diversity

MDPI

Perspective

Hemi- and Homonyms in the Big Data Era

Jorge Rubén Sánchez-González [1,2]

[1] Section of Quality, Environmental Evaluation and Environment—TRAGSATEC-SEPI, C/Julian Camarillo, 6B, Sector D, 28037 Madrid, Spain; jorge.r.sanchez.gonzalez@gmail.com; Tel.: +34-973-702-902

[2] Department of Animal Science-Wildlife Section, University of Lleida, Av. Alcalde Rovira Roure 191, 25198 Lleida, Spain

Received: 23 October 2020; Accepted: 11 December 2020; Published: 13 December 2020

Abstract: The issue of hemi- and homonyms is an unsolved topic in the Big Data era, where informatics and technicians, rather than biologists or taxonomists, analyze huge datasets. Nowadays, taxonomic nomenclature is ruled by four independent international codes, and according to them, the existence of hemihomonyms and homonyms is accepted under some conditions as an exception to the general rule. This situation entails confusion, disagreements, and a plethora of problems whose consequences could worsen in the near future within the framework of the big data era. Moreover, the increasing use of big databases and analyses, data science, bioinformatics, biological monitoring, and bioassessment has shown such exceptions to be inconvenient, since these exceptions to homonyms are considered as duplicates by databases and statistical software, which are handled by non-taxonomist experts. International Codes of Nomenclature must change within the new context of big data analysis. This work aims to propose the elimination of any exception to the presence of homonyms and to evaluate whether the Independence Principle makes sense within this new context. Increasing coordination between several independent nomenclatural systems is essential and, perhaps, we must conduct our efforts towards a universal species list, finishing with the historical schism between Codes.

Keywords: taxonomic nomenclature; big data era; hemi- and homonyms; data science; BioCode; Darwin Core; Globally Unique Identifiers; Life Science Identifier

1. Introduction

The classification of organisms is ruled by the International Code of Nomenclature for algae, fungi, and plants (ICN) [1], the International Code of Zoological Nomenclature (ICZN) [2], the International Code of Nomenclature of Prokaryotes (ICNP) [3] and the International Code of Virus Classification and Nomenclature (ICVCN) [4]. Moreover, these different international codes of nomenclature operate from different perspectives, applying different criteria and ruling, almost, in total independence, and, in consequence, uncoordinatedly, despite the International Committee on Bionomenclature's (ICB) efforts. As a result, the regulatory framework is restricted to the group of organisms within the same nomenclatural jurisdiction.

Within this context, where different international codes of nomenclature operate almost uncoordinatedly, many names have been assigned unintentionally to two or more different taxa. These cases are considered homonyms, i.e., identical accepted names applied to unrelated taxa but within the same nomenclatural jurisdiction, or hemihomonyms [5], existing across different nomenclatural jurisdictions, for which no replacement names are proposed for the second usage [6]. In most cases, names of taxa can be distinguished thanks to the author and year of description when they are included in the citation and, of course, thanks to the context where they are placed.

The existence of homonyms and hemihomonyms is a consequence of this uncoordinated regulatory framework and a classical question that, despite the uncountable efforts made, remains unsolved [7–12].

The consequences of these hemi- and homonyms could worsen in the close future within the framework of the big data era. These problems might be caused by (i) homonyms within the same nomenclatural jurisdiction that have not been recognized or replaced; (ii) homonyms within the same nomenclatural jurisdiction that have been recognized and replaced but for which there is no easy resolution mechanism (i.e., homonyms from online taxonomic or nomenclatural databases); (iii) hemihomonyms, existing across nomenclatural jurisdictions and homonyms; and (iv) hemihomonyms that have not yet been created. International codes of nomenclature establish, with a set of rules and recommendations, how formal names are given to species and other taxa. Most of the hemi- and homonym cases are included in types i–iii and all the existing nomenclatural codes offer rules about them with different levels of accurateness and, in consequence, varying degrees of success. For example, the ICN is restrictive on the use of homonyms (Article 53) by considering illegitimate later homonyms [1]. Similarly, the ICNP [3] regulates homonyms (Rule 23a) by rejecting, by the Judicial Commission, earlier synonyms and homonyms. In the same sense, the ICVCN [4] prevents the existence of homonyms by applying rule 3.14, according to which "*New names shall not duplicate approved names*" and any new names might be approved by the International Committee on Taxonomy of Viruses (ICTV). Finally, the introduction to the ICZN lists eight principles; among them, the fifth principle states that "*To avoid ambiguity, the use of the same name for different taxa must not occur and is prohibited. This is the Principle of Homonymy*".

This regulatory framework is restricted to the group of organisms within the same nomenclatural jurisdiction (types i and ii). However, codes do not face properly with hemihomonyms that have not yet been created (type iv). In fact, according to the ICN [1], its Article 54 clearly indicates that "*consideration of homonymy does not extend to the names of taxa not treated as algae, fungi, or plants*". Similarly, the Independence Principle (Article 1.4 from ICZN, 2000) states that the zoological nomenclature is independent of other systems of nomenclature in that the name of an animal taxon is not to be rejected merely because of being identical to the name of a taxon that is not an animal. Fortunately, the contrary is also true: within the same kingdom, one name of a taxon must be applied only to a single taxon.

Traditionally, new errors of classification, synonyms, and/or misspelling could be produced by non-taxonomical experts by using, e.g., biological indicators. However, the existence of homonyms and/or hemihomonyms is a type of error that comes from the taxonomical nomenclature context, which could remain hidden and their consequences undetected, especially for non-taxonomical experts. Afterwards, these nominas are used in multivariate analyses and/or big data analyses, where they could be considered as the same taxonomic group or duplicates, when it is not true or, on top of this, their taxonomical distance is huge. Within this (uncoordinated) regulatory framework, the increasing use of big data analyses, bioinformatics, environmental DNA methods, the use of big databases to catalogue the existing biodiversity and alien invasive species, or within the context of biological monitoring and/or bioassessment of ecosystems (e.g., the Water Framework Directive (WFD) [13] in Europe or the National Water-Quality Assessment (NAWQA) in the United States), are bringing forth some limitations and errors in nomenclature. These errors are not directly related to classification errors or misspelling but to the presence of homonyms and/or hemihomonyms because different taxa from different nomenclatural jurisdictions are used simultaneously in a single enormous database.

Big data analysis, bioinformatics, environmental DNA methods, and integrated systems of monitoring and/or evaluation of ecosystems are currently providing interesting information, methods, new perspectives, and a wide range of potential. As a counterpart, the use of such technologies entails some restrictions, limitations, rules, and principles. In fact, these techniques require common criteria, homogeneous and standard codes, and unique and (as much as possible) stable nominas [14]. One of these restrictions, as it was pointed out previously, is based on the use of the same term to refer to different records: it is not possible to use the same name to refer to two different observations or variable names within a database. Most statistical software and databases require an identifier that must be unique for each observation—in this case, each taxon—and the same identifier must be always referred to the same observation (again, a taxon in this case). Therefore, it is not possible to use the same name to refer to two different taxa. When two elements are identical, they are considered as a

unique element or the same taxon and, in consequence, they can be managed as duplicates, and one of them will be removed, or they could be wrongly considered as two observations of a unique element, making an undetectable error.

To sum up, according to nomenclatural codes, the existence of hemi- and homonyms is regulated and contemplated under some specific conditions, ranging from absolutely rejecting any hemi- and/or homonym to considering exceptions to the general rule. Nevertheless, the increasing use of big databases and analyses, within the context of bioinformatics, biological monitoring, and bioassessment, has shown such exceptions to be inconvenient, since these exceptions to homonyms are considered as duplicates by databases and statistical software. Moreover, the use of accepted species by users without relevant taxonomic expertise would require a unified list of species instead of forcing them to select between different sources, lists of species, and international nomenclature codes which are difficult to understand [7].

2. The Problem in Practice

Different initiatives from different countries and international organizations are attempting to catalogue the biodiversity by constructing large datasets (Catalogue of Life, World Register of Marine Species, FishBase, AlgaeBase, Global Biodiversity Information Facility, and many others)—in some cases, open access and through the use of data standards, such as Darwin Core (DwC). Eidos [15] is just one of these initiatives, whose objective is inventorying wild species of Spanish wildlife with a database according to the Plinian Core Standard [16], based on DwC, developed by the Global Biodiversity Information Facility (GBIF) and managed by the Spanish Government. Eidos aims to group and homogenize under an international standard protocol the species information from different official databases, legislation, national inventories, etc., in coordination with other European databases like European University Information Systems Organization (EUNIS) and European Alien Species Information Network (EASIN).

As a result of analyzing and processing the information from different sources, databases, and catalogues, the presence of some homonyms and hemihomonyms was detected. After evaluating and researching these homonyms, most of them were old hemi- and homonyms, but many others were not real homonyms (i.e., identical nominas applied to taxa governed by the same code) but hemihomonyms. However, the mere presence of hemi- and homonyms could cause some technical problems by considering different taxa as the same element or duplicates. For example, *Tuber* P.Micheli ex F. H. Wigg. (1780) belongs to the Kingdom Fungi, *Tuber* Schröder, Medioli & Scott, 1989 belongs to the Kingdom Chromista, and *Volutella* Perry, 1810 belongs to the Kingdom Animalia; *Volutella* Tode belongs to the Kingdom Fungi, and *Volutella* (Chardez, 1972) belongs to the Kingdom Chromista; *Pustularia* Swainson, 1840 belongs to the Kingdom Animalia, and *Pustularia* Bonord. belongs to the Kingdom Fungi; *Spirospora* Scherff. (Doubtful) belongs to the Kingdom Fungi, and *Spirospora* R. R. Kudoa (Doubtful) belongs to the Kingdom Protozoa; *Crassula* Marwick, 1948 belongs to the Kingdom Animalia, and *Crassula* L. belongs to the Kingdom Plantae; and, finally, *Ludwigia* Bayle, 1878 belongs to the Kingdom Animalia, and *Ludwigia* DC. belongs to the Kingdom Plantae [17–19]. These generic names could thus be considered as hemihomonyms and, therefore, duplicates by databases and statistical software.

These are just some examples, and evaluating the magnitude of this problem is not easy. Even though, and to preliminarily evaluate the magnitude of this problem, a brief search for homonyms and hemihomonyms, entitled Hemi- and Homonyms Dataset, has been performed, using as a starting point the "Hemihomonyms database" [9], Hemihomonyms [20], the List of Valid Homonyms [21], personal communications, and individual contributions (see Data Availability Statement). This preliminary exercise of evaluating the magnitude of the hemi- and homonyms' problem provides interesting information. The last edition of Catalogue of Life contains 1,837,565 living and 63,418 extinct, 1,900,983 in total [18]. On the other hand, in the Hemi- and Homonyms Dataset (Data Availability Statement), 2887 nomina have been listed as homonyms and/or hemihomonyms (Figure 1b), i.e., 0.15%

of the total number of species. In total, there are 1432 homonyms and hemihomonyms. In 99.02% of cases, a taxon belonging to Kingdom Animalia is implied, and in 94.34% of cases, a taxon belonging to Algae, Fungi, and Plantae is implied (see Data Availability Statement). A preliminary basic analysis shows that 1373 cases (95.88%) are hemihomonyms, 38 cases (2.65%) are homonyms, 13 cases (0.91%) were registered as homonyms and hemihomonyms at the same time, and, finally, 8 cases (0.56%) of triple hemihomonyms were detected (Figure 1a). According to this review, it is evident that the main problem is not the homonyms but the hemihomonyms, so better coordination between Codes is critical.

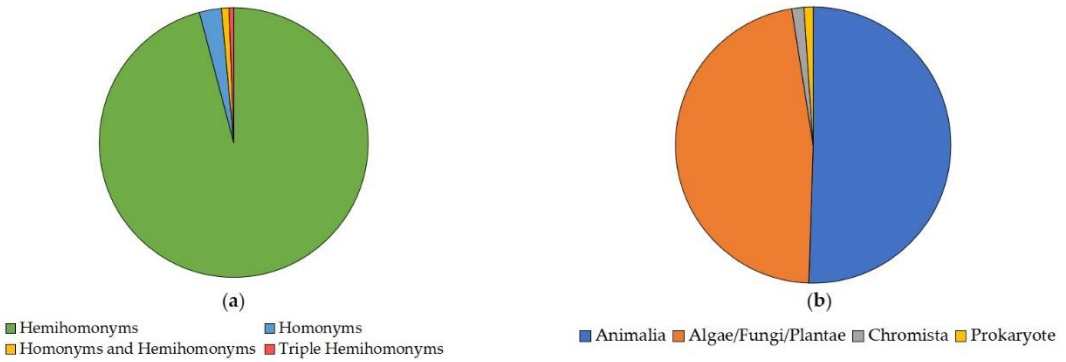

Figure 1. (**a**) Proportions of hemihomonyms (95.88%) in green, homonyms (2.65%) in blue, hemi- and homonyms (0.91%) in yellow, and triple hemihomonyms (0.56%) in red; (**b**) There are 2887 nomina that are hemi- and homonyms, 1458 (50.50%) belong to Kingdom of Animalia, 1357 (47%) to the Kingdoms of Algae, Fungi, or Plantae, 41 (1.42%) to the Kingdom of Chromista, and 31 (1.07%) to the Kingdom of Prokaryote.

3. Recommendations for the Future

This work aims to modestly contribute to the ongoing discussion about the fusion of the Codes or the creation of a single authoritative list of the world's species [7,8] within the new context of the Big Data era.

First of all, the strict application of Principle 5 from ICZN [2] could be considered essential, where it is unfruitfully expressed that "*to avoid ambiguity, the use of the same name for different taxa must not occur and is prohibited*", which is named, in fact, the Principle of Homonymy, and the recommendation 1A, where it is explicitly stated that "*Names already in use for taxa that are not animals. Authors intending to establish new genus-group names are urged to consult the* Index NOMINUM GENERICORUM (PLANTARUM) *and the Approved List of Bacterial Names to determine whether identical names have been established under the International Codes of Nomenclature relevant to those lists and, if so, to refrain from publishing identical zoological names*". On the other hand, Article 52.7, "*Homonymy with names of taxa which are not animals. The name of an animal taxon identical with the name of a taxon which has never been treated as animal is not a homonym for the purposes of zoological nomenclature [Arts. 1.4, 2.2]*", justifies not dealing with this issue, and, according to our preliminary results (Data Availability Statement), this laxity has produced nefarious results. Moreover, according to the obtained results, I consider it adequate to evaluate whether the Independence Principle (Article 1.4 5 from ICZN [2]) and Article 52.7 make sense within the current context of big data analysis and databases or whether they must be modified to state that a name of any taxon can be applied only to a unique taxon independently of the kingdom.

Secondly, other potential approximations to solve these problems could consider the use of common criteria or identifiers. Within this philosophy, Shipunov [9] proposed using a postfix for nomina covered by bacteriological ("b"), botanical ("p"), and zoological ("z"). However, this proposal could increase misspelling errors and using taxa as an identifier must be avoided because it is conceptually wrong and a source of errors. The use of Globally Unique Identifiers (GUID) such as the Life Science Identifier (LSID) is an initiative that has gone further. A GUID is a globally unique and persistent identifier that could provide more optimal results when identifying taxa in computer

systems; LSID is a universally unique identifier linked to a species or taxon and this LSID corresponds to the *scientificNameID* field in DwC. In fact, a potential solution could be based on following an international standard such as DwC, using common terms like *scientificNameID* or *scientificName* (*genus* + *specificEpithet* + *scientificNameAuthorship*) to avoid identical names. However, the presence of spaces, parenthesis, and other symbols could cause problems in most of the statistical and database software.

Another option could be the fusion of Codes or the creation of the first universally recognized list of species. With these actions, the historical schism [22] between nomenclature systems will be resolved. These ideas are not new but are currently a highly topical subject [7], even in news media [10]. This initiative will be associated with enormous disruption in the use of historical names; it will disrupt nomenclatural stability. However, the analysis of the advantages might be performed within the long-term context. Unifying the Codes will be a herculean task but essential in the future, with clear and positive global benefits in the long term. Nowadays, BioCode [11,12] is leading this initiative based on the convergence to a unique Code providing a novel basis for a unified nomenclature of organisms and solving some of the problems caused by the existence of different codes for classifying a single object of study: organisms. However, this is not a unique initiative: Catalogue of Life is a working project to create a global database of species that, unfortunately, has not been universally accepted. Currently, BioCode is a draft in active discussion, but, regarding homonymy, the major change proposed would be that it would operate across the kingdoms according to its Article 18.1 [11]. Since the existence of homonyms and/or hemihomonyms could be interpreted by software as the same taxon, then, in many cases, when this problem is detected, an alphanumeric code is commonly used as identifier (id), but this solution might be specific for each database or nomenclatural code, because a common alphanumeric code, such as LSID, will require a global coordination or an identifier of the taxonomic group.

To sum up, it seems that using a universally unique identifier such as LSID, or similar, is the fastest and probably the most optimal solution. However, this is a temporal solution that does not solve the real problem. A definitive solution requires the fusion of the Codes. Meanwhile, an increment and improvement in coordination mechanisms are considered essential, perhaps, with the elaboration of common articles or rules in order not to repeat a taxon name and to avoid hemi- and homonyms.

In case this proposal was accepted, the Principle of Priority or Priority rule to deal with the existing hemi- and homonyms might be followed.

4. Conclusions

To sum up, International Codes of Nomenclature must change within the new context of big data analysis and data science. This work proposes the elimination of any exception to the presence of hemi- and homonyms and to evaluate whether the Independence Principle makes sense within this new context. The existence of hemihomonyms and homonyms is a classical question that remains unsolved and the solution lies in an improvement of the Codes, their application, and coordination and cooperation between all the International Codes of Nomenclature, creating a unified global list of accepted species and, maybe, in the future, their fusion.

Funding: This research received no external funding.

Data Availability Statement: Hemi- and Homonyms Dataset deposited in Figshare at https://doi.org/10.6084/m9.figshare.11467374.

Acknowledgments: This manuscript was inspired by talks and discussions with Elena Herrero and it would not be possible without her contributions, reviews, and comments: To your daughter Irene. I thank Diego Barberán Molina for his essential helpful assistance, his comments and crucial aid, Lucy McKenna, Pablo Manzano, and Sergio Velasco for their comments and English revision, Rosario Rebole for providing some homonyms, and Alfredo G. Nicieza and Anabel Perdices for their reviews, comments, and recommendations. I want also to thank Elena Bermejo Bermejo and Blanca Ruíz Franco for their support and motivation. This manuscript was born within the context of the Project "Manage and development of applications for the Spanish Biodiversity Data Bank" 17MNES008 and supported by the Government of Spain and the Ministry for the Ecological Transition. The publication of this manuscript was possible with the financial support of my parents: Julio Sánchez González and Milagros González Laiz.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McNeill, J.; Barrie, F.R.; Buck, W.R.; Demoulin, V.; Greuter, W.; Hawksworth, D.L.; Herendeen, P.S.; Knapp, S.; Marhold, K.; Prado, J.; et al. (Eds.) *International Code of Nomenclature for Algae, Fungi, and Plants (Melbourne Code). Regnum Vegetabile 154*, 18th ed.; Koeltz Scientific Books: Melbourne, Australia, 2012; ISBN 978-3-87429-425-6.
2. ICZN. *International Code of Zoological Nomenclature*, 4th ed.; Ride, W.D.L., Cogger, H.G., Dupuis, C., Kraus, O., Minelli, A., Thompson, F.C., Tubbs, P.K., Eds.; The International Trust for Zoological Nomenclature 1999; The Natural History Museum: London, UK, 2000.
3. Parker, C.T.; Tindall, B.J.; Garrity, G.M. International code of nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **2019**, *69*, S1.
4. ICTV. The International Code of Virus Classification and Nomenclature. In *Virus Taxonomy—Ninth Report of the International Committee on Taxonomy of Viruses*; King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., Eds.; Elsevier: London, UK, 2011; pp. 1273–1277, ISBN 978-0-12-384684-6.
5. Starobogatov, Y.I. Problems in the Nomenclature of Higher Taxonomic Categories. *Bull. Zool. Nomencl.* **1991**, *48*, 6–18. [CrossRef]
6. Costello, M.J.; Bouchet, P.; Boxshall, G.; Fauchald, K.; Gordon, D.; Hoeksema, B.W.; Poore, G.C.B.; Soest, R.W.M.; van Stöhr, S.; Walter, T.C.; et al. Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. *PLoS ONE* **2013**, *8*, e51629. [CrossRef] [PubMed]
7. Garnett, S.T.; Christidis, L.; Conix, S.; Costello, M.J.; Zachos, F.E.; Bánki, O.S.; Bao, Y.; Barik, S.K.; Buckeridge, J.S.; Hobern, D.; et al. Principles for creating a single authoritative list of the world's species. *PLoS Biol.* **2020**, *18*, e3000736. [CrossRef] [PubMed]
8. Garnett, S.T.; Christidis, L. Taxonomy anarchy hampers conservation. *Nature* **2017**, *546*, 25–27. [CrossRef] [PubMed]
9. Shipunov, A. The problem of hemihomonyms and the on-line hemihomonyms database (HHDB). *Bionomina* **2011**, *4*, 65–72. [CrossRef]
10. Greenfield, P. Scientists Put forward Plan to Create Universal Species List. The Guardian. 2020. Available online: https://bit.ly/3qnALJk (accessed on 10 July 2020).
11. Greuter, W.; Hawksworth, D.L.; Mcneill, J.; Mayo, M.A.; Minelli, A.; Sneath, P.H.A.; Tindall, B.J.; Trehane, R.P.; Tubbs, P.K. Draft BioCode: Prospective International Rules for the Scientific Names of Organisms. *Bull. Zool. Nomencl.* **1996**, *53*, 148–166. [CrossRef]
12. Greuter, W.; Garrity, G.; Hawkworth, D.; Jahn, R.; Kirk, P.M.; Knapp, S.; McNeill, J.; Michel, E.; Patterson, D.J.; Pyle, R.; et al. Draft BioCode (2011) Principles and rules regulating the naming of organisms. New draft, revised in November 2010. *Bionomina* **2011**, *3*, 26–44. [CrossRef]
13. European Commission. WFD Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for the Community Action in the Field of Water Policy. *Off. J. Eur. Communities L. 327* **2000**, *17*, 1–73.
14. Winston, J.E. Twenty-First Century Biological Nomenclature—The Enduring Power of Names. *Integr. Comp. Biol.* **2018**, *58*, 1122–1131. [CrossRef] [PubMed]
15. MITECO; TRAGSATEC EIDOS: Base de Datos que Integra la Información Sobre Especies de Flora y Fauna Silvestres Presentes en España. Available online: https://www.miteco.gob.es/es/biodiversidad/servicios/banco-datos-naturaleza/Eidos_acceso.aspx (accessed on 30 August 2018).
16. Plinian Core Task Group Plinian Core. *Biodiversity Information Standards (TDWG)*. 2018. Available online: https://github.com/tdwg/PlinianCore (accessed on 24 September 2020).
17. GBIF.org GBIF Home Page. Available online: https://doi.org/10.15468/dl.example-donotcite (accessed on 14 July 2019).
18. Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist. Available online: www.catalogueoflife.org/annual-checklist/2019 (accessed on 2 August 2020).
19. WoRMS Editorial Board World Register of Marine Species. Available online: http://www.marinespecies.org (accessed on 14 January 2020).

20. Wikispecies Hemihomonyms. Webpage Database. Wikidata ID: Q36033662. Available online: https://species.wikimedia.org/w/index.php?title=Category:Hemihomonyms&pageuntil=Sclerochiton#mw-pages (accessed on 25 December 2019).
21. Edkins, K. List of Valid Homonyms. Wikimedia List Article. Available online: https://species.wikimedia.org/wiki/List_of_valid_homonyms (accessed on 25 December 2019).
22. Nicolson, D.H. A History of Botanical Nomenclature. *Ann. Missouri Bot. Gard.* **2006**, *78*, 33–56. [CrossRef]