



TaxonRank a synonymy ranking algorithm for earth science data networks

R. Huber (1) and J. Klump (2)

(1) Marum, University of Bremen (huber@stratigraphy.net), (2) Geoforschungszentrum Potsdam (jklump@gfz-potsdam.de)

Enough species have been described and investigated to make them extremely important tools in e.g. stratigraphy or paleoclimatology. Such large volumes of information however are only useful if effectively organized and made accessible, so that needed data can be quickly identified and retrieved. The classic method for doing this in paleontology is publishing primary taxonomic data in journals by carefully prepared synonymy lists. Most journals demand the use of the so-called ‘Open Nomenclature’, which allows working with taxonomic classifications that are unclear and allows the author to comment the identification of specimen of other authors.

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 should lend itself to bibliometric techniques to investigate the ranking of synonymy entries and used taxa, respectively. We present TaxonRank, a new ranking algorithm based on bibliometric analysis and internet page ranking technologies. TaxonRank uses published synonymy list data stored in TaxonConcept, a taxonomic information system which is part of the Stratigraphy.Net project. The bibliometric algorithm is additionally 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.

Synonymies are a well known a problem for earth science databases which mostly leave the original classification by the dataset’s author unchanged. Some databases are therefore difficult to query and might encounter difficulties when they try to build data networks. Technologies like TaxonRank will help those networks to create sophisticated ontologies which enable the data user to formulate adequate queries.