European Journal of Taxonomy 852: 43–56
https://doi.org/10.5852/ejt.2022.852.2007
ISSN 2118-9773
www.europeanjournaloftaxonomy.eu
2022 · Aguiar A.P. & Broad G.R.

Opinion article

urn:lsid:zoobank.org:pub:F96D8E4A-85E4-44E5-8277-105B621F7830

A new technique and software to optimize compression and data retrieval in the Material Examined section of taxonomic publications

Alexandre P. AGUIAR [1,*] & Gavin R. BROAD [2]

[1] Federal University of Espírito Santo, Dept of Biological Sciences, Zoology, Av. Fernando Ferrari 514, Goiabeiras, Vitória, ES, 29075−010, Brazil.
[2] Natural History Museum, Cromwell Road, London SW7 5BD, UK.

[*] Corresponding author: alexandre.aguiar@ufes.br
[2] Email: g.broad@nhm.ac.uk

[1] urn:lsid:zoobank.org:author:6CA248DA-56F5-45C3-9642-5B7B1F2A617A
[2] urn:lsid:zoobank.org:author:D06689DE-526F-4CFA-8BEB-9FB38850754A

Abstract. Elusive flaws are identified in techniques widely adopted to organize the Material Examined sections in taxonomic publications, mostly regarding the usage of the term *ibidem* and the nesting of information such as country and states. Logical errors are identified that prevent objective retrieval of the original information and can hinder or block its interpretation, even in case-by-case analyses. It is demonstrated that the free usage of *ibidem* in the sense of "same as previous except as follows" compromises the interpretation of data, characterizing bad practice. Solutions are proposed for the precise usage of both the term *ibidem* and the nesting technique. A new technique for organizing, compressing, and presenting information, called grid-setting, is described and evaluated. Its most notable practical effect is that the Material Examined section becomes literally a coded data sheet, which can be accurately converted back to spreadsheet format. In addition, the grid-setting technique was able to generate texts up to 30% shorter than those edited with the best-known traditional techniques. The new ideas and fixes are incorporated into a new software, flexible enough to process varied and unlimited data into largely user-defined texts, which remain nevertheless universal in their format and logical interpretation.

Keywords. Biodiversity, distribution, metadata, meta-analysis, time series.

Aguiar A.P. & Broad G.R. 2022. A new technique and software to optimize compression and data retrieval in the Material Examined section of taxonomic publications. *European Journal of Taxonomy* 852: 43–56. https://doi.org/10.5852/ejt.2022.852.2007

Introduction

Taxonomic publications, especially in the case of revisions, usually contain lists of information from the labels of the studied specimens, as well as other data such as sex, repository institution, notes on the integrity of the specimen and more. All this information is traditionally organized in a separate section for each taxon, usually subtitled "Material Examined" (hereafter ME). These sections can contain a minimum of text, as in the case of taxa known only from singletons, to cases where the ME section can

be much longer than the rest of the description (e.g., about 17 000 vs 7500 characters, or 2.25 × longer, for the treatment of *Digonocryptus crassipes* (Brullé, 1846) in Aguiar & Ramos 2011). The ME list can also represent the totality of the published information for a taxon, in cases where only new records need to be reported. In Telnov (2020) for example, ME lists are the only information presented for 87 species treated on pages 184–203.

Despite its importance, preparing the ME sections "can require a large amount of time in formatting species records" therefore contributing to worsening the "taxonomic impediment" problem, as already discussed by Brown (2013). It might thus seem surprising that even though ME lists are almost omnipresent in taxonomic publications, there is still no proposal for an easier and more universal standardization or rationalization of its format.

Chester *et al.* (2019) did propose a useful format, but it is quite elaborate and can be time-consuming to learn and reproduce correctly. Data recovery was also not primarily focused on individual users but rather aimed at external and centralized automatic parsing: "Using GoldenGATE and TaxPub, the taxonomic treatments and specimen data [...] are converted into Darwin Core (DwC) archives [...] [D]ata relating to the treatments and specimens [...] is accessible via Treatment Bank and GBIF [...]".

More generally, however, especially in the field of entomology, where records are often numerous, it is customary for the information in ME to be presented with some structure. A common pattern is to group data by countries and states, listed alphabetically (however, see Zanella *et al.* 2000), followed by other political subdivisions, and then by coordinates, elevation, date, collector name, collecting method, and institution code. This is the structure adopted for example by Brown (2013, 2021) and is also much like that in Chester *et al.* (2019). However, many variations are still being published, making it increasingly difficult to retrieve and compare the original information.

Aguiar (1998) proposed and used a standard to reliably list verbatim label data, but the idea was to encode historical records. This may be useful for type material but is neither practical nor easily retrievable with numerous specimens. Historically, each author has adopted the format they deem most appropriate, or the format required by the selected scientific journal, although only a few have defined or sufficiently precise requirements. This heterogeneity creates problems of its own. For example, the *European Journal of Taxonomy* adopts the precise but sophisticated requirements of Chester *et al.* (2019), which can increase the time needed to do the formatting (author) or to ensure that it is entirely correct (author and editors). *ZooKeys*, on the other hand, grants ample freedom to authors, supporting publications that cite, for example, only the database number of the studied specimens, storing in an external source all information regarding the collecting events.

The reasons for this general situation are probably varied, but it is possibly related to the large and growing variety of specimen data, which might create the impression that any attempt to propose a standard would result in something overly complex or incomplete.

Some problems generated by such wide freedom and variety of ME formats are notorious. The first is the time needed to create and perfect the format to be adopted, since each author and work will have their own idiosyncrasies. The second problem is a consequence of the first: the difficulty of recovering the original data, precisely because each format requires a different interpretation. Some formats can also result in or influence the organization of data in an incomplete, confusing, or even unresolvable way. Added to this there is the fact that ME lists, especially if long, are rarely the target of the scrutiny of reviewers and editors, and even when they are, there will hardly be an exhaustive or efficient check of the integrity of the ME list's usually compressed and confusing data, creating a circular problem. The central question is, then, why present long lists of ME, which take much time to prepare and a reasonable portion of a publication, without an assurance that the original information can be accurately retrieved?

The tendency of the ME section to be confusing is not merely because of the diverse information it contains, but also because of the structure and the way abbreviations are used. A seemingly obvious solution would be not to abbreviate anything, listing all specimen data fully. The extra space this would take in a publication can be easily and dramatically reduced using small font sizes, e.g., as done in Aguiar (1998), which is a viable option even for publication of extensive molecular data, as demonstrated by Aguiar (2013). Presenting ME data only as "supplementary files", on the other hand, incurs serious problems intrinsic to this practice (see Anderson *et al.* 2006; Seeber 2008; Kenyon & Sprage 2014; Pop & Salzberg 2015).

However, even if all specimen data are presented without abbreviations, the need for its strict organization remains. Without this, objective or automatic data recovery would remain difficult or impossible. Even so, whether for stylistic reasons, habit, personal preferences, space constraints, or other reasons, the extensive usage of abbreviations in ME persists as an editorial or scientific requirement, even in the absence of a precise and user-friendly method or standard to guide this practice.

The objective of this work is to identify flaws and propose adjustments and improvements for the construction of a logically correct and optimally compressed ME section, to propose a universal format for it, and to introduce a new software that incorporates the discussed techniques.

Material and methods

The total number of characters for all texts analyzed in this work was counted considering also all the blank spaces between words. The different types of information in a list of specimen data, separated by comma or semicolon, are referred to as variables (e.g., "Brazil, ES, Vitória, 27–Jun–2021" has four variables) and each unit of information itself is referred as the value of the variable (e.g., value "Brazil" or value "27–Jun–2021" in the previous example). The bullet point "•" is used throughout the work to separate data sequences from different specimens or collecting events, as proposed by Chester et al. (2019). The Supp. files 1 and 2 were used as default for tests, examples, and comparisons presented herein and carried out along the development of the new software.

Results

Traditional techniques

The traditional way of organizing and compressing information in the ME section involves three main techniques. The first and most obvious is simply to cite the number and sexes of the specimens for which the information is identical, e.g., "17 ♂♂, 26 ♀♀; Collection Naturhistorisches Museum Basel // Papua New Guinea Madang Prov. L. Cizek lgt. // Salemben village 145°24′ E 4°42′ S 16.XII.2000, 750 m; NHMB" (from Telnov 2020), which implies that all 43 specimens share the same exact label information and were deposited in the same museum.

The second is *nesting*, where information is arranged in a hierarchical structure. This is done using one or more highly inclusive variables (e.g., country and state), followed by a colon or dash, with each variable organized in its own alphabetical order. It is then assumed that all information cited after the colon shares the information before the colon. Thus, the structure "**ARGENTINA**: **Misiones**: Data1 • **BRAZIL**: **ES**: Data2 • Data3 • **SP**: Data4" indicates one collecting event in Argentina and three collecting events in Brazil, two of them in ES and one in SP.

The third technique is the use of the Latin term *ibidem* (often abbreviated to *ibid.*) to avoid repetition of one or more pieces of information from one collecting event that are identical in the next. There are two possible approaches. The first is to use *ibidem* strictly to replace each respective repeated piece of information. For example, in "Los Angeles, Malaise trap, 1000 ft., J.Smith • San Diego, *ibidem*, 20 ft., *ibidem*" the second collecting event would therefore correspond to "San Diego, Malaise trap, 20 ft., J.Smith".

Table 1. Excerpt of ME from Supeleto *et al.* (2019). The repository abbreviation is always repeated in this format, and the term *ibidem* is abbreviated as "ibid." (highlighted in red to call attention to the number of occurrences and localization).

BRAZIL – **São Paulo State** • 1 ♂; Luiz Antônio, Jataí Ecological Station; 21°37′26.1″S, 47°48′24.5″W; 1 Oct. 2008; NW Perioto shipping; riparian forest; Malaise; Pt1; IBRP • 1 ♂; ibid.; 13 Jan. 2009; IBRP • 1 ♂; ibid.; Malay; Pt2; IBRP • 1 ♂; ibid.; 12 Nov. 2008; IBRP • 1 ♂; ibid.; 15 Oct. 2008; IBRP • 1 ♂; ibid.; 3 Sep. 2008; IBRP • 2 ♂♂; ibid.; 17 Sep. 2008; IBRP • 2 ♂♂; ibid.; 24 Oct. 2007; IBRP • 2 ♂♂; ibid.; 29 Oct. 2008; IBRP • 4 ♂♂; ibid.; 27 Sep. 2007; IBRP.

The second approach is to use *ibidem* only once, implying that the data right after it is the "same as previous except as follows". In the example given above, the format for the second event would be "*ibidem* except San Diego, 20 ft." or simply "*ibidem*, San Diego, 20 ft." But this can get more complex. In the real case reproduced in Table 1, the first "ibid." refers to all information from the immediately preceding event, except the date, which is provided separately. With that, the information implied by the second "ibid." is different from that implied by the first, and a new change here ("Pt2") implies changes also to the third "ibid.", and so on. Many authors use a phrase such as "same data except…"; there are many ways of expressing the same intention.

Limitations

From the point of view of interpretation leading to complete and correct data recovery, the first technique works perfectly, but for the second and third there are logical problems that prevent or hinder its objective interpretation or the development of an algorithm for automatic interpretation.

With the strict usage of *ibidem* the interpretation of data is accurate and objective, but this format commonly results in numerous repetitions of *ibidem* (e.g., Table 1), which might become confusing, particularly in large ME sections. This approach is used in one of the options of the grid-setting technique, but with modifications (see the '*Grid-setting*' section below, The *strict* approach).

The usage of *ibidem* in the sense of "same as previous except as follows" leads to excellent compression, but the interpretation of the resulting text is subjective. This makes this format particularly difficult to be decoded by an algorithm or even by someone who is not familiar with the information being presented (e.g., obsolete names for locations, foreign alphabets). This is because new information inserted after each *ibidem* cannot be consistently and correctly classified without knowledge about what they are. The problem can be mitigated by adding an explicit identification, such as "State" in "São Paulo State", which allows for correct association between all equivalent information. But this is a verbose solution, not practical to use with all variables that describe specimen data. Without prior information, identifying the equivalence of terms such as "Near riverbank" vs "Canopy", or "Malaise trap" vs "ex. *Ficus* sp.", or the correct association of information from different classes, such as reserves vs collector names, etc., involve complex interpretations. The free usage of *ibidem* in the sense of "same as previous except as follows" is therefore classified here as bad practice and should be avoided.

The problem with the *nesting* technique is similar, because the way it is traditionally used in the construction of the ME produces results whose interpretation also depends on prior knowledge about the nature of each variable. For example, in the list "A: B: C • D: E: F • G • H: I" the precise relationships between all units cannot be fully decoded. This is the same structure as the example given above, for which the complete interpretation of the hierarchy between the variables is ((**ARGENTINA** (**Misiones** (Data1))) (**BRAZIL** (**ES** (Data2, Data3)) (**SP** (Data4)))).

This can, however, be fully deduced only if based on knowledge about the countries, provinces and states involved.

The problem can also be visualized if each unit of information is replaced by a letter (A, B, etc.), using the same letter within each collecting event. Consider for example the lists below:

BRAZIL: 1♀, ES: Santa Maria de Jetibá: Clarindo Krüger Farm, 6 Dec 2002 • Event2 • Event3 • 2♂♂, Conceição do Castelo: Ribeirão do Meio, 17–24 Mar 2007 • Event5 • Event6 • etc.

BRAZIL: 1♀, Espírito Santo: Santa Maria de Jetibá: Clarindo Krüger Farm, 6 Dec 2002 • Event2 • Event3 • 2♂♂, Pará: Serra Norte, 17–24 Mar 2007 • Event5 • Event6 • etc.

The structure of the information presented for the first and fourth collecting events (identified as A and D below) in both lists is as follows:

A: A, A: A: A, A

D, D: D, D

Since the full interpretation of the data from one collecting event is tied to the previous one (hence the *nesting*), the above structure leaves room for two interpretations of the data in the fourth collecting event:

A: A, A: A: A, A

A, D, **D:** A, D, D 1st possible interpretation

A, D, A, **D:** D, D 2nd possible interpretation

If only the structure of the text is considered, it will therefore be impossible to define whether "Conceição do Castelo" or "Pará" correspond to a state or a city. A correction for this problem is proposed in the section *Adequacy of the nesting technique*.

The problems described above support the perception that ME lists, as currently formatted, are confusing and susceptible to errors, both in their preparation and in their subsequent interpretation.

Grid-setting

The new technique presented below, here called grid-setting, is quite simple, but more elaborate schemes were also considered (e.g., the topological, used in the example above), without good results. The main idea of grid-setting lies in formatting the ME to make it equivalent to an encoded data sheet, but simple enough to remain human-readable, facilitating its conversion back to spreadsheet format, either manually or via software. There are two possible approaches; the first is described below together with an explanation of the grid-setting technique and its core algorithm.

The "*strict*" approach

In this case the term *ibidem* is used strictly (see the '*Traditional techniques*' section above) but with a simple correction, the quantification of repetitions, as described in point 5 below.

The algorithm for formatting with grid-setting involves only a few steps: (**1**) organize the available data in a spreadsheet, with specimens in rows and variables in columns (Fig. 1); (**2**) rearrange the columns according to the desired order of information in the final text, e.g., to start with country or institution code (Fig. 2); (**3**) make a "multiple sort" of the lines according to the selected columns, to group together identical values (Fig. 3); (**4**) replace all repetitions along each *column*, from top to bottom, for all columns, with a code (Fig. 4); and (**5**) join all sequences of identical codes on each *line*, preceding each repetition (ib) or absence (?) code with the number of uninterrupted occurrences; for example, "ib,

ib, ib, 20.III.1967, ?, ?, ib" becomes "3ib, 20.III.1967, 2?, ib" (Fig. 5); (**6**) join all lines with a bullet point, creating a single paragraph, as illustrated in Table 2.

The "*implicit*" approach

In this case, the term *ibidem* is used in the sense of "same as previous except as follows". This is a familiar usage, with excellent text-compression power, but requires a correction for the intrinsic problem of this approach, described in the '*Limitations*' section above. This can be achieved with a syntax of the type `i:text`, that is, the number of the variable (index `i`) that contains each value (`text`) listed after *ibidem*. If this is adopted, the example "3♂♂, *ibidem*, 3:São Roque do Canaan, Alto Misterioso, 8:C. Waichert exped., 2–11 Nov 2007", would indicate that all information after *ibidem* is identical to the previous collecting event, except for the variables in columns 3, 4, 8 and 9, for which the information

scientificName	sex	institutionCode	country	state	city	locality	date	elevation	lat	lon	collector	typeStatus
Distictus tibialis	1f	CNCI	ARGENTINA	Punta Lara	?	?	27–Jan–66	?	?	?	H.Townes and M	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Conceição d	Propriedade	17–24 Mar 2007	?	?	?	A.P.Aguiar expe	Other
Distictus tibialis	1f	CNCI	BOLIVIA	Noryungas	Coraico	El Bagante	18–Apr–97	150m	?	?	L.Masner leg.	Other
Distictus tibialis	3m	UFES	BRAZIL	ES	São Roque d	Alto Misteri	2–11 Nov 2007	?	?	?	C.Waichert expe	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Santa Teres	Res. Biol. Au	08–24 Oct 2016	764m	19°54′37.7″S	40°33′12.1″W	A.P.Aguiar expe	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Domingos N	Mata Pico d	03–10 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1f	UFES	BRAZIL	ES	Domingos N	Mata Pico d	Nov–03 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1f	UFES	BRAZIL	ES	Alfredo Cha	Picadão	8–15 Oct 2007	714m	20°27′53″S	40°42′35″W	C.O.Azevedo ex	Other
Distictus tibialis	1f	DZUP	BRAZIL	PR	Curitiba	?	1961	?	?	?	V.Graf leg.	Paratype
Distictus tibialis	1m	UFES	BRAZIL	ES	Domingos N	Mata Pico d	03–10 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1f	USUC	BRAZIL	CE	Serra do Ara	?	19–May–69	850m	?	?	M.Alvarenga leg	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Cariacica	RES. Biol. Du	21–30 Oct 2005	?	?	?	A.P.Aguiar expe	Other
Distictus tibialis	1f	CNCI	ARGENTINA	Missiones	Dos de May	?	Feb–67	?	?	?	?	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Domingos N	Mata Pico d	03–10 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1f	UFES	BRAZIL	ES	Conceição d	Propriedade	17–24 Mar 2007	?	?	?	A.P.Aguiar expe	Other
Distictus tibialis	1f	DZUP	BRAZIL	PR	Curitiba	?	1961	?	?	?	V.Graf leg.	Holotype
Distictus tibialis	1m	UFES	BRAZIL	ES	Domingos N	Mata Pico d	Nov–03 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1m	FSCA	ARGENTINA	Corrientes	Las Marias	ca. Virasoro	10–15 Nov 1969	?	?	?	C.Porter leg.	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Santa Teres	Res. Biol. Au	08–24 Oct 2016	775m	19°55′16.4″S	40°33′13.5″W	A.P.Aguiar expe	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Alfredo Cha	Picadão	8–15 Oct 2007	710m	?	?	C.O.Azevedo ex	Other
Distictus tibialis	1m	UFES	BRAZIL	ES	Domingos N	Mata Pico d	03–10 Dec 2004	?	20°22′17″S	40°39′29″W	M.Tavares expe	Other
Distictus tibialis	1m	CNCI	ARGENTINA	Punta Lara	?	?	31–Jan–66	?	?	?	H.Townes and M	Other

Fig. 1. Spreadsheet data from some of the specimens of *Distictus tibialis* (Brullé, 1846) cited in Supeleto *et al.* (2019). Mandatory columns and column names in the **Gridit** software marked in red.

scientificName	typeStatus	sex	country	state	city	locality	lat	lon	elevation	collector	date	institutionCode
Distictus tibialis	Other	1f	ARGENTINA	Punta Lara	?	?	?	?	?	H.Townes and M	27–Jan–66	CNCI
Distictus tibialis	Other	1m	BRAZIL	ES	Conceição d	Propriedade	?	?	?	A.P.Aguiar expe	17–24 Mar 2007	UFES
Distictus tibialis	Other	1f	BOLIVIA	Noryungas	Coraico	El Bagante	?	?	150m	L.Masner leg.	18–Apr–97	CNCI
Distictus tibialis	Other	3m	BRAZIL	ES	São Roque d	Alto Misteri	?	?	?	C.Waichert expe	2–11 Nov 2007	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Santa Teres	Res. Biol. Au	19°54′37.7″S	40°33′12.1″W	764m	A.P.Aguiar expe	08–24 Oct 2016	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1f	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	Nov–03 Dec 2004	UFES
Distictus tibialis	Other	1f	BRAZIL	ES	Alfredo Cha	Picadão	20°27′53″S	40°42′35″W	714m	C.O.Azevedo ex	8–15 Oct 2007	UFES
Distictus tibialis	Paratype	1f	BRAZIL	PR	Curitiba	?	?	?	?	V.Graf leg.	1961	DZUP
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1f	BRAZIL	CE	Serra do Ara	?	?	?	850m	M.Alvarenga leg	19–May–69	USUC
Distictus tibialis	Other	1m	BRAZIL	ES	Cariacica	RES. Biol. Du	?	?	?	A.P.Aguiar expe	21–30 Oct 2005	UFES
Distictus tibialis	Other	1f	ARGENTINA	Missiones	Dos de May	?	?	?	?	?	Feb–67	CNCI
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1f	BRAZIL	ES	Conceição d	Propriedade	?	?	?	A.P.Aguiar expe	17–24 Mar 2007	UFES
Distictus tibialis	Holotype	1f	BRAZIL	PR	Curitiba	?	?	?	?	V.Graf leg.	1961	DZUP
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	Nov–03 Dec 2004	UFES
Distictus tibialis	Other	1m	ARGENTINA	Corrientes	Las Marias	ca. Virasoro	?	?	?	C.Porter leg.	10–15 Nov 1969	FSCA
Distictus tibialis	Other	1m	BRAZIL	ES	Santa Teres	Res. Biol. Au	19°55′16.4″S	40°33′13.5″W	775m	A.P.Aguiar expe	08–24 Oct 2016	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Alfredo Cha	Picadão	?	?	710m	C.O.Azevedo ex	8–15 Oct 2007	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos N	Mata Pico d	20°22′17″S	40°39′29″W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1m	ARGENTINA	Punta Lara	?	?	?	?	?	H.Townes and M	31–Jan–66	CNCI

Fig. 2. Columns swapped according to the desired order for the information in the final text. The order of the column names (scientificName, typeStatus, sex, country, etc.) corresponds to the variable Display Order in the **Gridit** software.

given applies. Columns 4 and 9 can be inferred because there are two subsequent blocks of information, separated by a comma, after each index mark. Cells with missing information continue to be coded by the quantification of repetitions, thus avoiding cases such as "?, ?, ?" in favor of "3?".

A few but significant improvements can be added. (**1**) Once the exact position of all new information is indexed, there is no more need for citing *ibidem*, which becomes implied by the usage of the syntax `i:text`. (**2**) Most formats for citing dates (20–Mar–67, Apr 1992, etc) are also characteristic enough to reliably identify this type of information; these cases therefore do not need to be explicitly indexed, since its index can be figured from its first occurrence. (**3**) There is no need to cite the index for information from the first column since its position next to the bullet point and the very absence of a numeric index

scientificName	typeStatus	sex	country	state	city	locality	lat	lon	elevation	collector	date	institutionCode
Distictus tibialis	Holotype	1f	BRAZIL	PR	Curitiba	?	?	?	?	V.Graf leg.	1961	DZUP
Distictus tibialis	Paratype	1f	BRAZIL	PR	Curitiba	?	?	?	?	V.Graf leg.	1961	DZUP
Distictus tibialis	Other	1m	ARGENTINA	Corrientes	Las Marias	ca. Virasoro	?	?	?	C.Porter leg.	10–15 Nov 1969	FSCA
Distictus tibialis	Other	1f	ARGENTINA	Missiones	Dos de May	?	?	?	?	?	Feb–67	CNCI
Distictus tibialis	Other	1f	ARGENTINA	Punta Lara	?	?	?	?	?	H.Townes and M	27–Jan–66	CNCI
Distictus tibialis	Other	1m	ARGENTINA	Punta Lara	?	?	?	?	?	H.Townes and M	31–Jan–66	CNCI
Distictus tibialis	Other	1f	BOLIVIA	Noryungas	Coraico	El Bagante	?	?	150m	L.Masner leg.	18–Apr–97	CNCI
Distictus tibialis	Other	1f	BRAZIL	CE	Serra do Ara	?	?	?	850m	M.Alvarenga leg	19–May–69	USUC
Distictus tibialis	Other	1f	BRAZIL	ES	Alfredo Cha	Picadão	20°27'53"S	40°42'35"W	714m	C.O.Azevedo exp	8–15 Oct 2007	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Alfredo Cha	Picadão	?	?	710m	C.O.Azevedo exp	8–15 Oct 2007	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Cariacica	RES. Biol. Du	?	?	?	A.P.Aguiar expe	21–30 Oct 2005	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Conceição d	Propriedade	?	?	?	A.P.Aguiar expe	17–24 Mar 2007	UFES
Distictus tibialis	Other	1f	BRAZIL	ES	Conceição d	Propriedade	?	?	?	A.P.Aguiar expe	17–24 Mar 2007	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1f	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	6 Nov–03 Dec 2004	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	6 Nov–03 Dec 2004	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	?	M.Tavares expe	03–10 Dec 2004	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Santa Teres	Res. Biol. Au	19°54'37.7"S	40°33'12.1"W	764m	A.P.Aguiar expe	08–24 Oct 2016	UFES
Distictus tibialis	Other	1m	BRAZIL	ES	Santa Teres	Res. Biol. Au	19°55'16.4"S	40°33'13.5"W	775m	A.P.Aguiar expe	08–24 Oct 2016	UFES
Distictus tibialis	Other	3m	BRAZIL	ES	São Roque d	Alto Misteri	?	?	?	C.Waichert expe	2–11 Nov 2007	UFES

Fig. 3. Resulting order after multiple sorting according to the columns typeStatus, country, state, city and locality, in that order. This sequence of column names corresponds to the variable Sorting Order in the **Gridit** software.

scientificName	typeStatus	sex	country	state	city	locality	lat	lon	elevation	collector	date	institutionCode
Distictus tibialis	Holotype	1f	BRAZIL	PR	Curitiba	?	?	?	?	V.Graf leg.	1961	DZUP
ib	Paratype	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib
ib	Other	1m	ARGENTINA	Corrientes	Las Marias	ca. Virasoro	ib	ib	ib	C.Porter leg.	10–15 Nov 1969	FSCA
ib	ib	1f	ib	Missiones	Dos de May	?	ib	ib	ib	?	Feb–67	CNCI
ib	ib	ib	ib	Punta Lara	?	ib	ib	ib	ib	H.Townes and M	27–Jan–66	ib
ib	ib	1m	ib	ib	ib	ib	ib	ib	ib	ib	31–Jan–66	ib
ib	ib	1f	BOLIVIA	Noryungas	Coraico	El Bagante	ib	ib	150m	L.Masner leg.	18–Apr–97	ib
ib	ib	ib	BRAZIL	CE	Serra do Ara	?	ib	ib	850m	M.Alvarenga leg	19–May–69	USUC
ib	ib	ib	ib	ES	Alfredo Cha	Picadão	20°27'53"S	40°42'35"W	714m	C.O.Azevedo exp	8–15 Oct 2007	UFES
ib	ib	1m	ib	ib	ib	ib	?	?	710m	ib	ib	ib
ib	ib	ib	ib	ib	Cariacica	RES. Biol. Du	ib	ib	?	A.P.Aguiar expe	21–30 Oct 2005	ib
ib	ib	ib	ib	ib	Conceição d	Propriedade	ib	ib	ib	ib	17–24 Mar 2007	ib
ib	ib	1f	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib
ib	ib	1m	ib	ib	Domingos M	Mata Pico d	20°22'17"S	40°39'29"W	ib	M.Tavares expe	03–10 Dec 2004	ib
ib	ib	1f	ib	ib	ib	ib	ib	ib	ib	ib	6 Nov–03 Dec 2004	ib
ib	ib	1m	ib	ib	ib	ib	ib	ib	ib	ib	03–10 Dec 2004	ib
ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib
ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	6 Nov–03 Dec 2004	ib
ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	ib	03–10 Dec 2004	ib
ib	ib	ib	ib	ES	Santa Teres	Res. Biol. Au	19°54'37.7"S	40°33'12.1"W	764m	A.P.Aguiar expe	08–24 Oct 2016	ib
ib	ib	ib	ib	ib	ib	ib	19°55'16.4"S	40°33'13.5"W	775m	ib	ib	ib
ib	ib	3m	ib	ib	São Roque d	Alto Misteri	?	?	?	C.Waichert expe	2–11 Nov 2007	ib

Fig. 4. Repetitions in each column identified with the ib code for all columns (*strict* usage of the grid-setting technique).

can easily identify it. This is not in conflict with item 2 because dates can be recognized by their structure. So, for example, " • BRAZIL, ES,\..." implies that BRAZIL is the first information (`i=0`), but " • Apr 1985 • " implies that this collecting event differs from the previous one exclusively by the date, whichever value of `i` it might have. (**4**) The term *ibidem* is used exclusively in one rare situation: when the complete set of specimen data is identical for the holotype and one or more of the respective paratypes, such as in the example used in Figs 1–5 and Table 2 (see also line 17 of Fig. 5). These adjustments contribute to making the final ME even shorter and visually cleaner.

```
1  Holotype, 1f, BRAZIL, PR, Curitiba, 4?, V.Graf leg., 1961, DZUP
2  Paratype, 11ib
3  Other, 1m, ARGENTINA, Corrientes, Las Marias, ca. Virasoro, 3ib, C.Porter leg., 10-15 Nov 1969, FSCA
4  ib, 1f, ib, Missiones, Dos de Mayo, ?, 3ib, ?, Feb-67, CNCI
5  3ib, Punta Lara, ?, 4ib, H.Townes and M.Townes leg., 27-Jan-66, ib
6  ib, 1m, 8ib, 31-Jan-66, ib
7  ib, 1f, BOLIVIA, Noryungas, Coraico, El Bagante, 2ib, 150m, L.Masner leg., 18-Apr-97, ib
8  2ib, BRAZIL, CE, Serra do Araripe, ?, 2ib, 850m, M.Alvarenga leg., 19-May-69, USUC
9  3ib, ES, Alfredo Chaves, Picadão, 20°27'53"S, 40°42'35"W, 714m, C.O.Azevedo exped., 8-15 Oct 2007, UFES
10 ib, 1m, 4ib, 2?, 710m, 3ib
11 4ib, Cariacica, RES. Biol. Duas Bocas, 2ib, ?, A.P.Aguiar exped., 21-30 Oct 2005, ib
12 4ib, Conceição do Castelo, Propriedade Ribeirão do Meio, 4ib, 17-24 Mar 2007, ib
13 ib, 1f, 10ib
14 ib, 1m, 2ib, Domingos Martins, Mata Pico do Eldorado, 20°22'17"S, 40°39'29"W, ib, M.Tavares exped., 03-10 Dec 2004, ib
15 ib, 1f, 8ib, 26 Nov-03 Dec 2004, ib
16 ib, 1m, 8ib, 03-10 Dec 2004, ib
17 12ib
18 10ib, 26 Nov-03 Dec 2004, ib
19 10ib, 03-10 Dec 2004, ib
20 4ib, Santa Teresa, Res. Biol. Augusto Ruschi, 19°54'37.7"S, 40°33'12.1"W, 764m, A.P.Aguiar exped., 08-24 Oct 2016, ib
21 6ib, 19°55'16.4"S, 40°33'13.5"W, 775m, 3ib
22 ib, 3m, 2ib, São Roque do Canaã, Alto Misterioso, 3?, C.Waichert exped., 2-11 Nov 2007, ib
```

Fig. 5. Data from the spreadsheet in Fig. 4, excluding header and the column scientificName, copied and pasted into a text editor; resulting tabs (cells) replaced with comma. Sequences of "ib" and "?" in each row were grouped together with a preceding number (e.g., "11ib" in row 2) that indicates the total of subsequent repeats. Each row represents a unique collecting event. The final text generated from this file is shown in Table 2.

Table 2. Example of Material Examined text structured strictly with the grid-setting technique, as proposed in this work. Generated by grouping the collecting events (lines in the text in Fig. 5) in a single paragraph, separating each event with a bullet point. Selected variables highlighted in bold; m and f codes replaced by the respective male and female symbols.

Distictus tibialis

Holotype, 1♀, **BRAZIL**, **PR**, Curitiba, 4?, V.Graf leg., 1961, DZUP • **Paratype**, 11ib • **Other**, 1♂, **ARGENTINA**, **Corrientes**, Las Marias, ca. Virasoro, 3ib, C.Porter leg., 10–15 Nov 1969, FSCA • ib, 1♀, ib, Missiones, Dos de Mayo, ?, 3ib, ?, Feb–67, CNCI • 3ib, Punta Lara, ?, 4ib, H.Townes and M.Townes leg., 27–Jan–66, ib • ib, 1♂, 8ib, 31–Jan–66, ib • ib, 1♀, **BOLIVIA**, **Noryungas**, Coraico, El Bagante, 2ib, 150m, L.Masner leg., 18–Apr–97, ib • 2ib, **BRAZIL**, **CE**, Serra do Araripe, ?, 2ib, 850m, M.Alvarenga leg., 19–May–69, USUC • 3ib, **ES**, Alfredo Chaves, Picadão, 20°27′53″S, 40°42′35″W, 714m, C.O.Azevedo exped., 8–15 Oct 2007, UFES • ib, 1♂, 4ib, 2?, 710m, 3ib • 4ib, Cariacica, RES. Biol. Duas Bocas, 2ib, ?, A.P.Aguiar exped., 21–30 Oct 2005, ib • 4ib, Conceição do Castelo, Propriedade Ribeirão do Meio, 4ib, 17–24 Mar 2007, ib • ib, 1♀, 10ib • ib, 1♂, 2ib, Domingos Martins, Mata Pico do Eldorado, 20°22′17″S, 40°39′29″W, ib, M.Tavares exped., 03–10 Dec 2004, ib • ib, 1♀, 8ib, 26 Nov–03 Dec 2004, ib • ib, 1♂, 8ib, 03–10 Dec 2004, ib • 12ib • 10ib, 26 Nov–03 Dec 2004, ib • 10ib, 03–10 Dec 2004, ib • 4ib, Santa Teresa, Res. Biol. Augusto Ruschi, 19°54′37.7″S, 40°33′12.1″W, 764m, A.P.Aguiar exped., 08–24 Oct 2016, ib • 6ib, 19°55′16.4″S, 40°33′13.5″W, 775m, 3ib • ib, 3♂♂, 2ib, São Roque do Canaã, Alto Misterioso, 3?, C.Waichert exped., 2–11 Nov 2007, ib.

Corrected nesting

The nesting ambiguity characterized in the '*Limitations*' section above can be resolved by indicating how many variables (columns in Fig. 1) are implied *before* the level being modified. The most discreet way to do this is apparently by adding in front of the colon of the term in question as many other colons as there are previous levels. Thus, the representation "Info::" indicates that it is the second variable of two nested variables; "Info:::", the third of three, and so on. Since nesting is somewhat embedded in the *implicit* approach these techniques are mutually exclusive, and therefore the proposed usage of colons in both will never collide.

In the case of the two examples mentioned in the '*Limitations*' section above, the correction should be applied as follows (highlighted):

BRAZIL: 1♀, Espírito Santo: Santa Maria de Jetibá: Clarindo Krüger Farm, 6 Dec 2002 • Specimen 2 • Specimen 3 • 2♂♂, Conceição do Castelo ::: Ribeirão do Meio, 17–24 Mar 2007 • Specimen 5 • Specimen 6 • etc.

BRAZIL: 1♀, Espírito Santo: Santa Maria de Jetibá: Clarindo Krüger Farm, 6 Dec 2002 • Specimen 2 • Specimen 3 • 2♂♂, Pará :: Serra Norte, 17–24 Mar 2007 • Specimen 6 • Specimen 6 • etc.

The repetition of colons above provides the necessary information to interpret "Conceição do Castelo" as equivalent to "Fazenda Clarindo Krüger", and "Pará" as equivalent to "Espírito Santo", based exclusively on the structure of the text, without the need for external knowledge about the nature of each information.

In case the next label contains new information for multiple nested levels except the first one, only the first different level needs to be marked, avoiding situations like "NewState:: NewCity::: NewLocality::::" in favor of "NewState:: NewCity: NewLocality:". If the change occurs for the first term, nothing needs to be done.

Although useful on its own and producing even better results in association with the *strict* approach, it is important to note that nesting is already embedded in and cannot be used with the *implicit* approach.

Mixed approach

Although the result produced solely by the core algorithm for grid-setting (Figs 1–5 and Table 2) is already entirely self-sufficient and generates probably the easiest result to interpret, this usage will act on *all* data. This means that if some variable needs to be explicitly displayed for all specimens (e.g., as is commonly the case for specimen number, sex, and the institution code) then the result might be inadequate. The grid-setting technique is however compatible with this possibility, sufficing to limit coding to the desired variables. In addition, grid-setting is also compatible with the grouping of specimens with identical data, with the use of the corrected *nesting* technique (see previous section), and others.

The most notable practical result of any of the approaches discussed above is that the ME text literally becomes an encoded datasheet, which can be easily converted back to a spreadsheet. Due to the simplicity of the coding, the conversion can be done manually, or by find/replace commands in a text editor, or using a specific software, such as the one presented in the '*New software*' section below.

Compression

One of the goals valued in the formatting of the ME section is to minimize redundancy, reducing the size of the text to be published. The grid-setting technique has excellent performance here. For comparative purposes, the text from all ME sections published in Supeleto *et al.* (2019) (for species with 3 specimens or more, comments removed) add up to 15 702 characters, while the same information formatted with the grid-setting technique using the *implicit* approach generates a text with 11 328 characters, 4258 fewer,

or 28% smaller. If only the core algorithm is used (as in Figs 1–5), the final text is 13% smaller. If this is combined with the use of the corrected *nesting* for the variables country, state, and city, the total ME text size is 13 079 characters (17% smaller). These results show that grid-setting consistently outperforms the degree of compression obtained with traditional techniques.

The maximum compression that can be achieved with any of these techniques arises, in theory, by ordering and then sorting the variables (columns) from those with fewer to those with more unique values. For full efficiency this would have to be applied to each ME section separately, but this is just a theoretical consideration to help clarify the grid-setting logic. In practice, it will often be more important that the order of the variables be user-defined than strictly optimized for maximum compression. The use of grid-setting for formatting each ME separately in the same work would however only make them differently organized in relation to each other, but not incompatible – all ME sections could still be consistently converted back to a spreadsheet.

It is therefore important to note that grid-setting will generate different degrees of compression as a function of two user-defined choices: (**1**) Display Order, the order in which the variables appear, and (**2**) Sorting Order, the definition of which variables, and in which order, will be used to perform multiple sorting of the data. The same Display Order can produce higher or lesser compression depending on the diversity of values in each column and on the Sorting Order chosen. For example, if the first and second variables are respectively the coden (= institution code) and country, the final compression will tend to be larger / better if there are fewer unique codens than country names, but it will be less efficient if the variety of codens is greatest. The result will depend on the interaction between all the variables and the selected choices.

Best compression results seem to be produced with the *implicit* approach, but other than that the reasoning to choose between the different techniques or approaches seems to be more stylistic than technical.

The Automatex software (Brown 2013, currently at http://phorid.net/automatex/auto.php) was a pioneer in the production of formatted ME lists for publication, applying traditional techniques on a fixed number of variables (columns). For the ME data in Supeleto *et al.* (2019), filtered for the variables accepted by Automatex (Sup. file 2), the text generated with the "Format 3" option resulted in 16 697 total characters, which is greater than the number of characters for the respective data in the publication itself (15 825). With the grid-setting technique applied to the same data submitted to Automatex, the resulting text had 10 686 characters, 36% fewer. However, with the used file Automatex generated a text with some duplicated values (not counted) and other small problems of omission and / or repetition, which obviously interferes with the accuracy of the comparison presented above. It is not the aim of this work to present a detailed analysis of Automatex.

New software

The core algorithm of the grid-setting technique can be quickly applied with the aid of spreadsheets and a text editor, as for example shown in Figs 1–5, but its implementation in conjunction with other approaches is more time-consuming. For this, the **Gridit** application for desktop, written in Python, incorporates all the new ideas and fixes discussed in this work, and is freely available at https://www.systaxon.ufes.br/grd, with instructions. The software receives as input an Excel or CSV file containing label data, with specimens in rows and variables in columns, e.g., as illustrated in Fig. 1. The number of variables (columns), the name of each one and their order are user-defined, which allows for customized results. There are only four mandatory columns and column names (scientificName, typeStatus, sex, institutionCode), but the total is unlimited. These names and their meaning follow the DwC standard (https://dwc.tdwg.org/terms/), but all other column names and data types are ultimately user-defined. The variable sex also accepts a

combination of individualCount + sex, e.g., 3f2m to record 3 females and 2 males. Cells with no available data can be left empty, without any special markings, or filled in with a "?".

The `,():` characters are used by **Gridit** to build the formatted ME, and their use should therefore be avoided in the input file. If found in the input, they will be replaced respectively by `;[]/` during processing. Numbers must therefore not contain a comma.

The new software allows the use of the grid-setting technique alone or in conjunction with traditional techniques, in addition to allowing certain variables to never be abbreviated or coded (as commonly adopted, for example, for the number of specimens, sex and institution code). The various possible combinations of settings in **Gridit** generate customized ME texts, but all results are compatible with the conversion back to Excel or CSV spreadsheet format with the Convert option. For conversion, the ME of a given taxon must be copied and pasted in the respective text area and submitted (button Submit). Ideally, the Display Order list should be provided in the first line, to generate the respective column names. If not provided, the terms "Var01", "Var02", etc. will be used instead.

Notes and edits

Any variable or value can be annotated or edited in the input file or even in the final ME text, provided that the four reserved characters `,():` are avoided in the input. Annotations added to the final ME text will be included in the Excel spreadsheet generated by **Gridit**.

Discussion

In addition to the logical and structural improvements provided, there are other important advantages, listed below, with the adoption of the grid-setting approach. (**1**) Grid-setting offers ample freedom of choice, accepting any amount and nearly any kind of text information while still generating ME sections within a universally interpretable format. (**2**) The result produces excellent text compression and (**3**) generates a text which is both readable and ready to be consistently converted back to spreadsheet format. (**4**) The ease of converting to a spreadsheet provides a much more efficient method of review for authors, reviewers, and editors, supporting an efficient audit and therefore higher final quality of publications; for the same reason, (**5**) end-user access to the ME data is also facilitated. Furthermore, (**6**) data in spreadsheet form can be easily imported from and exported to databases. (**7**) The availability of software for encoding and decoding is another important advantage, but formatting or retrieving the data is not tied to the software itself, which is just a temporary resource; the most important point is that the founding ideas are outlined, ensuring that encoding and decoding will always be fully possible.

The main disadvantage is the repetition of coding characters, such as "?" and "ib". But this problem is equally common with traditional techniques, which also use codes (like *ibidem*, colons, dashes, etc.) and are less efficient in avoiding the repetition of terms, as shown. At the same time, the codes used by grid-setting are short, simple, and familiar: a "?" for missing or inapplicable information, and "ib" or the syntax `i:text` to encode repetitions. The proposed correction for the nesting technique can also display double or triple or even more chained colons, but there are few occurrences, and the visual impact of these extra colons is minimal.

A potential limitation is the eventual occurrence of label data that are identical to the "ib", "2ib", etc. codes, which would conflict with the *strict* approach. This could be worked around using the *implicit* approach, or with creativity, e.g., by surrounding the code with hyphens or brackets, etc. But in tens of thousands of labels checked, not even a single occurrence was found, indicating that this is a remote possibility.

Essential specimen information can now be retrieved easily from hyperlinks; for example, CETAF (The Consortium of European Taxonomic Facilities) member organisations have implemented a system of persistent identifiers for objects in collections (Güntsch *et al.* 2017). These Uniform Resource Identifiers follow a defined syntax and are potentially human readable, following a logical grammar. Some journals, such as *Biodiversity Data Journal*, now only accept specimen data in tabulated, machine-readable formats and the ME sections are generated automatically. These DwC tables can be directly exported to GBIF. DwC tables, in turn, are electronic files, and therefore cannot be in the printed publication itself. Accordingly, the BDJ itself states that "The table does not replace or exclude the detailed listing of specimen data (occurrences, label data) in the Materials examined sections of the taxon treatments" (BDJ 2022). Besides representing considerable additional work, preparing a DwC-ready table also requires strict adhesion to elaborate rules and standards, explained by the BDJ (2022) with 1429 words, two videos, a template spreadsheet, and even some programming code. This is likely to be time-consuming and prone to human error, discouraging its very audience. Using a program to format the ME data, on the other hand, requires little from the user. In the case of the grid-setting approach, it is also essential to note that the output is, at the same time, both text *and* table (= encoded spreadsheet), which solves the need for a compact, human-readable text in the publication and a table with data formatted to be accurately machine-readable.

A DwC archive (see Darwin Core Maintenance Group 2021) could also be used as ME. In its simplest form, it consists of rows of comma-separated values with the first row (column names) containing DwC term names, such as scientificName, typeStatus, etc. The archive itself cannot be incorporated in the printed paper, but its content is simple enough to be printed either raw or in tabular form. Note, however, that this is roughly equivalent to the first step of the grid-setting technique and would produce large, highly redundant texts (or tables). Furthermore, tables take much more space than text and could be cumbersome to fit after the treatment of each taxon.

Since the grid-setting technique is flexible (any data, any order, any size) and easy (load & convert in **Gridit**), it has good potential to bridge the ME text in the publication to the automatic generation of files in different formats. New formats would require additions to the software itself, with little or no impact on user-side demands in formatting. For example, to generate DwC tables, it should ideally suffice to use standard DwC names in **Gridit**.

It is, however, important to maintain that the present work is strictly presenting a new idea in its original form; it is ready for immediate practical use and comes with a tool to allow anyone to use it. Its adoption or incorporation into other efforts seem possible and potentially useful, but that is another step.

Acknowledgments

Two anonymous reviewers contributed with valuable comments and suggestions.

References

Aguiar A.P. 1998. Revision of the genus *Hemistephanus* Enderlein, 1906 (Hymenoptera, Stephanidae), with methodological considerations. *Brazilian Journal of Entomology* 41: 343–429.

Aguiar A.P. 2013. Publishing large DNA sequence data in reduced spaces and lasting formats, in paper or PDF. *Zootaxa* 3609 (6): 593–600. https://doi.org/10.11646/zootaxa.3609.6.5

Aguiar A.P. & Ramos A.C.B. 2011. Revision of *Digonocryptus* Viereck (Hymenoptera: Ichneumonidae: Cryptinae), with twenty six new taxa and cladistic interpretation of two species complexes. *Zootaxa* 2846 (1): 1–98. https://doi.org/10.11646/zootaxa.2846.1.1

Anderson N.R., Tarczy-Hornoch P. & Bumgarner R.E. 2006. On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics* 7: 260. https://doi.org/10.1186/1471-2105-7-260

BDJ – Biodiversity Data Journal. 2022. Instructions for Authors.
Available from https://bdj.pensoft.net/about#For-authors [accessed 30 Jun. 2022].

Brown B.V. 2013. Automating the "Material examined" section of taxonomic papers to speed up species descriptions. *Zootaxa* 3683 (3): 297–299. https://doi.org/10.11646/zootaxa.3683.3.8

Brown B.V. 2021. Automatex – Automated Material Examined.
Available from http://phorid.net/automatex/auto.php [accessed 20 Feb. 2022].

Chester C., Agosti D., Sautter G., Catapano T., Martens K., Gérard I. & Bénichou L. 2019. *EJT* editorial standard for the semantic enhancement of specimen data in taxonomy literature. *European Journal of Taxonomy* 586: 1–22. https://doi.org/10.5852/ejt.2019.586

Darwin Core Maintenance Group. 2021. Darwin Core text guide. Biodiversity Information Standards (TDWG). Available from http://rs.tdwg.org/dwc/terms/guides/text/2021-07-15 [accessed 30 Jun. 2022].

Güntsch A., Hyam R., Hagedorn G., Chagnoux S., Röpert D., Casino A., Droege G., Glöckler F., Gödderz K., Groom Q., Hoffmann J., Holleman A., Kempa M., Koivula H., Marhold K., Nicolson N., Smith V.S. & Triebel D. 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database* 2017: 1–9. https://doi.org/10.1093/database/bax003

Kenyon J. & Sprague N.R. 2014. Trends in the use of supplementary materials in environmental science journals. *Issues in Science and Technology Librarianship* 75. https://doi.org/10.5062/F40Z717Z

Pop M. & Salzberg S.L. 2015. Use and mis-use of supplementary material in science publications. *BMC Bioinformatics* 16 (237): 1–4. https://doi.org/10.1186/s12859-015-0668-z

Seeber F. 2008. Citations in supplementary information are invisible. *Nature* 451: 887.
https://doi.org/10.1038/451887d

Supeleto F.A., Santos B.F. & Aguiar A.P. 2019. Revision of *Distictus* Townes, 1966 (Hymenoptera, Ichneumonidae, Cryptinae), with descriptions of ten new species. *European Journal of Taxonomy* 542: 1–64. https://doi.org/10.5852/ejt.2019.542

Telnov D. 2020. A revision of the Maechidiini Burmeister, 1855 (Coleoptera: Scarabaeidae: Melolonthinae) from the Indo-Australian transition zone, and the first record of the tribe west of Wallace's Line. *European Journal of Taxonomy* 721: 1–210. https://doi.org/10.5852/ejt.2020.721.1127

Zanella F.C.V., Oliveira M.L. & Gaglianone M.C. 2000. Standardizing lists of locality data for examined specimens in systematic and biogeography studies of new world taxa. *Biogeographica* 76: 145–160.

Manuscript received: 9 March 2022
Manuscript accepted: 3 October 2022
Published on: 15 December 2022
Topic editor: Tony Robillard
Section editor: Frank Zachos
Desk editor: Radka Rosenbaumová

Supplementary files

Supp. file 1. Excel file with spreadsheet data from the ME sections, except comments and annotations, for species with three or more specimens treated in Supeleto *et al.* (2019), totalling 6 species, 293 specimens and 19 variables (columns).
https://doi.org/10.5852/ejt.2022.852.2007.8207

Supp. file 2. CSV file with adaptation of the previous file for processing with the Automatex software.
https://doi.org/10.5852/ejt.2022.852.2007.8209