) Polyphasic taxonomy of the genus *Vibrio*: numerical taxonomy of *Vibrio cholerae, Vibrio parahaemolyticus* and related *Vibrio* species. J Bacteriol 104:410–433

De Vos P, Thompson F, Thompson C, Swings J (2017) A flavor of prokaryotic taxonomy: systatics revisited. In: Microbial resources. From functional existence in nature to applications, 2. London Academic Press, E book, Chap, pp 29–44

Felis GE, Dellaglio F (2007) On species descriptions based on a single strain: proposal to introduce the status species proponenda (sp. pr.). Int J Syst Evol Microbiol 57:2185–2187

Hanson T (2000) Walter Goad, GenBank founder, dies. Newsbulletin: obituary. Los Alamos National Laboratory

Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T (2014) Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". Extremophiles 18:865–875

Huson DH, Mitra S (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol 856:415–429

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences, predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotech 31:814–821

Lapage SP, Sneath PHA, Lessel EF Jr, Skerman VBD, Seeliger HPR, Clark WA (1992) International code of nomenclature of bacteria, (1990) revision. *Bacteriological Code*. American Society for Microbiology, Washington, DC

Ma C, Smith HW, Chu C, Juarez DT (2015) Big data in pharmacy practice: current use, challenges, and the future. Integr Pharm Res Pract 4:91–99

Meng YC, Liu HC, Zhou YG, Cai M, Kang Y (2018) *Vibrio gangliei* sp. nov., a novel member of *Vibrionaceae* isolated from sawdust in a pigpen. Int J Syst Evol Microbiol 68:1969–1974

Modesto M, Michelini S, Stefanini I, Ferrara A, Tacconi S, Biavati B, Mattarelli P (2014) *Bifidobacterium aesculapii* sp. nov., from the faeces of the baby common marmoset (*Callithrix jacchus*). Int J Syst Evol Microbiol 64:2819–2827

Rosselló-Móra R, Trujillo ME, Sutcliffe IC (2017) Introducing a digital protologue: a timely move towards a database-driven Systatics of archaea and bacteria. Antonie Van Leeuwenhoek 110:455–456

Skerman VBD, McGowan V, Sneath PHA (1980) Approved lists of bacterial names (amended). ASM Press, Washington

Sokal R, Sneath PHA (1963) Principles of numerical taxonomy. W.H. Freeman, San Francisco

Techo S, Miyashita M, Shibata C, Tanaka N, Wisetkhan P, Visessanguan W, Tanasupawat S (2016) *Lactobacillus ixorae* sp. nov., isolated from a flower (West-Indian jasmine). Int J Syst Evol Microbiol 66:5500–5505

Tocchetti A, Donadio S, Sosio M (2018) Large inserts for big data: artificial chromosomes in the genomic era. FEMS Microbiol Lett 265(1):9

Verslyppe B, Kottmann R, De Smet W, De Baets B, De Vos P, Dawyndt P (2010) Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Res Microbiol 161:439–445

Wang Y, Song J, Zhai Y, Zhang C, Gerritsen J et al (2015) *Romboutsia sedimentorum* sp. nov., isolated from an alkaline-saline lake sediment and emended description of the genus *Romboutsia*. Int J Syst Evol Microbiol 65:1193–1198

Zentner A (2017) Big data and microbiology. *J Microbial Genetics* J113