Computers & Geosciences 35 (2009) 862–868

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Charting taxonomic knowledge through ontologies and ranking algorithms

Robert Huber[a], Jens Klump[b,*]

[a] MARUM, University of Bremen, Bremen, Germany
[b] Data Centre, GeoForschungsZentrum Potsdam, Telegrafenberg, 14473 Potsdam, Germany

ARTICLE INFO

Article history:
Received 30 April 2007
Received in revised form
29 January 2008
Accepted 3 February 2008

PACS:
07.05.Rm
89.20.Hh
07.05.Kf
91.65.n

Keywords:
Taxonomy
Paleontology
Knowledge management
Visualization

ABSTRACT

Since the inception of geology as a modern science, paleontologists have described a large number of fossil species. This makes fossilized organisms an important tool in the study of stratigraphy and past environments. Since taxonomic classifications of organisms, and thereby their names, change frequently, the correct application of this tool requires taxonomic expertise in finding correct synonyms for a given species name. Much of this taxonomic information has already been published in journals and books where it is compiled in carefully prepared synonymy lists. Because this information is scattered throughout the paleontological literature, it is difficult to find and sometimes not accessible. Also, taxonomic information in the literature is often difficult to interpret for non-taxonomists looking for taxonomic synonymies as part of their research.

The highly formalized structure makes Open Nomenclature synonymy lists ideally suited for computer aided identification of taxonomic synonyms. Because a synonymy list is a list of citations related to a taxon name, its bibliographic nature allows the application of bibliometric techniques to calculate the impact of synonymies and taxonomic concepts. TaxonRank is a ranking algorithm based on bibliometric analysis and Internet page ranking algorithms. TaxonRank uses published synonymy list data stored in TaxonConcept, a taxonomic information system. The basic ranking algorithm has been modified to include a measure of confidence on species identification based on the Open Nomenclature notation used in synonymy list, as well as other synonymy specific criteria.

The results of our experiments show that the output of the proposed ranking algorithm gives a good estimate of the impact a published taxonomic concept has on the taxonomic opinions in the geological community. Also, our results show that treating taxonomic synonymies as part of on an ontology is a way to record and manage taxonomic knowledge, and thus contribute to the preservation our scientific heritage.

1. Introduction

Since the beginning of geology as a modern science, biostratigraphy has been at the core of the discipline as a tool to establish chronostratigraphical relationships between strata. Besides their application in biostratigraphy, fossilized remains of organisms have become useful tools in the investigation of past climates and ecologies. Assigning a correct age to a stratum through biostratigraphy requires the correct identification of fossils used as chronostratigraphical markers, but the proper identification of fossil species is a common problem in biostratigraphy.

For a geologist not specialized in the taxonomy of the species in question it is difficult to follow the existing

[*] Corresponding author. Tel.: +49 331 288 1702; +49 331 288 1703.
E-mail addresses: rhuber@wdc-mare.org (R. Huber), jklump@gfz-potsdam.de (J. Klump).

0098-3004/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.02.016

taxonomic literature and correctly identify a fossil species. It is particularly difficult for non-taxonomists to find those taxa for which related data may exist and which might have been treated previously as synonyms of other taxa (Nimis, 2001), which makes it difficult to find all matching data in the literature. Furthermore, large databases, such as PANGAEA (http://www.pangaea.de) or CHRONOS (http://www.chronos.org), store taxon related data in their historical context 'as published' and leave the original classification by the data set author unchanged. However, queries on such databases need to include all known synonyms or else may deliver incomplete results. Therefore, it is essential for scientists to have access to taxonomic information that is organized effectively, so that the required data sets can be quickly identified and retrieved.

The common method of providing taxonomic information in paleontology is by publishing primary taxonomic data in journals or monographs. Traditionally, the careful preparation of synonymy lists is of great importance in these publications. During the last decades, a community wide agreement on the form of synonymy lists has been reached and most journals now demand the use of the 'Open Nomenclature' notation (Bengtson, 1988; Matthews, 1973). The use of the Open Nomenclature notation allows working with taxonomic classifications that are unclear and allows the author to comment on the identification of a specimen by other authors.

For instance, the use of the question mark in the name *Agenus? album* indicates a uncertainty of identification at the genus level.

> 1895 Agenus? album Aulus.-Bruno, Monogr. Agenidae, S.12 Taf. 3 Fig. 2.

Synonymy list annotations may also be used to review species identifications by other authors. For example, a synonymy list entry

> v 1895 Agenus album Aulus.-Bruno, Monogr. Agenidae, S.12 Taf. 3 Fig. 2.

expresses that the author has seen (v = vidimus) the material which relate to the cited work and the author agrees on this determination. In contrast the entry:

> non 1895 Agenus album Aulus.-Bruno, Monogr. Agenidae, S.12 Taf. 3 Fig. 2.

expresses that the cited species (or its illustration on page 12, plate 3, Fig. 2) cannot be compared with the discussed species or that the specimen shown in the illustration was wrongly identified.

Commonly, such a list contains all known occurrences of specimen in the literature matching the author's concept of a specific taxon. Synonymy lists in paleontology therefore rather express the taxonomic opinion or concept of the list's author on a specific taxon than represent the objective synonymy of a taxon and are therefore difficult to interpret for non-taxonomists. In addition, much of this information is scattered throughout the primary literature and sometimes difficult to access.

Open Nomenclature synonymy is highly formalized and therefore it is well suited for knowledge management. In a synonymy list the author describes a taxonomic concept and how it relates to the taxonomic concepts of other authors. A synonymy list is therefore part of an ontology as an explicit specification of a conceptualization (Gruber, 1993). In this ontology the individual taxonomic concepts serve as nodes in a network of directed, non-cyclical graphs inside a formal framework of rules. The body of synonymy lists on a group of organisms can be seen as an ontology that formally describes both corresponding and differing opinions on taxonomic concepts. These ontologies can be used to study how well founded synonymies are by using ranking algorithms to rank the quality of an identification. The resulting network of graphs is a chart of our knowledge on the identification of species, as it is recorded in synonymy lists in the scientific literature. It maps the strengths and weaknesses of our taxonomic identifications, and point out gaps in our taxonomic knowledge (Fig. 1. The expression of taxonomic synonymies as an ontology thus helps us to analyze domain knowledge, it enables reuse of domain knowledge, and it makes domain assumptions explicit.

In the following sections we will present some experiments on the data stored by our online taxonomic information system. TaxonConcept can be found at http://taxonconcept.stratigraphy.net. It is a system to store and provide taxonomic concepts expressed by authors of taxonomic literature without judging the validity of the authors' taxonomic concepts, even though some authors' concepts may be contradicting each other. TaxonConcept allows to store Open Nomenclature synonymy lists and now holds a sufficient number of records to perform computational experiments on this body of taxonomic information. We will demonstrate the application of this system on examples from the taxonomy of planktonic foraminifera and discuss the results from the application of ranking algorithms on published synonymies. Furthermore, we will show how taxonomy visualization and

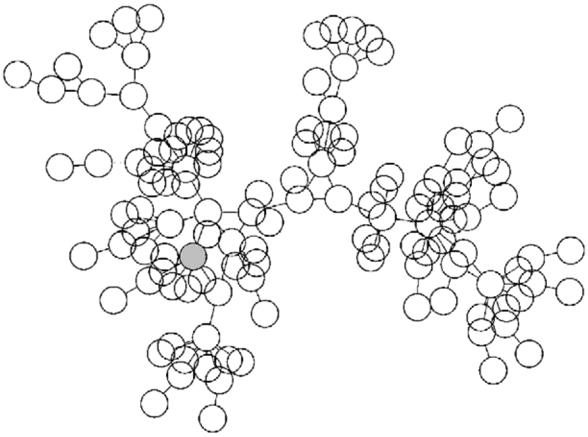

Fig. 1. Synonymy network for *Subbotina triangularis* based on TaxonConcept database entries. This synonymy network was calculated by linking all synonymy lists of *S. triangularis* or containing *S. triangularis* and the synonymy lists of all taxa listed in these synonymy lists, respectively. Approximately 120 taxon names are related through this network links to *S. triangularis*. A highlighted circle marks the position of *S. triangularis* within this network.

synonymy ranking can be used to assist researchers to manage and analyze the large amount of printed or data-based taxonomic information. Readers may run their own experiments using our web-based application and database in which we captured the taxonomic opinions on select groups of species from the taxonomic literature.

Capturing and analyzing taxonomic literature as a way of charting and preserving taxonomic knowledge without *prima facie* appraisal of the sources contradicts the notion of taxonomic authorities which has been alive in the paleontological community since the mid-19th century. The results of our computational experiments show that the output of the proposed ranking algorithm gives a good estimate of the impact a published taxonomic concept has on the taxonomic opinion in the geological community. Also, our results show that viewing taxonomic synonymies as part on an ontology is a way to record and manage taxonomic knowledge, which is an important part of preserving our cultural and scientific heritage.

There are several technologies available to represent ontologies, the most common representations are based on the standards RFD, OWL and SKOS specified by the W3C consortium (http://www.w3c.org). In our web-based taxonomic information system TaxonConcept data are stored in a relational database (http://taxonconcept.stratigraphy.net). It is planned to encode the synonymy information in RDF and provide it, together with taxonomic concept information, to outside systems through REST and SOAP web services. The authors have closely followed the activities of the Taxonomic Databases Working Group (TDWG, http://www.tdwg.org/). The TaxonConcept web services will be designed to conform with the taxonomic data transfer schema proposed by TDWG.

2. Ranking synonymies

Essentially, a synonymy list is a list of citations related to a taxon name, annotated in a specific way to express an author's opinion on these synonymies. The bibliographic nature of a synonym lists suggests that bibliometric techniques could be used to investigate the ranking of synonymy entries and used taxa, respectively. Bibliometry is a disputed topic and recent discussion in the literature suggests that impact factors may not be applicable in taxonomy (Krell, 2000). Instead of simply applying such a citation index we calculate a rank based on relations and links between synonymy lists and taxon names using a modification of the Internet page ranking algorithm, also known as the PageRank algorithm (Page et al., 1998; Brin and Page, 1998).

The PageRank algorithm assigns a numerical weight to web sites as a measure of their importance or popularity. PageRank exploits the interlinked nature of the world wide web and assumes that the popularity of a web site is reflected by the number of links pointing to a particular site, similar to bibliometric citation analysis (Garfield, 1972). However, PageRank captures more than the number of links pointing to a web page, it also analyze the page that points to the site. PageRank thus assumes that the importance of a web page is determined by the number and the importance of web pages pointing to it and thereby extends the bibliometric principle.

The idea behind TaxonRank (TR) is that some authors are more often cited than others and their species identifications might therefore have a stronger impact on a 'common taxon concept' than others. This can be the result of many factors, e.g. the quality of species illustrations, the reputation of the author or the availability of a publication. In analogy to PageRank we state that the rank of a synonymy list is determined by the rank of the synonymy cited in a particular synonymy list. To calculate the rank of a specific taxon name within a synonymy we assume that taxon names used in high ranked synonymy lists have a high rank. The bibliometric algorithm is then modified by the certainty of a species identification based on the Open Nomenclature notation used in the synonymy list, as well as other synonymy specific criteria which will be explained later.

2.1. Potential synonymy

Relations among taxon names can be quite complex. Differences in spelling and other variations, such as the optional use of a subgenus name, represent alternatives of a specific taxonomic name. A taxon name can be listed in a synonymy list of another taxon, and can in turn contain other taxon names in some of its synonymy lists. All these taxon names can again have alternative names and contain other taxon names in their synonym lists, and so on. This synonymy of synonyms represents a complex ontologic network of taxon names. We shall call this ontology, or set of related taxon names, potential synonyms P of a distinct taxon t_i where $t_j \in P(t_i)$ when t_j is an alternative name of any $t_k \in P(t_i)$ or t_j is listed in a synonym list of any $t_k \in P(t_i)$ and $t_i \in P$. P represents the set of all taxon names which will be considered for the calculation of the rank of a specific taxon name. These ontologies can be visualized and drawn as a network of directed graphs. For technical reasons the synonymy ontologies are currently still drawn as undirected graphs in our online application of TaxonConcept (Fig. 1).

2.2. SynonymyRank

For our synonymy ranking experiment on P we define a SynonymyRank (SR) as a variation of the PageRank algorithm. This algorithm is based on concepts and topology of the world wide web and therefore we first need to define 'pages' and 'links' between these pages. To apply the PageRank to synonymy lists, we define a synonymy list S_i for a taxon t published by author i as a Internet page containing an arbitrary number of pairs of synonymous names *syn* and the cited publication *doc* listed by author i as $l\{syn, doc\}$.

We further define such pairs as synonymy list entries. The order of such a synonymy list entries $o(\{syn, doc\}, S_i)$ is in turn defined by the publication year of the document containing the synonymy list.

A link within a synonymy list from S_i to S_j is present when the synonymy list entry $l\{syn, doc\}$ exists in S_i and in

S_j and $o(\{syn, doc\}, S_j) > o(\{syn, doc\}, S_i)$ and $syn \in P$, i.e. a synonym/publication pair has been used previously by the author of an older publication.

The set of all synonymy lists S_j of P is L_{S_i} and the number of links from S_j is N_j. Further, any pair $l\{syn, doc\} \ni L_{S_i}$ is defined as a synonym list having itself as only synonym list entry. The distance $dist_j$ between taxon t_j and taxon t_i within the ontological graph network P is calculated after Floyd (1962) and determines the strength $strength(S_i, S_j) = 1/dist_j$ of a link l leading from S_i to S_j.

The SR of a specific synonymy list S_i is defined analogous to the PageRank algorithm and calculated recursively using

$$\forall i \; SR_{k+1}(S_i) = \sum_{S_j \in L_{S_i}} \frac{strength(S_i, S_j) * SR_k(S_j)}{N_j} \qquad (1)$$

The rank of a synonymy list S_i is thus defined as the sum of the ranks of all synonymy lists pointing to list S_i, divided by the number of all links on S_j.

2.3. TaxonRank

To calculate the rank of a specific taxon within a synonymy list we included a pre-ranking derived from the Open Nomenclature notations used in the synonymy list of the author as well as in the citated source. In our ranking experiment we regard certain Open Nomenclature tags as indicators of confidence with respect to a species identification and assign a scalar value to each tag.

This scalar value is used as a confidence factor for each species determination of a synonymy list and represents a measure of the taxonomic expert knowledge. The rank of a taxon occurrence t_{i_k} in a synonymy list is calculated as the product of this confidence factor and the synonymy list rank SR:

$$TR(t_i, S_j) = \frac{\sum_{t_i \in S} SR(S_j) * conf(t_{i_k})}{n(t_i)} \qquad (2)$$

We can then calculate the rank of a taxon within a synonymy list $TR(t_i, S_j)$ using Eq. (2) as the mean of all instances of a taxon under consideration of the confidence factor in Table 1. As a first approach to determine the rank for a taxon as an element of P we can calculate the total TR as the sum of all $TR(t_i)$ of any synonymy list SR.

Table 1
Confidence factors applied to Open Nomenclature tags used to calculate TaxonRank

Tag	Meaning	Factor
sp or ?	Uncertain identification	0.5
?	Reference has doubtful identification	0.5
cf	Provisional identification (confer)	0.5
aff	Provisional identification (affinis)	0.5
non	Reference excluded from concept	0.25
p	Reference applies only in parts to concept (partim)	0.5

3. Results and discussion

The TR algorithm is based on the assumption that taxonomic concepts in one publication may have a stronger influence on the common taxonomic opinion than others. The TR is calculated iteratively by using the SR of the list entries in the synonymy list cited. As a result, the rank of a synonymy list is directly dependent on its precursors, and thus older synonymy lists will have higher impact on the TR score than those published more recently.

The higher scores of older taxonomic publications contradict the common expectation that the most comprehensive taxonomic expertise is found in the most recent literature. Even though this seems counter-intuitive, it is a result of the way taxonomic literature is commonly used. It takes time for new knowledge to spread, even among taxonomists. Meanwhile, especially non-taxonomists—or ‘name users’ (Nimis, 2001)—form the community that defines a ‘taxonomic common sense’. This community taxonomic concept, reflecting the common usage of taxon names, might not at all be influenced by the latest taxonomic insights. In addition, taxon related data are often published without a taxonomic section explaining the taxonomic concept adopted by the author. Taxon names are often used incompletely, i.e. without giving author and year, they are misspelled or—even worse—abbreviated. More and more data are being extracted from the literature and incorporated into large earth science databases, bearing exactly this weakness in their scientific documentation. This lack of proper taxonomic documentation further complicates the re-use of these data as it remains ambiguous which taxon the authors meant when they used a certain taxonomic name. Besides this ambiguity, the use of taxon related data is further complicated by changes in taxonomic classifications and resulting unclear synonymies.

Currently, a scientist looking for data relating to a specific taxon has to decide, which synonym names to be include in a search query in order to get the best possible and most complete result from a database. Are all relevant names included to compile the best possible biostratigraphy? Are there any isotope data besides those for taxon X and are all necessary taxa included for the planned paleoclimatological interpretation? In the Internet age some taxonomic assistance can surely be expected on any wired desktop by data mining technologies on biodiversity databases. TR is a first attempt to find synonymy top candidates for further investigation.

TR does not calculate the probability for a taxon A to be synonymous of taxon B. The calculated rank rather reflects a usage ranking of taxon names. A ranking of names frequently used in the context of a distinct taxon name and gives an indication which other taxa are important in relation to a specific taxonomic concept. Highly ranked taxa should therefore not expected to be ‘objective’ synonyms but should alert the user to taxonomic uncertainties that require additional information to be, ideally by consulting the related taxonomic literature.

How closely do automatically compiled synonymy lists reflect the taxonomic opinions expressed in the literature and present as implicit knowledge in the taxonomic community? In this paper we present a series of tests on TR and discuss how close to this vision we have come.

For our ranking test runs were performed on a body of approx. 6000 synonymy list entries on Cenozoic planktonic foraminifera, mostly on Paleocene to recent species. This relatively modern group is well studied and our database now contains high quality synonymy lists from publications covering the last 50 years. We chose the Paleogene Planktonic Foraminifera Working Group's (Olsson et al., 1999) synonymies on Paleocene foraminifers to test TR. This publication presents a recent taxonomic discussion on paleocene planktonic foraminifers and members of the working group are the regarded as foremost experts in this field.

As a first test run we chose the planktonic foraminifera species *Subbotina triangularis* (White, 1928). This example illustrates nicely how combining synonymy lists from several authors on related taxa can result in a complex taxonomic knowledge network: 123 taxon names are connected with this species name. These networks may contain both corresponding as well as differing taxonomic opinions, which can be extracted as ontologies and visualized as knowledge maps. Here we show only a simplified relation chart (Fig. 1). Despite the high numbers of related taxon names, some distinct clusters of taxonomic concepts can be identified. In the next step we performed a TR on *S. triangularis*. Fig. 2 shows a graphical representation of the TR results. The size of the circles in Fig. 2 reflect the rank of the synonym candidate taxa, which are also listed in Table 2. All highly ranked taxa plot in the cluster around the target species, which indicates that TR correctly identifies the most important taxa.

Testing the quality of TR is simplified by the fact that the most recent literature has little influence on the rank of a specific taxon. We can therefore test TR by comparing the calculation results—the top ranked taxa—with those taxa listed in the Olsson group's synonymies (Olsson et al., 1999). To make the ranking results comparable, all TRs are normalized to 100. From several earlier experiments with TR we found that the distribution of TR values shows a characteristic pattern. In most cases we observed a marked decrease in the assigned TR from values of $TR = 70\ldots60$ and another steep decrease at $TR = 30\ldots40$. We therefore regard ranks with $TR < 30$ as most probably not

Table 2
Example test run result

Taxon name	TaxonRank
Globigerina triloculinoides	100
Globigerina pseudotriloba	97.3
Globigerina gerpegensis	84.7
Subbotina triangularis	84.7
Globigerina uruchaensis	84.7
Subbotina triangularis	84.7
Globigerina triangularis	74.1
Globigerina inaequispira	61.1
Globigerina bacuana	47.3
Subbotina patagonica/triangularis group	37.4
Eoglobigerina trivialis	5.3
Subbotina velascoensis	5.3
Parasubbotina varianta	4.2
Subbotina patagonica	4.2

TaxonRank calculated ranks of top rated taxa related to *S. triangularis*. Results are normalized to 100.

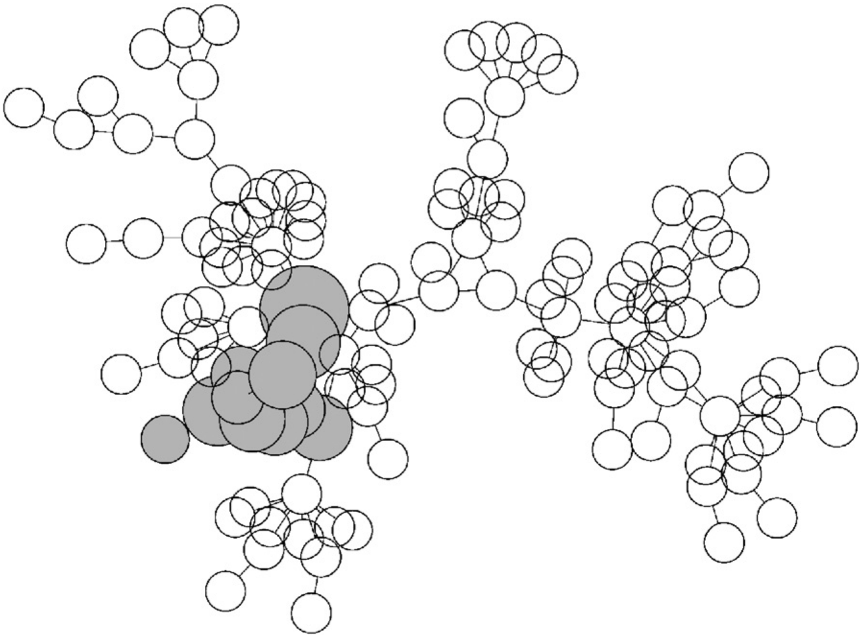

Fig. 2. Synonymy network for *Subbotina triangularis*. Top ranked taxa in this network are highlighted and are found clustered around *S. triangularis*. Circle sizes are drawn in proportion to the calculated TaxonRank.

significant and taxa showing ranks $TR>30$ as indicating potential synonyms.

For our above example of *S. triangularis* (Table 2), we find all taxa, for which a rank $TR>30$ was calculated, in the synonymy list of the Olsson group (Olsson et al., 1999). Only in one case a rank $TR>30$ was calculated for a species that does not appear in the Olsson synonymy list. However, this species, *Globigerina bacuana*, was listed in another list as synonym of *S. triangularis* by Berggren and Norris (1997). The high rank for *G. bacuana* therefore seems plausible, indeed. The high ranks for *Globigerina triloculinoides* and *Subbotina triloculinoides*, on the other hand, appear to be overestimated in this example. *G. triloculinoides* is listed by the Olsson synonymy only once and indicates the name as only partially matching the concept. The high rank of this taxon might be explained by its frequent use in other synonymy lists. Nevertheless, a high rank indicates that the synonymy of this taxon has to be considered carefully before it can be excluded from the list of 'real' synonyms.

The examples given above show that the evaluation of the quality of TR values is not trivial and the discussion of our test results becomes 'taxonomic' almost immediately. However, three objective test parameters can be used to judge the overall quality of TR:

(1) the total number of calculated synonyms ($TR>30$) which are matching those proposed by the paleocene foraminifera working group,
(2) the number of additionally proposed synonyms which do not appear in the Olsson list, but show TaxonRank $TR<30$, and
(3) the number of taxa showing a TaxonRank $TR<30$ do appear in the Olsson list.

We performed additional test runs to further investigate the reliability of TR results, using these criteria and comparing the calculated synonyms with every synonymy published by Olsson et al. (1999). A TR could be calculated for 62 taxa (out of 66 taxa), for the remaining four species TaxonConcept contained only too short synonymy lists to calculate any meaningful ranking (Table 3).

Overall, the synonymies estimated from our test runs on the basis of TR matched the Olsson synonymy lists very well. TR calculated less synonyms than actually do occur in the Olsson list in only seven cases, resulting in an overall percentage of not detected synonyms as low as 2.6%. In contrast, the proportion of synonyms calculated by TR which do not occur in the Olsson synonymy lists is considerably higher (17%). A closer look at those species with a high numbers of synonyms outside the Olsson synonymy lists shows that a significant number of these taxon names are very similar to synonyms on the Olsson lists. Frequently, the second epithet of the calculated name is also included in one of the names of the experts' synonym lists. For example, TaxonConcept proposed the taxa *Turborotalia compressa*, *Globigerina compressa var. compressa* to be synonyms for *Globanomalina compressa*. After excluding such similar names from our mismatch calculations, the rate of additional calculated synonyms is

Table 3
TaxonRanks (TR) for 62 taxa compared to their synonymy list counterparts in Olsson et al. (1999)

Taxon name	A	B	C	D	E
Acarinina coalingensis	14	10	0	0	0
Acarinina mckannai	19	6	1	0	1
Acarinina nitida	11	4	0	0	1
Acarinina soldadoensis	8	5	0	0	0
Acarinina strabocella	6	6	0	0	0
Acarinina subsphaerica	17	13	0	0	0
Chiloguembelina crinita	2	2	0	0	0
Chiloguembelina midwayensis	5	3	0	0	0
Chiloguembelina morsei	5	4	1	1	0
Chiloguembelina trinitatensis	2	2	0	0	0
Chiloguembelina wilcoxensis	2	2	0	0	0
Eoglobigerina edita	12	11	1	1	0
Eoglobigerina eobulloides	6	5	0	0	0
Eoglobigerina spiralis	7	3	1	1	0
Globanomalina archeocompressa	2	2	0	0	0
Globanomalina australiformis	4	3	0	0	0
Globanomalina chapmani	11	8	0	0	1
Globanomalina compressa	17	11	4	3	0
Globanomalina ehrenbergi	12	8	3	1	0
Globanomalina imitata	4	3	1	0	0
Globanomalina ovalis	3	2	0	0	0
Globanomalina planocompressa	9	7	3	2	0
Globanomalina planoconica	3	2	0	0	0
Globanomalina pseudomenardii	8	6	1	1	0
Globoconusa daubjergensis	11	9	0	0	0
Guembelitria cretacea	8	3	0	0	3
Hedbergella holmdelensis	3	1	0	0	0
Hedbergella monmouthensis	7	2	0	0	0
Igorina albeari	11	8	3	2	0
Igorina pusilla	14	11	3	3	0
Igorina tadjikistanensis	10	9	2	2	0
Morozovella acuta	18	7	0	0	1
Morozovella acutispira	8	6	1	1	0
Morozovella aequa	17	4	0	0	0
Morozovella angulata	24	6	0	0	0
Morozovella apanthesma	19	7	3	0	0
Morozovella conicotruncata	14	4	0	0	2
Morozovella gracilis	9	8	2	1	0
Morozovella occlusa	15	12	4	2	0
Morozovella pasionensis	4	3	0	0	0
Morozovella praeangulata	8	5	1	0	0
Morozovella subbotinae	13	5	0	0	1
Morozovella velascoensis	10	7	0	0	0
Parasubbotina pseudobulloides	13	9	2	1	0
Parasubbotina varianta	17	13	8	2	0
Parasubbotina variospira	3	3	0	0	0
Parvularugoglobigerina alabamensis	2	2	0	0	0
Parvularugoglobigerina eugubina	24	15	0	0	0
Parvularugoglobigerina extensa	23	9	0	0	0
Praemurica inconstans	31	26	10	8	0
Praemurica pseudoinconstans	6	4	0	0	0
Praemurica taurica	5	4	1	0	0
Praemurica uncinata	10	8	2	1	0
Rectoguembelina cretacea	3	3	0	0	0
Subbotina cancellata	5	3	0	0	0
Subbotina triangularis	15	10	2	0	0
Subbotina triloculinoides	13	7	1	0	0
Subbotina trivialis	6	5	0	0	0
Subbotina velascoensis	13	7	2	2	0
Woodringina claytonensis	5	3	0	0	0
Zeauvigerina virgata	2	2	0	0	0
Zeauvigerina waiparaensis	4	3	0	0	0

A, number of taxon names proposed by TaxonRank with $TR>0$; B, number of calculated synonyms (CS) with $TR>30$; C, number of CS proposed by TaxonRank (rank>30) but not included in Olsson lists; D, number of additional CS (see column E) with taxon names similar to names of other CS included in the Olsson lists; E, number of synonyms not found by TaxonRank but included in the Olsson lists.

much lower (7.4%). To clarify whether these additional synonyms indicated by the ranking algorithm should be considered as synonyms—or not—would require a detailed taxonomic analysis, which is beyond the scope of this paper. Our results show that, in general, the top ranked taxon names calculated by TR have a good potential to be regarded as 'objective synonyms', or should at least be considered for further investigation when assembling a synonymy list.

Additional test runs can be performed by anyone interested by visiting the TaxonConcept homepage at http://taxonconcept.stratigraphy.net. A TR for a taxon will be calculated at the TaxonConcept homepage after selecting the link 'TR' on the taxon summary page.

4. Conclusions

Biostratigraphy and paleontology have been at the core of geology since its inception as a modern science. The correct application of paleontology to biostratigraphy, however, requires an up-to-date knowledge of the taxonomy of the species under consideration. To the non-taxonomist this knowledge is not available and it is difficult—even for experts—to follow the development of synonymous uses of taxonomic names in the literature over time. Our analysis of existing synonymy lists showed that it is the older literature that has the most impact on the taxonomic concepts applied today.

TaxonRank gives an overview and rating of synonymous names, by making use of the ontological network nature of synonymy lists in the taxonomic literature. These networks of non-cyclic, directed graphs with taxonomic concepts at the network nodes can be used to chart the taxonomic opinions expressed in the scientific literature, help identify valid synonymies and point to gaps in our taxonomic knowledge. The application of ranking algorithms on the ontological network encoded in the taxonomic literature also shows a way to make the tacit knowledge inside taxonomic literature available to a wider community of users. In this way, it may also help to preserve our scientific heritage in taxonomy, accumulated in centuries of careful and systematic work.

In our experiments we showed that the application of a ranking algorithm on an ontology is helpful in dealing with the inaccuracies encountered in real-world ontologies which basically represent opinions rather than 'truth'—here applied to systematic biology and paleontology. However, the concept could also be transferred to other applications, such as the identification and naming of geographical objects, or to the analysis of social networks. Wherever the opinions on concepts expressed in an ontological network have to be weighed against each other, ranking algorithms can be useful to assess the relevance of an ontological network node to a specific question.

References

Bengtson, P., 1988. Open nomenclature. Paleontology 31 (1), 223–227.

Berggren, W., Norris, R., 1997. Biostratigraphy, phylogeny and systematics of paleocene trochospiral planktic foraminifera. Micropaleontology 43 (Suppl. 1), 1–116.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30 (1–7), 107–117.

Floyd, R.W., 1962. Algorithm 97: shortest path. Communications of the ACM 5 (6), 345.

Garfield, E., 1972. Citation analysis as a tool in journal evaluation. Science 178, 471.

Gruber, T., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5 (2), 199–220.

Krell, F.-T., 2000. Impact factors aren't relevant to taxonomy. Nature 405 (6786), 507–508.

Matthews, S., 1973. Notes on open nomenclature and on synonymy lists. Paleontology 16 (4), 713–719.

Nimis, P.L., 2001. A tale from bioutopia. Nature 413 (6851), 21.

Olsson, R., Hemleben, C., Berggren, W., Huber, B., 1999. Atlas of Paleocene Planktonic Foraminifera. Contributions to Paleobiology, vol. 85. Smithsonian Institution Press, Washington, DC.

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The pagerank citation ranking: bringing order to the web. Technical Report, Stanford Digital Library Technologies Project, Palo Alto, CA.