Geologos, 2009, 15 (3–4): 181–188
doi: 10.2478/v10118-009-0002-1

GEOLOGOS

Easily-accessible digital palaeontological databases – a new perspective for the storage of palaeontological information

Paweł Wolniewicz

Geological Institute, Adam Mickiewicz University, Maków Polnych 16, PL-61-606 Poznań, Poland;
e-mail pawelw@amu.edu.pl

Abstract

Techniques that allow to render diverse types of palaeontological data as publicly available internet resources are described. In order to develop an easily accessible digital palaeontological database, three steps should be followed: (1) digitization of the studied specimens, (2) acquisition of morphometric data, and (3) contribution of the data to open and searchable geoinformatic (palaeontological) databases. Digital data should be submitted to internet databases that allow a user to fetch various types of information from dispersed sources (semantic web services).

Keywords: digitization, morphometrics, palaeoinformatics, palaeontological collections, palaeontological databases, semantic web

Introduction

Since the beginning of palaeontology and biostratigraphy, a huge number of fossil species were identified and described. The frequent changes in taxonomic nomenclature of fossil organisms are, unfortunately, often difficult to follow, since taxonomic information is being spread over numerous journals and monographs. In addition, access to many of these scientific works is limited. Data retrieval therefore requires substantial time and effort. According to Di & McDonald (1999), at least 70% of scientist's time is consumed by the data discovery and preprocessing.

The availability of palaeontological data is further hampered by the limitations of the traditional printed media. The limited number of reference specimens presented through published photographs may not provide sufficient information to make objective decisions about particular taxa. Brief diagnoses of species and their descriptions commonly do not reflect true intraspecific variation. Moreover, the morphometric data gathered from individual specimens is generally not included in published works or restricted to estimates of means.

Some of the limitations mentioned above were eliminated with the foundation of electronic journals and internet databases. A new subdiscipline of palaeontology, namely palaeoinformatics, aims to improve the management and retrieval of information (MacLeod & Guralnick, 2000). However, these sources of palaeontological information are dispersed and heterogeneous. In contrast, the ideal model requires information management across multiple databases that pass information to

one another in the fly, thus providing scientists continuously with new taxonomic and biostratigraphic data (MacLeod & Guralnick, 2000). The computer-aided scientific-information management should also integrate a wide diversity of data types used by palaeontologists (synonyms, diagnoses, microphotographs, morphometric measurements, etc.) and be accessible for all researchers, regardless of their technical knowledge.

It is important that the information stored electronically be accurate, objective and up-to-date. However, palaeontological databases mostly include taxon-related data in a historical context, omitting synonyms and revisions (Ruban & Van Loon, 2008; Huber & Klump, 2009). Queries addressed to such databases have to include a detailed list of synonyms or authors' names. Another problem is caused by incorrect identifications of species. If such data are entered uncritically into a database, it affects consecutive studies, for example the estimates of species diversity (Stearn, 1999).

The objective of the present contribution is to describe a set of techniques that allow to render diverse types of palaeontological data as easily accessible, permanent publicly available internet resources, avoiding any inconsistency of data. Thanks to this approach, the information concerning fossil species provided by a taxonomist becomes accessible to all interested specialists, facilitating their further studies and revisions, and fostering collaboration within the scientific community.

Methods

Inaccessible palaeontological collections are useless (MacLeod & Guralnick, 2000). Unfortunately, computerization of palaeontological material, which consists of type specimens, polished slabs, acetate peels and thin sections, is a complex task. In order to convert collections into digitized datasets and to contribute them into geoinformatic databases, three fundamental steps should be followed (Fig. 1):

1. mass digitization of the studied specimens;
2. acquisition of reliable morphometric data from the fossil organisms, that allow other palaeontologists to pursue further qualitative and quantitative analyses,
3. contribution of data to open and searchable internet databases that allow the user to fetch various types of information from dispersed sources (semantic web services).

In the present contribution, these fundamental steps are demonstrated using the author's collection of Famennian (Late Devonian) stromatoporoids from southern Poland (Wolniewicz, 2009). It must be noted here that similar techniques may also be applied to other groups of fossils.

Digitization of specimens and thin sections

For the purpose of archiving and for automated quantification of the properties of rocks and fossils, digital photomicrographs are crucial. A short summary of several possible uses of digital images was provided by Choh & Milliken (2004). Lamoureux & Bollmann (2004) reviewed several of the main image-acquisition methods used chiefly for sedimentary samples. Similar techniques can be employed in palaeontological studies. Methods such as photography, scanning and scanning-electron-microscopy allow to obtain 2-D digital images of microfossils, polished slabs, thin sections and acetate peels. Palaeontological specimens that have complex 3-D shapes can be digitized using 3-D computed-tomography (CT) techniques (Molineux et al., 2007), point digitizers (Wilhite, 2003), high dynamic range imaging (Theodor & Furr, 2009) and 3-D laser scanners (Lyons et al., 2000; Smith & Strait, 2008).

In the case of Famennian stromatoporoids from southern Poland, 2-D digital images of thin sections were acquired using a film scanner. Modern 35-mm film scanners allow to capture high-resolution images of entire thin sections or acetate peels (De Keyser, 1999). Quantitative studies of thin sections require image resolutions with a pixel size of less than 5 μm (Lamoureux & Bollmann, 2004). An optical resolution of 4800 dpi, used in the present

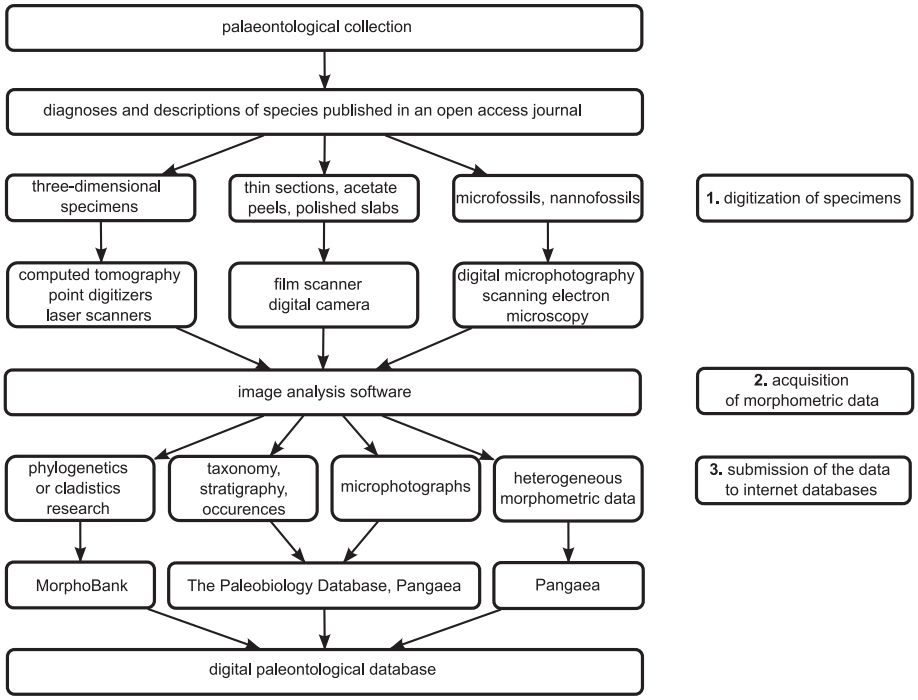

Fig. 1. Theoretical flowchart for the digitization of a palaeontological collection.

study, is therefore sufficient. Digital photomicrographs of thin sections and acetate peels can also be made by a digital camera attached to a microscope.

Acquisition of morphometric data

Contemporary palaeontological collections are usually taxon-based, which means that they contain information linked to the names of taxa. However, specimen-based databases (Berendsohn, 1995) may be more appropriate, since studies that are not linked directly to taxa should be less error-prone, allowing to avoid incorrect identifications of species made by previous investigators or changes in the taxonomic nomenclature. Specimen-based palaeontological databases should ideally contain not only digitized images of individual specimens but also descriptive detailed morphometric data obtained from these images. The descriptive information would allow to investigate a given taxonomic group with a set of diagnostic characters.

Morphometric data is collected from digitized images of palaeontological specimens and thin sections using image-analysis software. An overview of image-measurement procedures was provided by Pirard (2004). Individual researchers should follow the same measurement procedures, in order to obtain comparable results. This can be achieved using software which automatically collects the data from digital images. The public domain program ImageJ (available online at http://rsbweb.nih.gov/ij/) developed at the U.S. National Institutes of Health, and its precedessor, NIH Image, may be used. These software packages were already employed in the studies of petrographic thin sections (White et al., 1998). An investigation of particular fossil groups may require other specialized software applications.

Contribution of data to searchable internet databases

Morphometric data gathered from studied specimens should be made available, as well as

be made easily accessible for the broad scientific community. It is not sufficient to upload the entire data set to the website of a university or scientific journal, because such web resources will remain largely undiscovered. Rather than being dispersed and heterogeneous, the palaeontological resources need to be fully integrated into the global systematics reference system (MacLeod & Guralnick, 2000). Several large-scale databases of such a kind already exist (Fig. 2).

Instant availability of data published in printed journals is a second important issue. Numerous works are not yet available in a digital format, being rare and/or inaccessible. It is therefore essential to publish the results of the research in journals that are available in a digital format, preferably free of charge. Some of the most accessible palaeontological journals are featured in an informal survey 'Open Access Paleontology Journals' (http://openpaleo.blogspot.com/2009/04/open-access-paleontology-journals.html). However, this approach does not solve the problem of limited availability of works published decades ago.

The morphometric data and information concerning published works need to be integrated into one system. This was made possible thanks to the foundation of semantic web services, which introduced the concept of ontology (Lutz, 2007). Ontology implies the sharing of knowledge among different data sources (Chandrasekaran et al., 1999), which allows to access multiple, and to search dispersed sources of knowledge.

Methods for ontology-based integration of geoscience and palaeontological data sources were developed by the Geosciences Network (GEON), a project aiming at facilitating interoperability between geoscientific databases. It integrates the PaleoIntegration Project (PIP), which provides access to five global-scale fossil and sedimentary-rock databases (http://portal.geongrid.org/gridsphere/gridsphere?cid=geonpaleo). Palaeomapping tools and web services are also available, allowing not only for fast data retrieval, but also for plotting the locality palaeocoordinates on the palaeogeographic maps. The PaleoIntegration Project includes The Paleobiology Database (http://paleodb.org/), which provides occurrence and taxonomic data as well as statistical tools. All above mentioned data resources are integrated within the GEON project, thus representing an important step towards semantic interoperability between geoscientific databases.

Sources of palaeontological data integrated in the GEON project are taxonomy- and nomenclature-oriented. Incorrect identifications of species entered into the Paleobiology Database may therefore affect further studies. Careful preparation of detailed lists of synonyms could be an appropriate solution to a problem. However, the presence of many published synonymy lists and taxonomy concepts for the same groups of organisms makes the data difficult to map to a relational database. A rank based on relations between synonymy lists could be used in such situations (Huber & Klump, 2009) since impact factors are not applicable (Krell, 2000), presumably as a consequence of the lower citation rate of taxonomic articles in comparison to other studies (Valdecasas et al., 2000).

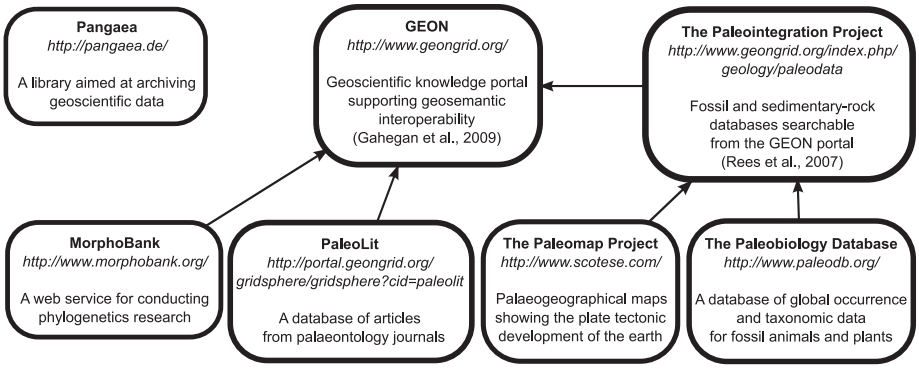

Fig. 2. Some large-scale palaeontological databases. The arrows indicate how the various databases interact to facilitate semantic interoperability between geoscientific internet resources.

Another solution is the usage of specimen-based databases, the contents of which could be searched not only by the names of taxa, but also referring to their diagnoses and to the descriptions of individual specimens (Berendsohn, 1995). However, the development of a universal data-exchange format for quantitative morphometric data from different groups of fossil organisms is difficult, since each taxonomic group is described using other sets of characteristics. Due to the heterogeneity of palaeontological data, existing geoscientific ontologies and markup languages (GeoSciML; http://www.geosciml.org/) are therefore taxon-based.

A compromise solution between linking information entirely to specimen data or taxa can be sought. For example, the concept of 'potential taxa' was proposed for the use in botanical databases (Berendsohn, 1995). However, this solution requires the development of a dedicated web-based and database system. The use of existing relevant ontologies is therefore strongly encouraged whenever possible.

A case study

Methods allowing to create easily accessible and open palaeontological databases were evaluated using a test set comprised of 75 specimens of Famennian stromatoporoids from the Cracow (Kraków) Upland, southern Poland, collected by the author (Fig. 3). The studied specimens were assigned to the genera *Gerronostroma* Yavorsky, 1931 and *Stylostroma* Gorsky, 1938. Two new species, *Stylostroma multiformis* and *Gerronostroma raclaviense*, were established. The detailed diagnoses and descriptions of the studied species were published by Wolniewicz (2009) in an open-access journal which follows the guidelines of the Budapest Open Access Initiative (http://www.doaj.org/). All contents of the journal are available online in full text, free of charge, thus being accessible for all researchers.

The studied collection consists of 160 thin sections from 75 stromatoporoid specimens, stored in the Institute of Geology, Adam Mickiewicz University, Poznań, Poland. The thin

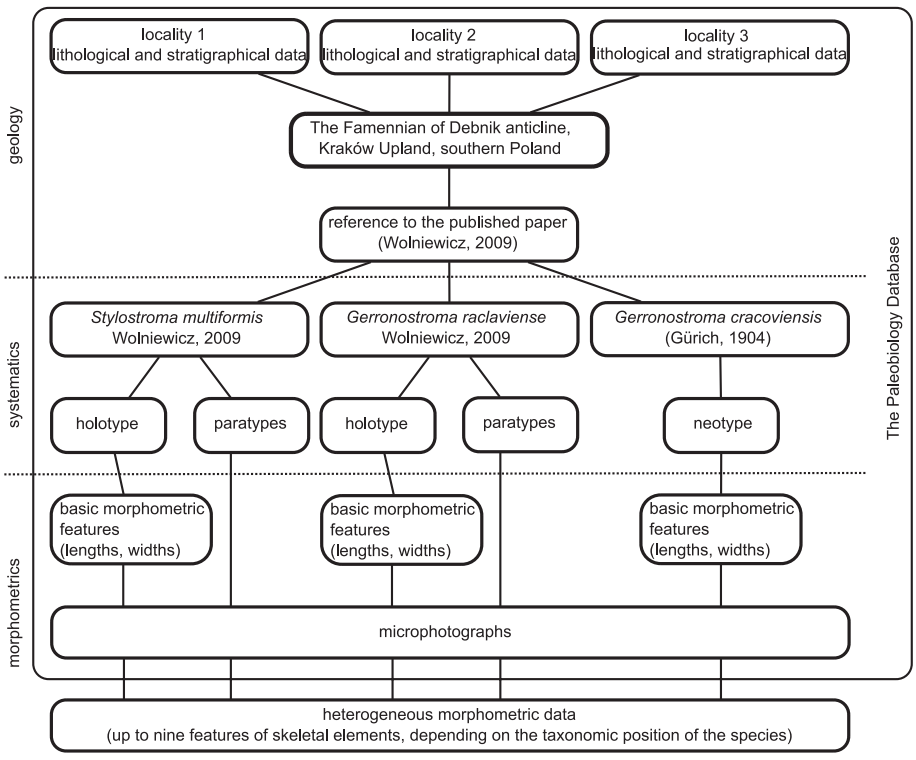

Fig. 3. The types of resources contributed to searchable internet databases in the case of Famennian stromatoporoids from southern Poland. Morphometric data submitted to the Paleobiology Database are searchable from the GEON portal. All resources, including detailed morphometric data, are available on the Pangaea website.

sections were digitized using an Epson Perfection V200 Photo flatbed scanner with a built-in transparency unit for slides. Images were acquired at a resolution of 4800 dpi.

Digital microphotographs of the studied stromatoporoids were used to obtain detailed morphometric data. Strommetric 1.0, a software package developed by the author for the analysis of the internal structure of stromatoporoid skeletons, was employed in order to perform morphometric measurements. In total, 36,159 measurements of 15 features were obtained. The data were saved in common CSV (Comma Separated Values) format, which is supported by most spreadsheets and database-management systems. All image-analysis procedures are performed by the software, thus allowing to obtain objective and reproducible data.

Microphotographs and morphometric measurements were subsequently submitted to palaeontological databases. The data are now available in two widely used web services for sharing palaeontology collections data (Table 1). The Paleobiology Database, integrated into the GEON portal, is focused on taxonomy and phylogeny, whereas Pangaea includes a wider range of information, including geoscientific and environmental data. Projects available via GEON represent an important step towards easily accessible semantic web services. However, submission of highly heterogeneous data, such as measurements of many morphometric features that apply to small groups of taxa only, is difficult. The Paleobiology Database allows only to enter the values of selected parameters (length, width, height) of the body parts specified by the user. Batch uploads of pre-existing data files are possible but not recommended. Thus, large and heterogeneous data sets containing measurements obtained from the studied specimens were submitted to the Pangaea information system. This web service does not support, however, detailed inquiries into the taxonomy, synonyms and taxonomic occurrences, which are being processed by the Paleobiology Database. Furthermore, the Pangaea system is not provided with palaeomapping tools.

Conclusions

Easy digitization of palaeontological collections is now possible due to the availability of digital cameras, scanners and advanced techniques for 3-D imaging. Digital images allow to acquire valuable and precise morphometric data. These resources should be made available to other researchers. Works with taxonomic descriptions and key illustrations published in open science journals are preferred, whereas supplementary information (including microphotographs and morphometric measurements) should be submitted to interoperable semantic web services.

To avoid possible inconsistencies within existing palaeontological web resources, caused by incorrect identifications of species, specimen-based databases could be used. Their efficiency is, however, limited due to the heterogeneity of palaeontological data. Nonetheless, researchers should make available not only the names of the taxa, but also morphometric data, images and other supplemental data, contributing these resources to the most widely used web services for sharing palaeontology-collections data. This would facilitate further studies and revisions and would allow to detect incor-

Table 1. A web-accessible, digitized collection of Famennian stromatoporoids from southern Poland.

Types of data	The Paleobiology Database	Pangaea
homepage (all data are accessible through these links)	http://paleodb.org/cgi-bin/bridge.pl?action=displayReference&reference_no=30167	http://doi.pangaea.de/10.1594/PANGAEA.724454
localities, stratigraphy	collections: 70068, 77910, 90033, 90034	dataset 724454
stromatoporoid taxonomy	taxon numbers: 148565, 148566, 148567	no data
microphotographs		dataset 724453
detailed morphometric data	no data	datasets 723765, 724366, 724367, 724368, 724369, 724370, 724371, 724372

rect identifications of species. When submitting the data to a purely taxon-based database (e.g. to the Paleobiology Database), carefully prepared lists of synonyms should be provided.

Acknowledgements

I would like to thank Hannes Grobe (Alfred Wegener Institute, Bremerhaven, Germany) and Wolfgang Kiessling (Museum für Naturkunde, Berlin, Germany), who assisted me during the entire task of data entry to the Paleobiology Database and the Pangaea library. I am also grateful for the helpful reviews and suggestions for improvement provided by Piotr Łuczyński (Institute of Geology, University of Warsaw, Poland) and Dmitry A. Ruban (Geology & Geography Faculty, Rostov State University, Russia).

References

Alroy, J., Aberhan, M., Bottjer, D.J., Foote, M., Fürsich, F.T., Harries, P. J., Hendy, A.J., Holland, S.M., Ivany, L.C. , Kiessling, W., Kosnik, M.A., Marshall, C.R., McGowan, A.J., Miller, A.I., Olszewski, T.D., Patzkowsky, M.E., Peters, S.E., Villier, L., Wagner, P.J., Bonuso, N., Borkow, P.S., Brenneis, B., Clapham, M.E., Fall, L.M., Ferguson, C.A., Hanson, V.L., Krug, A.Z., Layou, K.M., Leckey, E.H., Nürnberg, S., Powers, C.M., Sessa, J.A., Simpson, C., Tomasovych, A. & Visaggi, C.C., 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science* 321, 97–100.

Berendsohn, W.G., 1995. The concept of "potential taxa" in databases. *Taxon* 44, 207–212.

Chandrasekaran, B., Johnson, T. & Benjamins, V., 1999. Ontologies: what are they? Why do we need them? *IEEE Intelligent Systems and their Applications* 14, 20–26.

Choh, S.-J. & Milliken, K.L., 2004. Virtual carbonate thin section using PDF: new method for interactive visualization and archiving. *Carbonates and Evaporites* 19, 87–92.

De Keyser, T.L., 1999. Digital scanning of thin sections and peels. *Journal of Sedimentary Research* 69, 962–964.

Di, L. & McDonald, K., 1999. Next generation data and information systems for earth sciences research. *Proceedings of the First International Symposium on Digital Earth*. Science Press, Beijing, China, 92–101.

Gahegan, M., Luo, J., Weaver, S.D., Pike, W. & Banchuen, T., 2009. Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Computers & Geosciences* 35, 836–854.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 199–220.

Huber, R. & Klump, J., 2009. Charting taxonomic knowledge through ontologies and ranking algorithms. *Computers & Geosciences* 35, 862–868.

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

Lamoureux, F. & Bollmann, J., 2004. Image acquisition. [In:] P. Francus (ed.): *Image analysis, sediments and paleoenvironments*. Springer Science+Business Media, Dordrecht, 11–34.

Lutz, M., 2007. Ontology-based descriptions for semantic discovery and composition of geoprocessing services. *Geoinformatica* 11, 1–36.

Lyons, P.D., Rioux, M. & Patterson, T., 2000. Application of a three-dimensional color laser scanner to paleontology: an interactive model of a juvenile *Tylosaurus* sp. basisphenoid-basioccipital. *Palaeontologia Electronica* 3 (2), 16 pp.

MacLeod, N. & Guralnick, R., 2000. Paleoinformatics. [In:] R.H. Lane, F.F. Steininger, R.L. Kaesler, W. Zeigler & J. Lipps (eds): *Fossils and the future: Paleontology in the 21st century*. Senckenberg Museum, Frankfurt, 31–36.

Molineux, A., Scott, R.W., Ketcham, R.A. & Maisano, J.A., 2007. Rudist taxonomy using X-ray computed tomography. *Palaeontologia Electronica* 10, 6 pp.

Pirard, E., 2004. Image measurements. [In:] P. Francus (ed.): *Image analysis, sediments and paleoenvironments*. Springer Science+Business Media, Dordrecht, 59–86.

Rees, P.M., Alroy, J., Scotese, C., Memon, A., Rowley, D.B., Parrish, J.T., Weishampel, D.B, Platon, E., O'Leary, M.A. & Chandler, M.A., 2007. Phanerozoic earth and life: the Paleointegration Project. *Abstracts, GSA Geoinformatics Division, San Diego (May 2007)*, Paper No. 5–9.

Reitsma, F., Laxton, J., Ballard, S., Kuhn, W. & Abdelmoty, A., 2009. Semantics, ontologies and eScience for the geosciences. *Computers & Geosciences* 35, 706–709.

Ruban, D.A. & Van Loon, A.J., 2008. Possible pitfalls in the procedure for paleobiodiversity-dynamics analysis. *Geologos* 14, 37–50.

Smith, N.E. & Strait, S.G., 2008. PaleoView3D: from specimen to online digital model. *Palaeontologia Electronica* 11, 17 pp.

Stearn, C.W., 1999. Easy access to doubtful taxonomic decisions. *Palaeontologia Electronica* 2, 4 pp.

Theodor, J.M. & Furr, R.S., 2009. High dynamic range imaging as applied to paleontological specimen photography. *Palaeontologia Electronica* 12, 30 pp.

Valdecasas, A. G., Castroviejo, S. & Marcus, L. F., 2000. Reliance on the citation index undermines the study of biodiversity. *Nature* 403, 698.

White, J.V., Kirkland, B.L. & Gournay, J.P., 1998. Quantitative porosity determination of thin sections using digitized images. *Journal of Sedimentary Research* 68, 220–222.

Wilhite, R., 2003. Digitizing large fossil skeletal elements for three-dimensional applications. *Paleontologica Electronica* 5, 10 pp.

Wolniewicz, P., 2009. Late Famennian stromatoporoids from Dębnik Anticline, southern Poland. *Acta Palaeontologica Polonica* 54, 337–350.

Manuscript received 16 July 2009; revision accepted 26 August 2009.

Appendix: A compact glossary of technical terms related to databases and used in this paper

Geosciences Network (GEON; http://www.geongrid.org/): a project started in 2002 and funded by the National Science Foundation in the U.S.A. GEON aims at facilitating interoperability between geoscientific databases. For this purpose, a cooperation network has been established with other projects in archaeology, earth sciences and palaeontology. GEON includes a collection of over 5000 datasets (Gahegan et al., 2009).

Markup languages: coding systems used for annotating and structuring the text. Markup languages are widely used in the computer sciences, with HyperText Markup Language (HTML) being the core markup language of the World Wide Web.

Ontology: a formal representation of a vocabulary for a shared domain of discourse (Gruber, 1993). In computer sciences, ontology is a model to describe an object using sets of types and properties.

Paleobiology Database (http://paleodb.org/): a database containing taxonomic and distributional information about animals and plants of any geological age. The project also integrates web-based software for statistical analysis of the data. The Paleobiology Database includes over 40,000 collections and nearly 300,000 fossil occurrences (Alroy et al., 2008).

Pangaea (http://pangaea.de/): a library aimed at archiving and publishing data from earth-system research. A web-based information system stores a wide range of geoscientific and palaeontological data, including morphometric measurements, occurrences and microphotographs of fossil specimens. Pangaea is hosted by the Alfred Wegener Institute for Polar and Marine Research (Bremerhaven, Germany) and the Center for Marine Environmental Sciences (University of Bremen, Germany).

Semantic interoperability: semantic integration across heterogeneous resources (Reitsma et al., 2009). Interoperable databases pass information to one another, thus allowing researchers to gain the knowledge from dispersed data sources.

Semantic web services: web services that use markup languages in order to translate data into machine-readable form.