Turk J Bot
28 (2004) 129-138

Computer Technology for the Future of SW Asiatic Botany

R.J. PANKHURST
Royal Botanic Garden, 20A Inverleith Row, Edinburgh EH3 5LR, United Kingdom
email: richard@rbge.org.uk

Received: 05.11.2002
Accepted: 29.01.2003

Abstract: Much computer software already exists for floristic and monographic studies. Database software can hold all the data required for a Flora, and can process it much more accurately. The fundamental data tables are those for names, authorities, people, specimens, bibliography, geography and distribution. To these can be added other data such as descriptions, keys, images, maps, chromosomes and vernacular names. The same data system can be used for collection management and making herbarium labels. It is very important to use data standards such as for books, journals and authorities. Computerised descriptions and keys can also be constructed and held in the database. Data from the Flora of Turkey will be used to demonstrate the PANDORA database and PANKEY software.

Key Words: Flora of Turkey, databases for plants, taxonomic databases, PANDORA taxonomic database, PANEKY identification software

Introduction

This paper will explain how the flora of SW Asia may be more effectively known with the help of computer technology. In fact, much the same remarks will apply to monographic studies, and to the floras of other areas. The data for the demonstration has been taken from the family *Rosaceae* from the Flora of Turkey (Davis, 1972). The author has not himself taken part in any aspect of the preparation of this Flora, and acknowledges the contributions of those present at the conference who did do so. The author's own software, the PANDORA taxonomic database (Pankhurst, 1993) and the PANKEY identification program suite (Pankhurst, 1995) have been used. The PANDORA software is also being used in the Euro+Med project and holds the standard version of the data from the Flora Europaea (Tutin et al., 1964-1993) at the Royal Botanic Garden Edinburgh. A PANDORA database of the published volumes of the Med Checklist (Greuter et al., 1984-) will also contribute to the Euro+Med database (www.euromed.org.uk). Some of the nomenclatural data relevant to Turkey was downloaded from the global Rosaceae PANDORA database which is currently in preparation at Edinburgh.

This paper will concentrate on better ways of achieving botanical projects using computers, and rather little will be said about the computing technology. However, it is worth pointing out that the PANDORA database application was created for use either by a single researcher working on a desktop, or to be shared over a local network by a team of workers. It can also be accessed and shared over the internet. Each worker or each institute can have one to many separate PANDORA databases for different purposes, and data can be transferred from one database and merged into another. PANDORA is based on the Advanced Revelation relational database management system. This is a sophisticated but little known development system, superior in important ways to more familiar database software (Pankhurst, 1993). Other taxonomic databases and other identification packages exist, but will not be discussed here. PANDORA is still unusual in the extent to which it covers a wide range of different types of data, and most especially the morphological descriptions and identification keys. This means that it can accommodate all the data published in such texts as the Flora of Turkey.

While it is perfectly sensible to use a word processor to prepare the text of a Flora, that is not at all what is meant here by computerisation. It is important to realise that much of the virtue claimed here for a computerised Flora depends on the use of a database, and that that

database should be relational. There are technical definitions of this, which need not concern us now, but there are two important points to make. The first, which makes obvious good sense, is that data items are only stored once, and then referred to repeatedly when used. For example, the data concerning a person, full name, date of birth, and so on, is only entered in one place (one record in one table). Then if a spelling correction is needed later, this only needs to be made in one place, and then the corrected name appears automatically in every other place in which it is used. Secondly, the more accurately the structure of the real data is reflected in the structure of the data, via related tables in the database, then the easier it is to realise unanticipated future uses of the data. PANDORA is unusual in that in certain tables of data, particularly in nomenclature, geography and uses, it is capable of expressing the inherant hierarchies.

Professor Heywood referred in his talk to the statement 'The discipline [of taxonomy] will have to reinvent itself if it is to survive and flourish' (Godfray, 2002). There is no question that basic taxonomic information, i.e. the body of data published over several centuries, will continue to be indispensable for biology, and that molecular and cladistic studies will only serve to supplement it. The essential redefinition must be in the organisation and presentation of information. Hence, the published Flora or monograph will in future take second place to the information system on which it is based. This is already happening with the Euro+Med project. The first Flora Europaea was a completed publication before the team at Reading under Professor Heywood computerised it. Now with Euro+Med the priority is reversed. First create the database and then use it to produce such publications as may be required. In fact this change is part of a much wider revolution, affecting many other professions, in which the use of books and libraries is being supplanted by the use of databases. This change may be as fundamental as the change that took place in medieval times, when handwritten parchment was supplanted by printed books. In an even wider sense, it may be that the computer revolution will come to be seen as an extension to human intelligence. The revolution in method that is needed should be carried out by the 'Young Turks' in our profession; that is, those (of any age) who want to make a permanent change towards a much better way of doing things.

Practical details

General

It is no exaggeration to say that all the information in a traditional Flora can be stored in a database. The backbone of a taxonomic database is the scientific nomenclature (including synonyms, and other kinds of names). This in turn is based on the people, who in turn published the names in the literature. The definitions of the taxa also depend on the herbarium and other specimens used, and the specimen label data enables us to map the geographic distribution, based on the geography. We must also have the morphological descriptions for identification. Hence biography, nomenclature, literature, specimens, geography and descriptions are the most fundamental types of data which need to be represented in database tables. To this may be added a more or less endless variety of other kinds of information, such as images, vernacular names, uses, chromosome numbers, conservation data, ecology, breeding data and so on. There are many interconnections between these tables, such that for example, nomenclature depends on biography and literature. The principle referred to above about storing information once and referring to it as often as needed is applied here. For example, people (or strictly groups or 'committees' of one or more people) are referred to as authorities for taxa, authors or editors of books or articles, and as collectors and determiners of specimens. Standards for data are used throughout, following the recommendations of the International Working Group on Taxonomic Databases (TDWG, see the website WWW.TDWG.ORG). Many of these are standards for abbreviations, such as TL2, BPH and IPNI.

Nomenclature

In order to capture the hierarchy of names, the names are stored in an unfamiliar way. Each name is represented by 3 items; the actual name string e.g. 'vulgaris', the rank e.g. 'species', and a reference to the next higher taxon e.g. the genus. There is a separate table for ranks which can be added to when unusual ranks are needed. The ranks of family, genus and species are fixed into the system and are not optional, but others can be added or deleted as needed. In a particular project, as in the Flora Europaea, the only ranks that were officially required in the text are the family, genus, species and subspecies, but about 30 other ranks entered from the synonymy. PANDORA is deliberately designed to follow only one taxonomy, as is the practice in actual floras or

monographs. This is not a shortcoming of the design, but a conscious and deliberate choice, in order to encourage the creation of a single consistent taxonomic framework for the convenience of users. Hence names can be one of 4 different types; accepted, synonymous, provisional and external. The provisional category is for names whose type has not yet been decided. The external type is for names which need to be included for various reasons but which do not occur in the region, e.g. homonyms. Names of all ranks and types are all included in the same table. If a name is a combination based on another taxon (the basionym), this can be recorded. Similarly for the parents of hybrids. The citations of the place of publication of the taxa refer to the bibliography file. See Figure 1.

People and authorities

The file for people is one of the basic files of the database and does not depend on any other. An example is shown in Figure 2. People often work in groups (which we call 'committees') and this is expressed in an intermediate file. Committees are referred to in many different places, such as authorities, collectors, names of authors of publications and so on. Authority abbreviations for plant names are kept in a separate file, together with a reference to the committee representing the people concerned. The abbreviations for authorities follow the standard provided by Brummitt & Powell (1992, also on the IPNI website).

Bibliography

These are mainly books or journals, but it is also possible to have articles in journals or chapters in books, so this file is hierarchic to a limited degree. Figure 3 shows the entry for Boissier, Flora Orientalis. Citations are created in the usual way with an abbreviation for the book or journal, and volume number, part and page numbers. The titles and abbreviations for books are taken, wherever possible, from TL2 (Stafleu & Cowan, 1976-) and for journals, from BPH (Lawrence et al., 1968).

Geography

The gazetteer file is expressed hierarchically and with geographical synonyms, in a manner similar to the taxon file. Each place has a name and a level and can belong to a higher level place. For example, Anatolia belongs to Western Asia, and both of these are represented by separate records in the gazetteer. The record for Bolu is shown in Figure 4.

Fig. 1. Nomenclature record for *Amygdalus communis* L.

Figure 2. A record from the PERSONS file for Peter Davis.

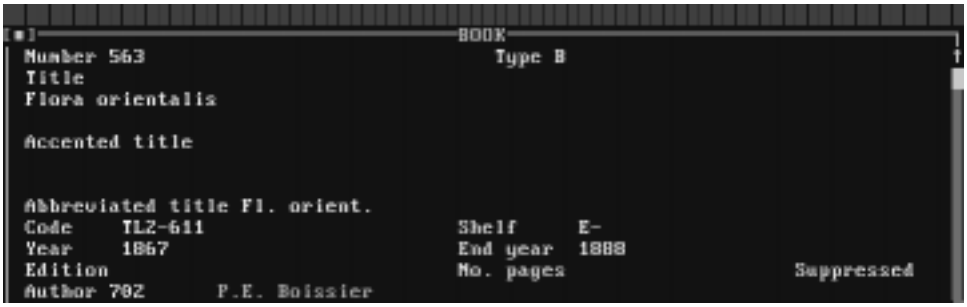

Figure 3. Bibliography record for Boissier, Flora Orientalis.

Figure 4. Gazetteer record for Vilayet Bolu.

Records

The RECORDS file contains the details of all specimens, including types, referring to the place of collection in the gazetteer, the collector as a committee, the date and all of the types of information that regularly appear on herbarium labels. Figure 5 shows type specimen data for an Amygdalus. PANDORA also includes a complete system for printing herbarium labels, based on information stored in the database. In addition, the DISTribution file holds statements of the occurrence of a taxon, referred to the nomenclature file, which are found in an area (in the gazetteer) with its status. These statements are not necessarily based on specimens but may come from publications.

Figure 5. Type specimen details for *Amygdalus trichamygdalus* var. *elongata* Browicz.

Figure 6. PANKEY showing spine character for *Amygdalus* L.

Descriptions

Descriptions of the morphology of the taxa are expressed as states of characters which are coded in the DELTA format (Description Language for Taxonomy, Dallwitz & Paine, 1986. A TDWG standard). There is a file for the definitions of characters (qualitative with states or numerical) and a data matrix, which is a rectangular table of taxa and characters. Descriptions of specimens can also be stored and used to compile the descriptions of the corresponding taxa. This information can be processed to produce diagnostic keys, either automatically or interactively, printed descriptions (as in the printed Flora) and other identification aids by the PANKEY program package (Pankhurst, 1995). Figure 6 displays the variation of a character for Amygdalus species in the interactive identification program. These programs can be called from within the PANDORA system.

Images

All kinds of images can be stored with the database and viewed via the database. Obvious examples are maps (Figure 7), pictures of specimens and of drawings, paintings and photographs, and text of original descriptions (Figure 8). These kinds of information are usually only available at the original institution or in libraries, and it is much better to make them generally available to everyone.

Reporting

There is practically no limit to the useful reports that can be obtained from a taxonomic database. The system provides some reports ready made, and others can be created as required in a variety of different ways. Examples shown here are Figure 9, a report on the nomenclature of *Amygdalus* L., Figure 10, with a list of distribution records for Turkey, and Figure 11, showing vernacular names for *A. communis* L.

Advantages

Speed

One of the first benefits claimed for working with computers is speed, and this is true when applied to plant taxonomy as to anything else. However, since much of the data processing involved is text oriented, the benefits of speed are usually evident at a later stage, once sufficient data has been gathered. So, for example, if data from herbarium specimen labels has been computerised and used to plot distribution maps, and some specimens are redetermined, then the subsequent reconstruction of the maps can be very rapid. Similarly, after a change of synonymy, a fresh report of the current taxonomy can be prepared very quickly.

Copying

Once a database is established, it is extremely easy and cheap to make multiple copies, both for sharing among members of a project and for the purposes of publication. A traditional Flora may well be a splendid production, as is for example the Flora of Turkey, but its price puts it out of the reach of many potential users. Also, as it goes out of print, it will cease to be available at all, whereas the computer data can be kept indefinitely, and provided that it is actively maintained and used, will remain available always. It hardly needs to be said that Flora data should be put onto the internet, for general access.

Capacity

Modern data storage is cheap and abundant, so there does not need to be any limit on what can be included for reasons of limited space or overriding cost. Authors of conventional Floras have made a virtue of conciseness,

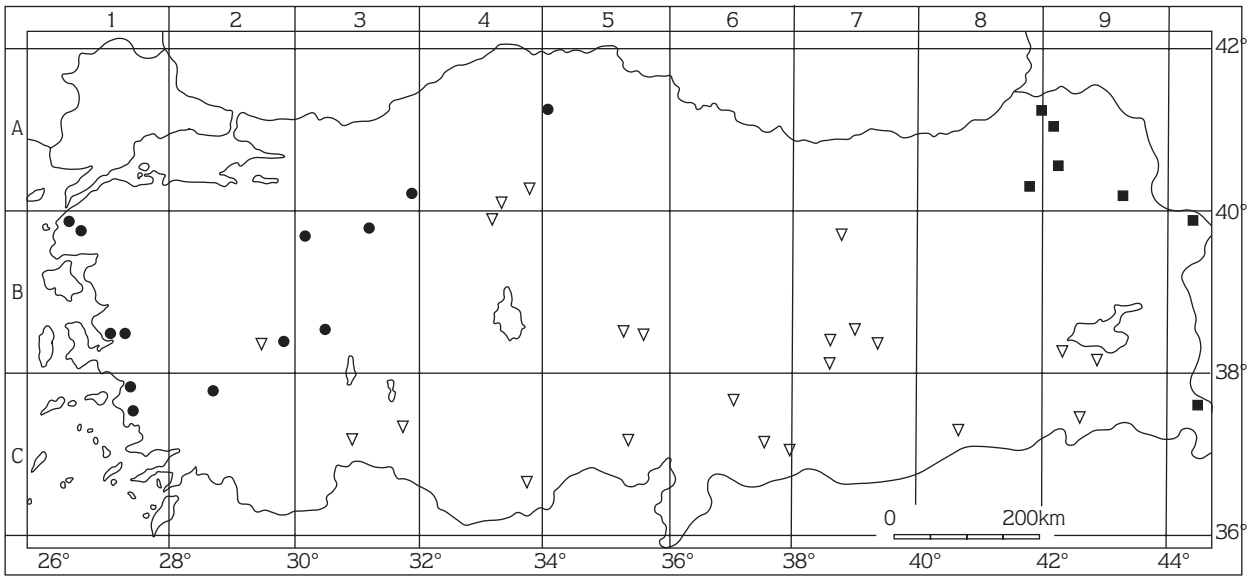

Figure 7. Distribution maps of *Amygdalus.*

****Prunus** ***(Amygdalus) trichamygdalus*** **Hand.-Mzt., sp. nova** ***(Amygdalus communis*** **Boiss., Fl. or. II, p. 641 p. p., non L.) (Taf. II, Fig. 2).**

Arbor elata ramis elongatis ascendentibus brunnescenti et griseo corticatis, laevibus et glabris, inermibus. Ramuli abbreviati copiosi, 5—8 mm longi, cicatricibus densis nigrelli, glabri, apice fasciculum foliorum patulorum gerentes. Stipulae lanceolatae, scariosae, brunneae, 1·5—3 mm longae, glanduloso-fimbriatae. Folia petiolis brevissimis latis superne profunde canaliculatis, 1·5—3 mm longis suffulta; parva; elliptica, 7 × 19 usque 11 × 29 mm, basi subrotundata, apice recurvulo subcomplicato acutiusculo, toto margine densissime subtiliter glanduloso-crenata, glandulis basilaribus utrinque singulis vel binis magnis aurantiaco-brunneis; rigidiuscula, concolori dilute et subcaesio-viridia, opaca, superne vix nitidula; nervo mediano tenui, inferne basin versus incrassato et valde prominente, nervis lateralibus utrinque ± 8, leviter prorsus curvatis, ante marginem arcuato-confluentibus, cum venularum reti denso in facie superiore magis conspicuis; in pagina inferiore pilis brevibus saepe nitidulis large crispule puberula. Gemmae inflorescentiarum parvae, ovato globosae, bracteis triangularibus, siccis, brunneis, margine dense et breviter albo-ciliatis, interioribus toto dorso breviter velutinis. (Flores ignoti.) Drupa nutans, sessilis, sicca, dura, ovata, 20—22 mm longa, 15 lata, parum complanata, basi transverse truncata, apice breviter acuminata, ventre anguste sulcata, brunnea, rugulosa, utrinque pilis brevissimis griseis patentibus densissime velutina.

In Wäldern auf steinigem Boden im Tale von Urik am Nemrud Dagh bei Kjachta im kataonischen Taurus, Kalk, 1200—1400 m, 12./VII. 1910 (Nr. 2134).

Fig. 8. Original description of *A. trichamygdalus* (Hand.-Mazz.) Woron.

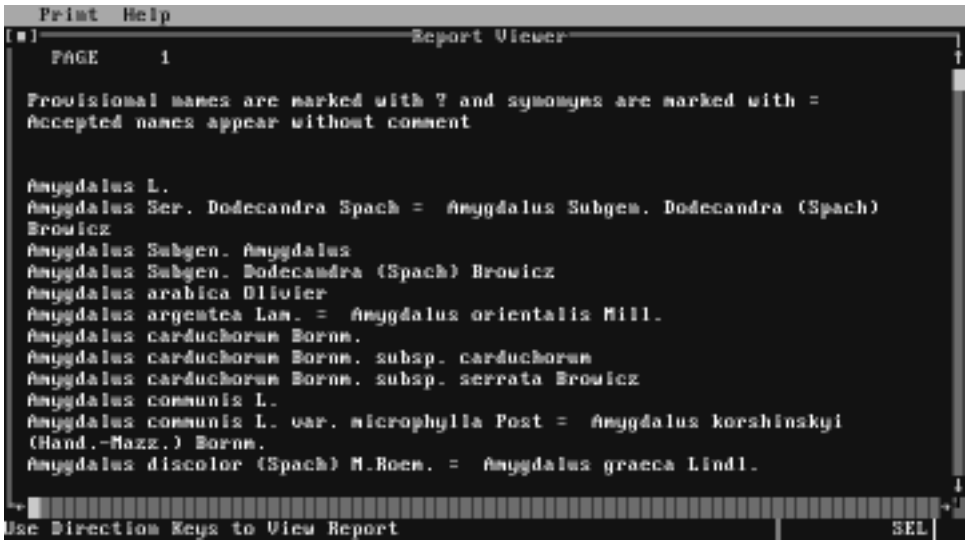

Figure 9. Nomenclature of Amygdalus from Flora of Turkey.

which is obviously what users need in some circumstances, but all too often the result is just brevity, and excessive brevity at that. When attempting to convert text descriptions of taxa from Floras into DELTA format, it has often been this author's experience that there is only just enough information to distinguish each taxon from every other by only one character, or worse, there is less than this minimum level of information, so that the

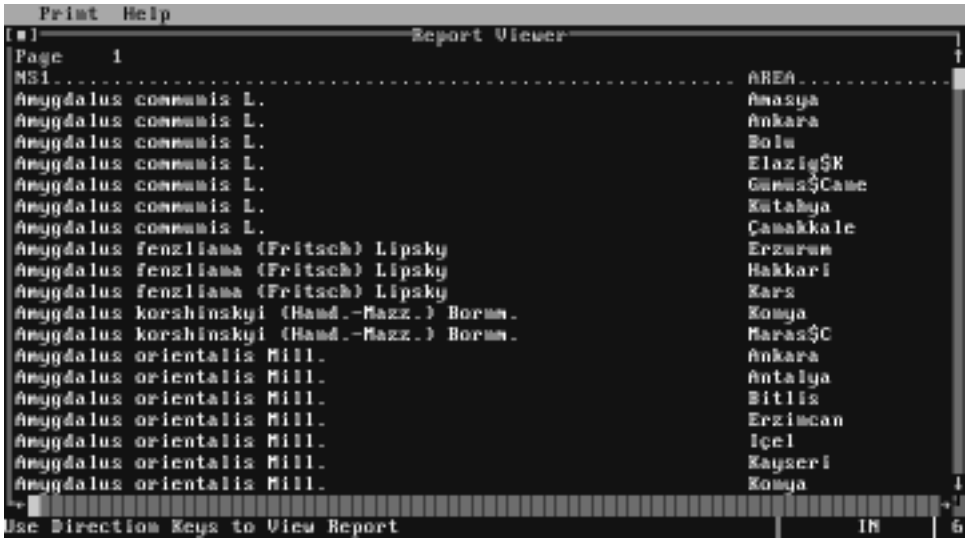

Figure 10. Distribution of Amygdalus in Turkey.

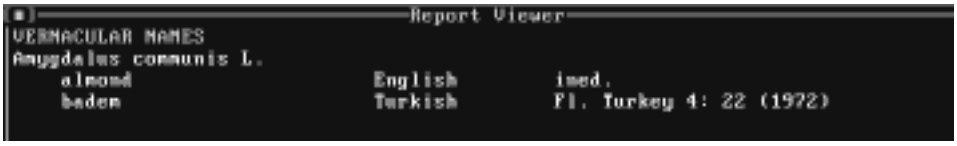

Figure 11. Vernacular names for *A. communis* L.

taxa cannot all be distinguished from one another. When storing plant data in computers, there is now every reason to include as much data as possible, especially bulky items such as images, maps and specimen label data. Conciseness, or the selection of those parts of the data that are of immediate interest for some specific purpose, can then be managed by software. Flora projects should always keep all their materials so that workers in later projects can use the information that already exists and not have to waste effort in recollecting what is already known.

Accuracy

The use of computers does not of itself automatically guarantee any degree of accuracy, but a well-managed database with careful data checking can achieve a much greater level of accuracy than any method based on traditional editorial methods. This is not meant as a criticism of the quality of work carried out by the editors of traditional Floras, and editorial diligence is still needed when compiling a database, but it is simply the result of better methods. One such improvement is that data can often be checked for accuracy at the moment of entry, and immediately corrected, rather than waiting for manual proofreading or editorial checks at a later stage. For example, the year of publication in a citation can be immediately compared with the known date or range of dates of a book or journal, which are also stored in the database. There are many ways in which data can be used to make crosschecks within a database. As another example, the date of publication of a taxon can be compared with the known date or range of life dates of the author. When this was done for Boissier, it was found that taxa from this author continued to be published until 10 years after his death!

Flexibility

A printed Flora is fixed in its scope and presentation, and therefore cannot serve very many purposes at once. As an example, it would be possible to produce checklists of taxa for the individual administrative units of Turkey (vilayets) from the printed Flora, as the necessary information is in fact given, but it would be laborious, since the data is ordered the other way round, by species

and then by district. A database could be very easily made to produce such reports. In fact, data can be retrieved in any order or combination that is desired, with selections by date, author, geography and so forth. Provided that the database is properly maintained, then the floristic information can be kept up to date more or less continuously, whereas the data in a published Flora is frozen at a particular moment of time.

Consistency

It is not only desirable but also vitally necessary to follow data standards when compiling databases. For example, the authority strings should be entered according to the abbreviations given in Brummitt & Powell. Abbreviations for the names of books might be made to follow those given in the standard reference, TL2, and for journals in BPH. This not only contributes to accuracy, as explained above, but makes it possible to combine data from different databases. For example, once the authority abbreviations from Flora Europaea and from Med-Checklist were converted to the standard they could then be merged for the Euro+Med Plantbase.

Disadvantages

Data collection

It is sometimes said that data collection is harder by computer than by manual methods. This might be because the computer is less tolerant of errors, and if so, then the possibility of achieving increased accuracy may make the effort worthwhile. On the other hand, instant error correction by computer shortens the task. Probably, data collection is about equally laborious either way.

Training

Staff need to be properly trained in order to make a success of computerised Floras. Experience with the Euro+Med suggests that most of the training is needed to create an understanding of the relations and structure of the data, rather than in learning how to use a specific software interface. While some potential users have expressed nervousness about the amount of learning required in order to become familiar with PANDORA, the data entry staff on Euro+Med found a 2-day course to be sufficient.

Disastrous mistakes

It is well known that hardware error or human carelessness with computer systems can cause immense destruction of data. This can be avoided by proper management, such as by making regular backups.

Instability

While it can be an advantage to be able to keep a database continuously updated, users may find the constant change to be an annoyance. This may be overcome by producing numbered versions at known dates and with a proper version history, as is done with the ILDIS database for the Leguminosae (www.ildis.org/LegumeWeb).

Conclusions

While the revolution will not take place immediately, the replacement of traditional methods of writing Floras by taxonomic databases is seen as irresistable and inevitable.

References

Brummit RK & Powell CE (1992). *Authors of Plant Names*. Kew: Royal Botanic Gardens. See also the International Plant Name Index (IPNI) website at WWW.IPNI.ORG/IPNI/QUERY_AUTHOR.HTML

Dallwitz MJ & Paine TA (1986). *User's guide to the DELTA system – a general system for processing taxonomic descriptions.* Third edition. CSIRO Div. Entomology, Canberra, Australia, Report 13. See also the URL BIODIVERSITY.BIO.UNO.EDU/DELTA

Davis PH (ed.) (1972). *Flora of Turkey and the East Aegean Islands*. Vol. 4, Rosaceae to Dipsacaceae. Edinburgh: Edinburgh University Press.

Godfray HCJ (2002). Challenges for taxonomy. *Nature* 417: 17-19 (2nd May 2002).

Greuter W, Burdet HM & Long G (1984-) *MED-CHECKLIST. A critical inventory of the vascular plants of the circum-mediterranean countries.* Conservatoire et Jardin botaniques, Genève.

Lawrence GHM, Buchheim AFG, Daniels GS & Dolezal H (1968) *B-P-H. Botanico-Peridocum-Huntianum*. Hunt Botanical Library, Pittsburgh, and Supplementum B-P-H/S in 1991.

Pankhurst RJ (1993) *Taxonomic Databases: the PANDORA System*. In Advances in Computer Methods for Systematic Biology ed. R. Fortuner, pp. 229-240. Baltimore & London: Johns Hopkins UP. Also see website WWW.RBGE.ORG.UK.

Pankhurst RJ (1995) *PANKEY, Programs for the identification and description of plants or animals*. New York: Exeter Software, Setauket. [Website at WWW.EXETERSOFTWARE.COM]

Stafleu FA & Cowan RS (1976-1988) Taxonomic Literature. Edition 2. Volumes 1-7. Utrecht: Bohn, Scheltema & Holkema, and Supplements 1- (1992-), Koelz, Germany. Often known for short as TL2.

Tutin TG, Burges NA, Chater AO, Edmondson JR, Heywood VH, Moore DM, Valentine DH, Walters SM & Webb DA (1964-1993). *Flora Europaea*. Vols. 1-5 and Vol.1, ed.2. Cambridge: Cambridge University Press.

For databases at Royal Botanic Garden, Edinburgh, see WWW.RBGE.ORG.UK