Simple and practical techniques to manage small databases, illustrated by a case study: bibliographic data from the "*Fossil Cnidaria & Porifera*" newsletter (1972-2010)

Katarzyna Zalecka [1]

Tomasz Wrzołek [2]

Bruno Granier [3]

Abstract: Small databases, *i.e.*, with less than 15,000 entries, are sometimes handled using inappropriate, complex, and often expensive data management systems. We present and briefly discuss a few types of proprietary and open-source, relational and non-relational, server-based versus portable databases and specific tools to handle the latter. With a collection of nearly 7,000 bibliographic notes during its 40-year history "*Fossil Cnidaria & Porifera (FC&P)*", the newsletter of the "*International Association for Study of Fossil Cnidaria and Porifera*", was chosen as a case study. The analysis of the temporal trends in the FC&P bibliographic database shows a decrease over the years in the number of publications effectively reported in FC&P. Almost all relevant papers for the decade 1981-1990 are reported, but this good coverage ratio falls down to less than 50% after 2000; accordingly, the concern about the data representativeness is addressed in our interpretation. Besides the classical database management systems and spreadsheet software, which were originally used with the FC&P case study, we present two discrete, open-source, flat and portable options where data can be displayed using any widely available Internet browser, and that are suitable to handle most small databases (XML or JS files) as documented herein.

Key Words: XML; JavaScript; bibliographic database; corals; sponges; reefs; fossils.

Citation: Zalecka K., Wrzołek T. & Granier B. (2015).- Simple and practical techniques to manage small databases, illustrated by a case study: bibliographic data from the "*Fossil Cnidaria & Porifera*" newsletter (1972-2010).- *Carnets Géol.*, Madrid, vol. 15, n° 2, p. 13-19.

Résumé : *Techniques simples et efficaces pour gérer de petites banques de données, illustrées par une étude de cas : données bibliographiques issues du bulletin d'information "Fossil Cnidaria & Porifera" de 1972 à 2010.-* La gestion de banques de données de taille modeste, c'est-à-dire comportant moins de 15000 entrées, est parfois réalisées par le biais de systèmes complexes, parfois inappropriés, et souvent onéreux. Nous présentons et abordons brièvement quelques exemples de banques de données, relationnelles ou non, et leurs systèmes d'exploitation, à code protégé ou ouvert, interrogeant sur un serveur distant ou embarqués sur l'ordinateur de l'utilisateur. Avec quelques 7000 notices bibliographiques collectées pendant une quarantaine d'années d'existence, "*Fossil Cnidaria & Porifera (FC&P)*", le bulletin d'information de l' "*International Association for Study of Fossil Cnidaria and Porifera*", a été sélectionné comme exemple d'application. L'analyse des tendances temporelles dans la banques de données bibliographiques FC&P montre une diminution au cours du temps du nombre de publications effectivement enregistrées dans FC&P. Presque toutes les références pertinentes de la décennie 1981-1990 y figurent, mais le bon taux de couverture tombe à moins de 50% après l'an 2000 ; par conséquent, la question de la représentativité est traitée dans notre interprétation. À côté des logiciels classiques de gestion de données et de type tableur, qui ont été utilisés dès le départ dans notre étude de cas, nous présentons deux options embarquées distinctes, plates et à code ouvert, dont les données peuvent être lues en utilisant n'importe quel navigateur Internet, et qui sont tout à fait adaptées pour gérer la plupart des petites bases de données (documents de type XML ou JS) telles que celle présentée dans cette note brève.

Mots-clefs : XML ; JavaScript ; base de données bibliographiques ; coraux ; éponges ; récifs ; fossiles.

1. Introduction

Although many organizations, in academia and industry, use MySQL, a popular Open Source database management system, there are needs in our communities (geologists, paleonto-logists, biologists, and others) for simple and practical tools to build and handle small databases (*e.g.*, Benton, 1993, 1995, with "*The Fossil Record 2*", which is merely a spreadsheet; Löser, 2004, with PaleoTax, a software he developed; Steemans & Breuer, 2007, with

[1]   ul. Kokoszki 10/19, 44-100 Gliwice (Poland)
mkat.zalecka@gmail.com

[2]   Silesian University, Department of Earth Sciences, ul. Będzińska 60, 41-200 Sosnowiec (Poland)
tomasz.wrzolek@us.edu.pl

[3]   Department of Ecology and Evolutionary Biology, The University of Kansas, 1200 Sunnyside Avenue, Lawrence, Kansas 66045 (USA)
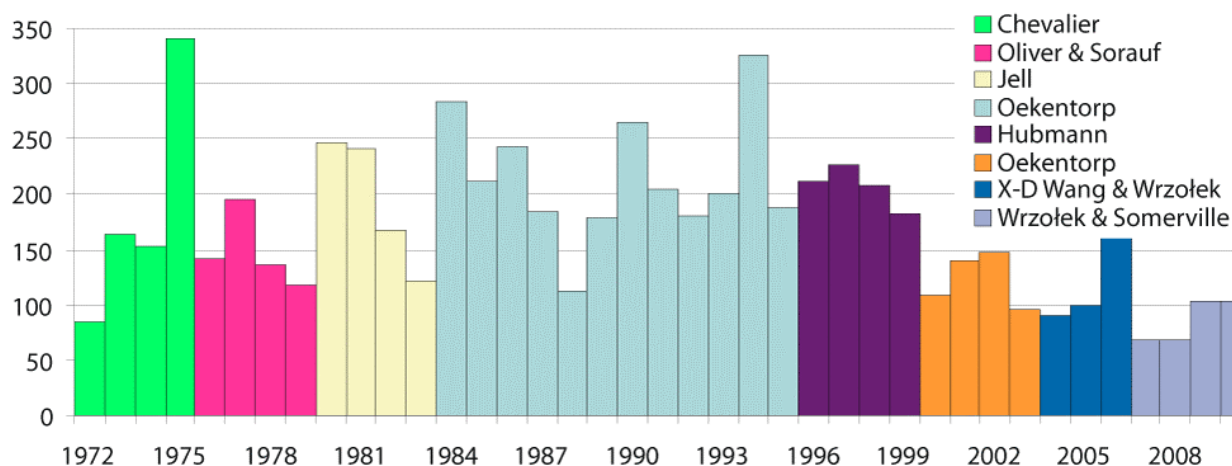bgranier@ku.edu

Figure 1: Numbers of papers on fossil corals, sponges and reefs (CSR) as reported by *Fossil Cnidaria & Porifera* per year (1972-2010), with indicated editors-in-chief of the newsletter.

PalyWeb, which is a MediaWiki project; or KERNER *et al.*, 2011, with "Archaeocyatha - a knowledge base", which uses Xper[2]). We are considering alternative flat or relational, portable or non-portable tools to launch simple queries. For example, the "*International Association for Study of Fossil Cnidaria and Porifera*" publishes a newsletter, namely "*Fossil Cnidaria & Porifera (FC&P)*", and over the years its editors have compiled summary data of nearly 7,000 bibliographic references dealing with fossil corals and sponges and the related topics. Ultimately, it represents a "small" database that is used hereafter as our practical case study.

2. The case study: the FC&P database

2.1. Background

"*Fossil Cnidaria & Porifera (FC&P)*", the newsletter of the "*International Association for Study of Fossil Cnidaria and Porifera*", has been established as a means of dissemination of information among specialists studying fossil sponges, corals and reefs. Over the years, 37 volumes have been published.

Besides reporting on current research and publishing original short papers, each issue of FC&P includes an update of the currently published literature relevant to the topics followed by FC&P. We gathered these bibliographic notes primarily as a sort of memorial to the past 40 years of activity of several editors (Fig. 1) and many correspondents. A secondary goal was to create a tool for future editors to avoid multiple presentations of the same material. Last but not least, we wanted to build an interactive catalogue of literature for paleontologists studying fossil corals and sponges. The result of our efforts was initially presented as a poster during the 11th Symposium on Fossil Cnidaria and Sponges in Liège, Belgium (ZALECKA & WRZOŁEK, 2011).

Here, we present the contents of the FC&P database, discuss temporal trends observed, and finally analyze its representativeness and quality, estimated on the basis of the ratio of coverage of all publications concerning fossil corals and sponges.

2.2. Nearly 7,000 bibliographic notes

The bibliographic notes of the 36 issues of FC&P have been gathered into the database. Only current bibliographic entries for the period 1970-2010, were considered. The contents of the last issue, *i.e.*, the 37th volume of FC&P, which covers only the beginning of the last decade, were added to the latest versions of our database, but was not included in our decennial analyses.

Entries in our database concern only publications. Consequently, we avoided abstracts or posters as not being publications *sensu stricto*, at least not in the sense of the International Code of Zoological Nomenclature (Article 9.9, 1999 edition amended). On the other hand, we deliberately listed unpublished theses which contain a significant body of data. In our opinion, they deserve a mention to permit, at best, upgrading by their original author(s) to the status of effective publications or, at least, being quoted by other authors in their own publications sometime in the future.

The recorded features include: author(s), title, year of publication, taxonomic group, stratigraphic interval, geographic area, publication source (journal, book), abstract, and eventually DOI number for some recent entries.
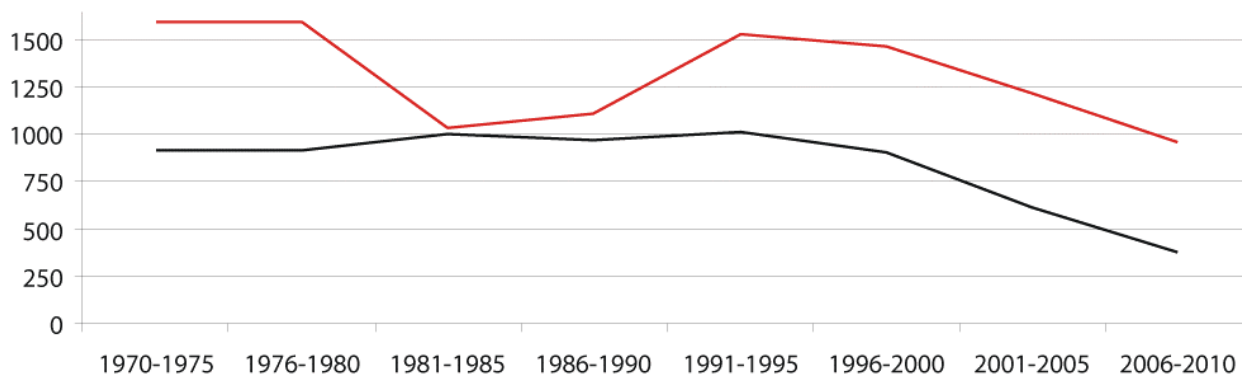
Figure 2: Numbers of papers on fossil corals, sponges and reefs (CSR) - supposed temporal trends. Lower (black) line for CSR papers reported by *Fossil Cnidaria & Porifera* newsletter; upper line (red) for hypothetical, corrected numbers of CSR papers (taken from Table 1, row 7).
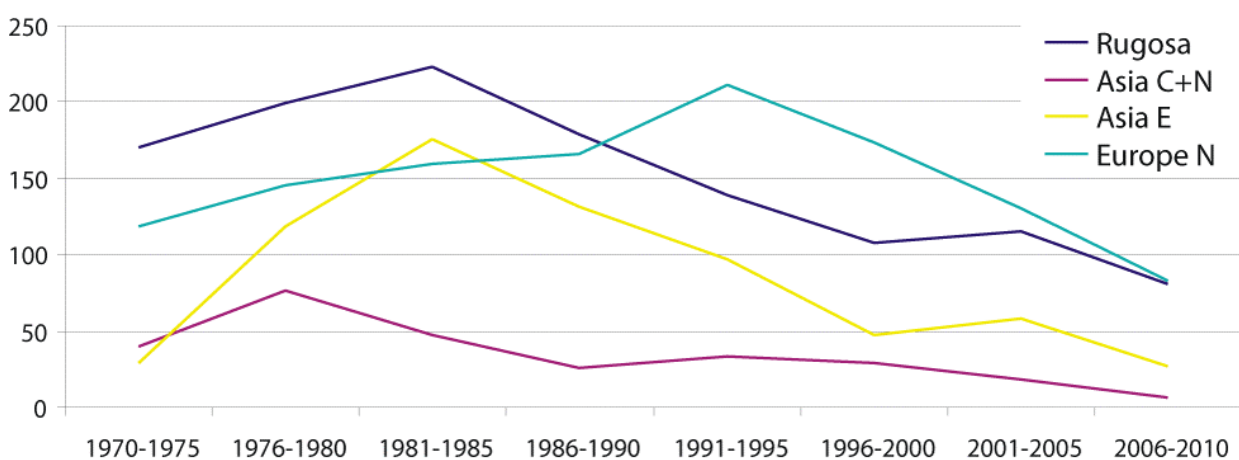

Figure 3: Temporal trends in numbers of papers reported by FC&P, dealing with selected research areas; particular trends are overall similar to each other and to the general trend, as presented in Figure 2.

2.3. Thematic contents

Taxonomic subjects of the reported papers, not surprisingly, concern sponges (20%) and cnidarians (80%). Material published on sponges mostly deals with stromatoporoids (62%), archaeocyathans (22%), chaetetids (12%) and sphinctozoans (4%). Cnidarian publications almost wholly deal with corals (93%), among which 47% describe Rugosa, 27% Scleractinia, and 21% Tabulata.

The main general subjects of publications in the FC&P database are reefs (45% - also papers on carbonate sedimentology and diagenesis), taxonomy (41%) and other geological subjects (14%).

With respect to stratigraphy, about 60% of papers deal with the Paleozoic, and 19% for both the Cenozoic (which per definition includes the Quaternary) and the Mesozoic. Most articles concern material from the Devonian System (18%), the Neogene (16%, including Recent biota), and the Carboniferous (12%).

Regarding geography, 87% of the publications refer to continental areas and the rest to oceans and seas. Localities listed are mostly in Europe (41%), then Asia (33%), and finally the Americas (19%). As for the marine areas, the most cited are in the Pacific Ocean (32%), followed by the Indian Ocean (30%) and the Caribbean portion of the Atlantic Ocean (17%).

2.4. Temporal trends

Besides the topics of publications, the FC&P database allows us to analyze temporal trends, which can be adjusted for the coverage ratio of the publications concerning fossil corals and sponges, *i.e.*, actually the number of publications reported by FC&P versus the estimated number of all publications. In addition, it offers us an opportunity to test Sando's (1993, 1997; see also Turnsek & Löser, 1993, and Wrzołek, 2007) predictions of a fall in numbers of publications on fossil corals and sponges due to decline in funding, manpower, *etc*.

The number of bibliographic notes published in FC&P varies yearly in a rather erratic way (Fig. 1). To smooth this, we used five-year intervals in our "final" analysis (Table 1, row 1; Figs. 2 - 3). The resulting trend is represented in Figure 2 by the lower line that clearly documents a drastic decline.

	1970-1975	1976-1980	1981-1985	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010
1	914	912	999	973	1011	910	610	373
2	159	179	171	255	239	198	221	195
3	91	103	166	225	157	123	111	77
4	0.57	0.57	0.97	0.88	0.66	0.62	0.50	0.39
5	2016	2112	2328	2653	2797	2922	3522	4286
6	0.08	0.08	0.07	0.10	0.09	0.07	0.06	0.05
7	1600	1600	1030	1110	1530	1470	1220	960
8	690	690	30	130	520	560	610	590

Table 1: Numbers of papers on fossil corals, sponges and reefs (CSR) per 5-years intervals (1970-2010):
1 CSR papers reported in *Fossil Cnidaria & Porifera (FC&P)* newsletter
2 CSR papers in 9 paleontological journals (listed in "Temporal trends" in text)
3 FC&P reports of CSR papers in 9 journals
4 calculated coverage rate of CSR papers by FC&P (**3** over **2**)
5 total number of papers in 9 journals
6 ratio of CSR papers in 9 journals (**2** over **5**)
7 estimated full numbers of CSR papers (**1** over **4**, rounded to nearest 10)
8 estimated numbers of CSR papers missing from FC&P reports (**7** minus **1**, rounded to nearest 10)

In search for factors responsible for lower numbers of papers, we analyzed temporal trends in selected research areas (Fig. 3). Overall, the trends are similar, with only minor differences between the groups analyzed: we did not observe any specific research area that was neglected and responsible for the general negative trend.

On the other hand, however, we noticed that the raw data of our database (Table 1, row 1) indicate only reported publications; the numbers are obviously partial, and what is missing is a more or less significant number of unreported publications. In an attempt to obtain corrected numbers, we checked the contents of nine paleontological journals and compared results with those listed in the database of FC&P. The journals taken into consideration for this test were *Acta Palaeontologica Polonica*, *Acta Palaeontologica Sinica* (after a break, new volumes since 1976), *Facies* (established in 1979), *Journal of Paleontology*, *Lethaia*, "*Palæogeography, Palæoclimatology, Palæoecology*", *Palaeontology*, *Palaios* (established in 1986), and *Paleontologicheskiy Zhurnal*. The number of papers on fossil corals, sponges and reefs in nine journals are listed in Table 1, row 2. Estimated coverage rates (Table 1, row 4) were calculated by comparison of numbers of papers on fossil corals, sponges and reefs in the nine journals and those reported for the same journals by FC&P (Table 1, rows 2 and 3 respectively). These data were used subsequently to obtain corrected values of publications (Table 1, row 7).

In the nine journals analyzed, papers on fossil corals, sponges and reefs represented about 1 in 10 papers in the period 1970-1995. Thus, ratio dropped to about 1 in 20 in the years 1996-2010 (Table 1, row 6). These values possibly reflect the general trend of decline in taxonomic studies in favor of more general topics, and possibly indicate an important general trend in paleontological research and publications.

The corrected numbers of publications (Fig. 2, upper, red line), display similar trends to the uncorrected data, confirming a marked decline in the numbers of published papers in the decade 2001-2010.

As for the trends in FC&P coverage rates, our analysis suggests that during the decade 1981-1990 most published papers were reported and listed by FC&P. In the subsequent decade (1990-2000) the rate of coverage fell to about 60-70% and in the decade 2001-2010 it fell to 40-50%. One possible explanation of this decrease might be the growing body of commercial databases (GeoRef, Pascal, Web of Science®, *etc.*) and the consequence that many authors assume they do not need to report their new contributions.

Taking into account coverage rates as estimated in Table 1 and total numbers of publications reported by FC&P we estimated "missing" numbers (Table 1, row 8). When summed, these estimate over 3,800 "missing" publications, which is a surprisingly high number for 7,000 reported publications! Although it is impossible to be sure how accurate this number is, at the moment, we feel fairly certain that there is a large body of bibliographic data unreported by FC&P.

3. FC&P database

3.1. The original FC&P data structure

After transcription into a MySQL® server-based database, a preliminary version was made available online in November 2011. It was checked for multiple records, possible errors and omissions. In April 2012, it was transferred to the web server of the University of Silesia, Sosnowiec (Poland), at http://kse.wnoz.us.edu.pl/sql/index.php. There were two problems with this arrangement. First,

some knowledge in PHP language is required to administer the MySQL database. Secondly, it is not a portable database, but rather a server-based database. For these reasons, two extra files with the basic information attached, *i.e.*, a Microsoft® Excel® file and an Access® file (which can also be read with LibreOffice Calc and Base, respectively), were made available for download. Both files, the XLS (Appendix 1) and the ACCDB (Appendix 2), are set to allow anyone to easily access and analyze these data in all their aspects provided the appropriate programs are already installed on the user's personal computer, either the proprietary programs of Microsoft or their free counterparts with LibreOffice (LibO).

When one requires a relational database (with multiple tables) to store a very large amount of data (thousands of entries) consisting mostly of text (not numbers), either MySQL, Access or LibOBase could be used. However, the current version of the FC&P database, with only about 7,000 records, does not nearly exceed the 15,000-row and 256-column upper limits recommended for Excel or Calc spreadsheets. Even though both spreadsheet programs are not database management systems, they can be used to handle a small flat (non-relational) database like the FC&P one and display it as a simple table. As spreadsheets are commonly used to handle integer strings (numbers), not character strings (text), there is a 255-character limitation in the number of characters authorized in the cells (if this number is exceeded, some characters can be lost when copied and pasted from one cell to another). Consequently, we have been looking for alternative portable versions without such limitations.

3.2. Alternative FC&P file formats and related tools to handle them

As a result, in addition to the original database, we present two new discrete types of portable flat databases and the tools to operate them:

- the first (FCnP.xml: Appendix 3) is written in XML (Extensible Markup Language). It is accessible by any computer using many widely available Internet browsers (Microsoft Internet Explorer®, Mozilla® Firefox®, Google® Chrome®, Apple® Safari®) and requires either a browser with a built-in XML parser to display its whole contents as an XML file, or a JS (JavaScript™) parser to be able to read the entries as individual HTML files [Note];

- the second file type (tip_data.js: Appendix 4) comprises only Java-Script™ elements, for both the database and its dedicated HTML search engine. It only requires a browser to return search results.

XML files can easily be converted, either manually or through a JS parser, into XLS files for Microsoft® Excel® file or LibreOffice Calc.

3.3. Technical specification requirements of the XML and JS databases

XML has a special set of characters that are invalid and cannot be used, in normal XML strings. Accordingly these 5 special characters should be converted:

- " into " or "

- & into & or &

- ' into ' or '

- < into < or <

- > into > or >

Most modern browsers have a built-in XML parser to read and manipulate XML. First the XML document must be loaded. Then the parser reads it and converts it into an XML DOM (Document Object Model), which can then be manipulated with JavaScript. Using two simple JS files, browse.js and display.js (Appendix 3), we are able to display an HTML-interpreted version of the XML database (Appendix 3) and browse through it (see the permalink: http://paleopolis.rediris.es/cg/1502/browse.html).

Individual entries are accessed, starting from entry number 0. When clicking on the information displayed (author, year, title) the whole entry content, *i.e.*, the full available information (including the references, the abstract, *etc*.), is made visible. Discrete references can be browsed with an increment or a decrement of one, five or twenty. One can also go directly to a specific entry when typing its ordering number.

Alternatively one can use the JavaScript search engine. It includes two simple JS files, tip_form.js and tip_search.js, slightly adapted from "Tipue Search" (URL: http://www.tipue.com/search/), an Open Source search engine, the code of which was released under the MIT License. The flat database, tip_data.js, is also written in JavaScript. The display page is an HTML interface (Appendix 4). Once again, this database is a portable, platform independent search engine (see the permalink: http://paleopolis.rediris.es/cg/1502/search.html).

As for the XML database presented above, in the flat JS (JavaScript) database, some characters must be replaced: " (because this symbol is used twice in each data entry to frame it) into " and & into &

To complete this chapter, another script is made available (see the permalink: http://paleopolis.rediris.es/cg/1502/tool.html) to write additional entries for both databases, either the XML or the JS.

4. Discussion

Hopefully, the FC&P database will be maintained and even expanded with the addition of the contents of the future volumes of "*Fossil Cnidaria & Porifera*". The first two authors (K.Z. & T.W.) hope that users will help in detecting errors and data gaps in the current versions.

In the illustrated case study of the FC&P database, there are nearly 7,000 entries, which qualify it as a relatively "small" database. People interested in working with small, non-relational (flat) databases with less than 15,000 entries (rows), which can be displayed in the form of a table, should consider using a portable, non-server-based data management system. Access and LibOBase are portable and allow complex queries but they are not user-friendly and require the software to be pre-installed on the user's personal computer. Excel and LibOCalc are spreadsheet softwares that also require preinstallation, and are better for handling numbers than text. On the other hand, the XML and JS database can be easily copied and run from many widely available browsers.

One of the two tools presented here, *i.e.*, the JavaScript tool, proves to be very efficient as a search engine for references through specific words or groups of words. The flat JS database and the associated search engine are probably the most comprehensive and efficient data management system to handle bibliographic data, as demonstrated herein with the FC&P database, but also small fossil collections and other small databases. It is only 6.8 MB (7.4 MB for the .XML file) versus 15.5 MB for the original Access file (.ACCDB) and 7.4 MB for the original Excel file (.XLS). On the other hand, the XML database is more versatile and scalable but it may require an adapted parser (*i.e.*, some JS programming) to address the specific queries of its end users.

In conclusion, before using a database one should always question the pros and cons of adopting a particular version and its attached tools, *i.e.*, the components of its database manager.

Bibliographic references

BENTON M.J., ed. (1993).- The fossil record 2.- Chapman & Hall, London, 845 p. URL: http://www.fossilrecord.net/fossilrecord/index.html

BENTON M.J. (1995).- Diversification and extinction in the history of life.- *Science*, Washington, vol. 268, p. 52-58.

International Commission on Zoological Nomenclature (1999).- International Code of Zoological Nomenclature (4th edition).- The International Trust for Zoological Nomenclature, The Natural History Museum, London. URL: http://www.nhm.ac.uk/hosted-sites/iczn/code/index.jsp

KERNER A., VIGNES LEBBE R. & DEBRENNE F. (2011).- Computer-aided identification of the Archaeocyatha genera now available online.- *Carnets Géol.*, Madrid, Letter 2011/02 (CG2011_L02), p. 99-102.

LÖSER H. (2004).- PaleoTax - a database program for palaeontological data.- *Computers & Geosciences*, Stanford, vol. 30, p. 513-521. URL with a link to download the software: http://www.paleotax.de/

SANDO W.J. (1993).- Late Paleozoic coral research: past, present, and future. *In*: Proceedings of the VI International Symposium on Fossil Cnidaria and Porifera (Münster, 9-14 September 1991).- *Courier Forschungsinstitut Senckenberg*, Frankfurt am Main, vol. 164, p. 21-36.

SANDO W.J. (1997).- A silver platter-history of the International Association for the Study of Fossil Cnidaria and Porifera and trends in cnidarian and poriferan research, 1971-1994. *In*: Proceedings of the VII International Symposium on Fossil Cnidaria and Porifera (Madrid, 1995).- *Boletin de la Real Sociedad Española de Historia Natural*, Madrid, (Seccion Geologica), vol. 91, issue 1-4, p. 5-33.

STEEMANS P. & BREUER P. (2007).- PalyWeb: a palynomorph database project on the web. *In*: STEEMANS P. & JAVAUX E. (eds.), Recent

advances in palynology.- *Carnets Géol.*, Madrid, Memoir 2007/01, Abstract 08 (CG2007_M01/08), p. 48-52.

TURNSEK D. & LÖSER H. (1993).- The history of Mesozoic coral research after 1940. *In*: Proceedings of the VI International Sympo-sium on Fossil Cnidaria and Porifera (Münster, 9-14 September 1991).- *Courier Forschungsinstitut Senckenberg*, Frankfurt am Main, vol. 164, p. 37-46.

WRZOŁEK T. (2007).- Perspectives of research on fossil corals and sponges. *In*: Proceedings of the 9th International Symposium on Fossil Cnidaria and Porifera (Graz 2003).- *Öster-reichische Akademie der Wissenschaften, Schriftenreihe der Erdwissenschaftlichen Ko-missionen*, Wien, vol. 17, p. 517-521.

ZALECKA K. & WRZOŁEK T. (2011).- The database of *Fossil Cnidaria & Porifera* newsletter, 1972-2010. *In*: Abstracts volume of 11th Symposium on Fossil Cnidaria and Sponges (Liege, August 19-29, 2011).- *Kölner Forum für Geologie und Paläontologie*, vol. 19, p. 196.

Appendices

1) XLS database

http://kse.wnoz.us.edu.pl/sql/base.xls
(external link, regularly updated)

http://paleopolis.rediris.es/cg/1502/base.xls

2) ACCDB database

http://kse.wnoz.us.edu.pl/sql/base.accdb
(external link, regularly updated)

http://paleopolis.rediris.es/cg/1502/base.accdb

3) XML database

The XML file for FC&P:

http://paleopolis.rediris.es/cg/1502/FCnP.xml

The printable version of the XML file for FC&P (10.0 MB, *i.e.*, 2,548 pages):

http://paleopolis.rediris.es/cg/1502/pdf/FCnP.pdf

The HTML code for the parsed page:

http://paleopolis.rediris.es/cg/1502/pdf/browse.html.pdf

The codes for JS files used to parse the XML file:

http://paleopolis.rediris.es/cg/1502/pdf/browse.js.pdf
http://paleopolis.rediris.es/cg/1502/pdf/display.js.pdf

4) JS database

The JS file for FC&P:

http://paleopolis.rediris.es/cg/1502/tip_data.js

The printable version of the JS file for FC&P (6.2 MB, *i.e.*, 1,646 pages):

http://paleopolis.rediris.es/cg/1502/pdf/tip_data.pdf

The HTML code for the search engine page:

http://paleopolis.rediris.es/cg/1502/pdf/search.html.pdf

5) Editing tool

The HTML code of the form page to write new XML and JS entries:

http://paleopolis.rediris.es/cg/1502/pdf/tool.html.pdf