Reflections on the introduction of the Digital Protologue Database – a partial success?

Ramon Rosselló-Móra[1*] and Iain C. Sutcliffe[2]

[1] Marine Microbiology Group, Department of Ecology and Marine Resources, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Balearic Islands, Spain
* For correspondence: ramon@imedea.uib-csic.es

[2] Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, UK
** For correspondence: iain.sutcliffe@northumbria.ac.uk

Modern science revolves around databases, be they the massive (e.g. NCBI) or the bespoke (e.g. EzBioCloud). There are enormous databases covering the sequence world and the protein world but what of the organisms from which they are derived? With this is mind, we have argued (Sutcliffe et al. 2012; Rossello-Mora, 2012; Rossello-Mora and Amann, 2015; Sutcliffe, 2015; Rossello-Mora and Whitman, 2019) that microbial systematics needs to become a database driven science. After all, if it has taken more than a century to characterise <20,000 of the estimated >10m prokaryotic species (<0.2%), then a flexible repository will be needed if we are to complete a timely systematic census of the microbial world. An ideal database would integrate information on the characteristics of a taxon with nomenclatural information and links out to other databases, particularly for sequence data, and back to the original data source (primary publication). Entries would range from the minimal information needed to delineate a novel taxon through to maximal descriptions of well characterised taxa.

In an attempt to achieve this (or at least set the ball rolling), in 2017 we established the Digital Protologue Database (DPD) (Rossello-Mora et al., 2017a and 2017b). This database was intended to be reminiscent of the public repositories compiling genes, genomes and proteins, wherein the entries were ordered by unique identifiers (Taxonumbers). Data was entered in fields to capture all the relevant information that is normally given as a text in the protologues of taxonomic papers (i.e. etymology, diagnostic properties and designated type material). However the Digital Protologues (DPs) were intended to capture much other very relevant metadata on the described taxa such as the geographical origin, kind of sample, gene and genome entries in public repositories, as well as other medical and ecological properties. Authors describing taxa in *Antonie van Leeuwenhoek* and *Systematic & Applied Microbiology* were encouraged to populate the nascent database with information on the taxa being described. Indeed, *Systematic & Applied Microbiology* required their authors to compulsorily fill in the forms, and also to substitute the written protologues in their paper, which are often redundant in the contribution, for protologue tables directly extracted from the DPD. This was also intended to help guarantee the accuracy of the entries. There was some early successes – shortly thereafter, the DPD was also recommended to authors of papers in *Archives of Microbiology*, *Current Microbiology* and more recently at *New Microbes New Infections* (as part of "New Species Announcement 2.1") (Stackebrandt and Smith 2017a.; Stackebrandt and Smith 2017b.; Drancourt and Fournier, 2018). Relatively quickly the DPD has grown to include almost 1000 entries and almost 750 registered users in just 2 years. A feature of the design of the database is that only those entries curated as representing effectively or validly published taxa are in the public domain – although adding publication details has proven onerous for the database editors, there has been a steady growth in the release of validly and effectively published entries.

Despite this progress, there have been challenges associated with operating the DPD. Its initial configuration provided a fairly basic level of functionality. Quality control of entry information (curation), including fixing incomplete and/or erroneous entries, along with updating the DPD with citation details following effective or valid publication of taxa has proven onerous. Regrettably, we have been unable to secure the funding or support of a larger organisation that would allow us to improve the DPD via curators and database architects. Funding is also lacking for information scientists to employ use machine-learning approaches to backfill the database with information on the ca. 15,000 historically described taxa. Perhaps if we had adopted a different model for database operation (e.g. wiki style editing by the community) some of these challenges could have been addressed. However, also problematic has been that we have been unable to secure the support of the editorial board of the International Journal of Systematic & Evolutionary Microbiology (IJSEM). We fully understand the workload concerns of the IJSEM editors and acknowledge their important contributions to the field. Nevertheless, as ca. 75% of the taxa described each year are published in IJSEM, the lack of input from this major journal remains a significant limitation of the database.

These challenges will inevitably be amplified if we succeed in our goals of shifting microbial systematics toward becoming a database driven field, especially if we see an anticipated (and indeed hoped for) step-change in much volume of taxonomic activity, such that we can classify and name perhaps 10-fold more taxa per annum. Therefore we have been forced, reluctantly to conclude that the DPD cannot be maintained in its current form. Consequently, the editors of *Antonie van Leeuwenhoek* will no longer insist that authors describing taxa also create entries in the DPD, although we will recommend this and hope that many will continue to do so. On the other hand, the editors of *Systematic & Applied Microbiology* will continue to ask their authors to fill a streamlined version of the DPD with the purpose of the entries becoming the metadata support for the Microbial Genome Atlas (MiGA; Rodriguez-R et al., 2018) of any genome, metagenome assembled genome (MAG) or single amplified genome (SAG), published in this journal. In any case, we will also maintain an archive of the information stored such that, at some future point, this can be used to populate any new database established for this type of activity.

Despite this set back we remain convinced that the diversity microbial of the world must eventually be captured in a functional and interactive database. Ultimately we hope that there will be change in the publication ‘habits’ of the microbial taxonomy community such that the current formulaic species description papers are no longer viewed as the “currency unit” for building careers. Instead, we would encourage a shift towards minimal database entries and/or species ‘announcements’ (diagnosis) that map the microbial world and are then complemented by retrospective comprehensive analyses (description) of representative, significant or problematic taxa and characteristics of interest (Table 1).

Its limitations and flaws notwithstanding, we hope our DPD experiment has been instructive and useful project that may, ultimately, inspire others to attempt to succeed where we have not. Indeed, to end on a positive note, we have been impressed and sustained by the enthusiasm and support of user community, who we thank greatly for their efforts to date. We also greatly thank Pierre-Edouard Fournier and Erko Stackebrandt and for their ongoing support on behalf of, respectively, *New Microbes New Infections* and *Archives of Microbiology*/*Current Microbiology*.

References

Drancourt, M., Fournier, PE (2018) New species announcement 2.1. New Microbes New Infect 25: 48.

Rodriguez-R LM, Gunturu S, Harvey WT, Rossello-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. (2018) The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nuc Acid Res 1 doi: 10.1093/nar/gky467.

Rosselló-Móra, R. (2012). Towards a taxonomy of *Bacteria* and *Archaea* based on interactive and cumulative data repositories. Environ Microbiol 14:318-334.

Rosselló-Móra, R., Amann, R. (2015) Past and future species definitions for *Bacteria* and *Archaea*. Syst Appl Microbiol 38:209-216.

Rosselló-Móra, R., Trujillo, M., Sutcliffe, I.C. (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria. Syst Appl Microbiol 40: 212-122.

Rosselló-Móra, R., Trujillo, M., Sutcliffe, I.C. (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria. Antonie van Leewenhoek. 110: 455-456.

Rossello-Mora R, Whitman WB (2019) Dialogue on the nomenclature and classification of prokaryotes. Syst Appl Microbiol In Press

Stackebrandt, E., Smith D. (2017a) Expanding the 'Digital Protologue' database (DPD) to 'Archives of Microbiology': an offer to scientists and science. Arch Microbiol 199: 519-520.

Stackebrandt, E., Smith D. (2017b) Expanding the 'Digital Protologue' database (DPD) to 'Current Microbiology': an offer to scientists and science. Curr Microbiol 74(9):1003-1004

Sutcliffe I.C. (2015) Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: rip it up and start again. Frontiers in Genetics 6, 218.

Sutcliffe, I.C., Trujillo, M.E. and M. Goodfellow (2012). A call to arms for systematists: revitalising the purpose and practises underpinning the description of novel microbial taxa. Antonie van Leeuwenhoek 101: 13-20.

Table 1 Traditional versus modernised approaches to describing the microbial world

Approach	**Status quo**	**Modernised 'high throughput'**
Method	Polyphasic	Genome-based
Rate	~1000 species per annum	10-fold increase?
Process	Characterise ► Classify ► Name ► Publish	Sequence ► Classify ► Name ► select for in-depth characterisation ► Publish
Primary forum	Journal publication	Database entries ► DOI assignation or similar microattribution
Output	Typically formulaic single strain species descriptions	Publications synthesising knowledge and insights at different taxonomic levels
Bottlenecks	Already a major editorial and peer review burden	Requires development of sophisticated machine-learning database technology
	Staid academic community will become a barrier to early career recruitment?	Promoting an engaging field attractive to future early career scientists