Information standards in botanical databases – the limits to data interchange

Barry J. Conn

Abstract

Conn, Barry J. (Royal Botanic Gardens Sydney, Mrs Macquaries Road, Sydney, NSW 2000 Australia). Information standards in botanical databases – the limits to data interchange. Telopea *10(1) 53–60.* Computers and databases are generally intolerant of inconsistency. Strict standards in data structure and content are necessary to ensure effective management and retrieval of information and communication between databases. The problem in taxonomic botany is not the lack of standards but that there are so many of them from so many different sources that knowing which ones to choose for a particular database situation is difficult. The choice between alternative, competing and often contradictory standards is confusing and can require substantial time in the development of botanical database applications. Botanical databases have been developing rapidly in functionality and content over the past two decades, drawing on international and local standards from other disciplines and creating standards of their own. This paper discusses the practical difficulties of implementing the 'Herbarium Information System and Protocols for Interchange of Data' (HISPID) including problems associated with: data dictionary variants within the standard; implementation of localised standards and styles; data processing errors and impact on data transfer scripting programs; professional work preferences (particularly, the conflict between data capture versus data interpretation, and published versus unpublished data).

Introduction

The Taxonomic Databases Working Group (TDWG) evaluates and endorses standards relevant to biological databases. This Working Group facilities and coordinates the development of new standards when these are needed. The use of such standards is critical to the success of any collaborative *Flora Malesiana* information management project, as it has been for the developing *Australia's Virtual Herbarium* project (Barker 1998). However, the consequence of strict prescriptive adherence to rigid standards is that although it ensures consistency, it constrains innovation. Therefore, a balance between strict adherence to standards and the flexibility offered by non-standardised applications is required. Innovation results in new applications and improved ways of managing botanical data. However, the continual search for improvements reduces the usefulness of standards.

This paper uses as a basis for its discussion the 'Herbarium Information System and Protocols for Interchange of Data' (HISPID) (Croft 1989, Whalen 1993, Conn 1996, 2000).

Institutional abbreviations follow Holmgren et al. (1990).

Importance of Standards

As for other disciplines, data standards provide the common language, rules and protocols for the sharing and interpretation of information. Herbarium collection databases have attempted to incorporate these standards into the structure and management of electronic collection information. However, even when existing databases are unable to comply with the standards, the data interchange protocols (in

this case HISPID) have been developed to employ these standards as the common interchange 'language'. For example, even though most Australian herbaria record date information in the form of either DDMMMYYYY (namely, 02 Sep 1978) or DDMMYYYY (02/09/1978), these data are interchanged in HISPID format according to the International time notation (ISO 8601, 2000) of YYYYMMDD (namely, for the above example, 19780902) [HISPID field transfer codes[1] *cdat*, *cdat2*, *credit*, *loadat*, *texcdat*, *texvdat*, *vdat*]. Nomenclatural standards (rules and protocols) (Greuter et al. 2000, Trehane et al. 1995) and standard author abbreviations (Brummit & Powell 1992, Brummit & Davies 1998-) have provided an important framework for referring to plant names within databases. Bisby (1995) developed a specific database information model for plant names to cope with nomenclatural data within a database. Other standards that have been incorporated into database structures or interchange protocols include:

- standard abbreviations for scientific publications (Bridson & Smith 1991; Lawrence et al., 1968; Stafleu & Cowan 1976-);
- official organisational abbreviations (Holmgren et al. 1990) [HISPID – *don*]; and a standard for recording relationships between plants and places (WCMC 1995) [HISPID – *posnat*, *poscul* and *posint*];
- 'Dublin Core' metadata that have been incorporated into the HISPID transfer protocols for describing multimedia resources (Anonymous 1996);
- transfer of spatial data based on the 'Spatial Data Transfer Standard' (Anonymous 2002). The 'Australian and New Zealand Land Information Council' (ANZLIC) has supported this spatial standard (ANZLIC 2002) for use in Australia and New Zealand.

One of the strengths of the HISPID interchange standard is that it has been able to draw on the expertise incorporated into the many specialised international standards. The actual transfer format of HISPID is based on the 'Abstract Syntax Notation One' standard (ISO/IEC 8824, 1990) that aims at specifying data used in communication protocols. However, the difficulty of incorporating so many standards into a single standard is that they are frequently not completely compatible. Furthermore, some important standards appear to be of limited long-term value. For example, the geographic scheme for describing plant distributions (Hollis & Brummitt 1992) appears to have been largely superseded by more modern geographic information system schemas. The 'Open GIS Consortium' is an industry association that provides a consensus process for adopting spatial data specifications for interfaces and protocols that enable interoperable geoprocessing services, data, and applications (OGC 2002).

Sharing of electronic accession data

The need to manage the storage and retrieval of Australian biological data was first recognised in the 1970s as an important requirement for exchanging data between institutional datasets (Busby 1979). The 'Australian Biotaxonomic Information System' (ABIS) was proposed as a distributed database model, with data management being the responsibility of participating agencies (*loc. cit.*). The ABIS standard was a refinement of the core data standard developed by the 'Australian Biological Resources Study' (Underwood 1975). ABIS standard was further developed through usage by the Australian Museum, the Queensland Herbarium and the Western Australian Museum. Potential participants in ABIS were also involved in its development. The need to share electronic data between herbaria was more generally

[1]Note: The three- to seven character abbreviations (in italics) refer to the transfer codes used in tag-formatted HISPID3- and HISPID4-files (Conn 1996 & 2000, respectively).

first recognised by the major Australian herbaria in the 1980s. HISPID was first published in 1989 to provide a standard electronic interchange framework for herbarium collections information (Croft 1989). The rationale for sharing electronic herbarium data between Australian herbaria was based on concerns about unnecessary data-entry of replicate material held at each herbarium. This replication of effort not only introduces potential data processing errors, but it also consumes considerable personnel and financial resources, as well as time (Conn 1998). The detailed analysis of data-entry costs provided by Conn (1998) clearly illustrates the financial savings resulting from minimising the unnecessary data processing of the same herbarium label information held in several institutions.

HISPID is a tag-formatted accession-based interchange data-dictionary, even though many fields also refer to taxon-level attributes, such as nomenclature, bibliographic, and typification data (Whalen 1993; Conn 1996, 2000). The botanic gardens community has implemented a similar interchange format (Wyse Jackson et al. 1997). Although HISPID is concerned primarily with electronic data exchange, it has often also been used as a guide in developing database structure. Those agencies that have closely aligned their databases with the standard have been able to readily import and export data using HISPID. Currently, HISPID4 (Conn 2000) is being redrafted as an XML-formatted standard (refer HISPID5 Standard – http://sourceforge.net/projects/big). In part, this will overcome the challenge that HISPID4 presents for data managers, namely, the need to interchange data in a flat-file format. The new HISPID standard will improve the transfer of nested information and repeated elements, such as verification history of the plant name [HISPID – *vhist*] represented by the record.

Data dictionary variants within HISPID

The *Council of Heads of Australian Herbaria* (CHAH) based the development of HISPID on a list of core interchange data fields then recognised. There was unanimous agreement on some core fields, such as the plant name fields; however, other fields included in the standard reflect components of the various CHAH databases. This compromise solution resulted in different ways of transferring some of the same data components. Therefore, even though agencies may be HISPID compliant, data from one database may be difficult to import into another. For example, altitude data (in metres) held in the *NSW Collections* database of the Royal Botanic Gardens Sydney (NSW), including the National Herbarium of New South Wales, is stored as a single value with an associated accuracy field (in metres). Whereas, the *ADHERB* database of the State Herbarium of South Australia (AD) records these data in two fields that represent the altitudinal range of the collection; one for minimum altitude, the other for maximum altitude (both in metres). Therefore, the *NSW Collections* database stores the altitudinal range values of 300 m [minimum altitude: – HISPID transfer code – *alt*] to 500 m [maximum altitude: HISPID – *altx*] from *ADHERB* as an altitude of 400 m [HISPID – *alt*], with a precision code of 100 m (ie. ± 100 m) [HISPID – *altc*]. Although both methods are HISPID compliant, both require conversion by data load scripts prior to incorporation into the other database. Some databases (eg. *MELSIR* – National Herbarium of Victoria and *NSW Collections*) use the aggregate field concept of 'Collection Notes' [HISPID – *cnot*] to store habit/life form, frequency, phenology and other miscellaneous notes about the plant. These data can be transferred either in aggregated form [HISPID – *cnot*] or as the component fields of data [HISPID – *form, fre, phe*, *misc*]. Both formats are supported by HISPID. However, data sent in either formats may well be stored differently by recipient institutions. Such reorganisation of the data is liable to introduce errors.

Components of descriptive spatial data are another example of alternative data transfer definitions within the HISPID standard. In particular, the transfer of primary recording units (= *pru*) (eg. States or Provinces of Countries) and secondary recording units (= *sru*) (eg. subdivisions of States or Provinces) have two alternative possible formats. Valid values of *pru* are defined as either "… in full or [in] standard abbreviations accepted by …" (Conn 1996, p. 66). Likewise, the latter field (*sru*) is defined as "written in full or any valid regional code or abbreviation … (*ibid*., p. 67). Obviously, a local convention of using abbreviations will result in difficulties for an agency that expects the data in an unabbreviated form. Although the *HERBRECS* database of the Queensland Herbarium (BRI) stores unabbreviated *sru* values, this database appends this information with the qualifier 'District'. For example, *HERBRECS* records 'Burnett District' and 'Gregory District' for two of the official subdivisions of Queensland. These *sru* values are stored in *NSW Collections* database (NSW) as "Burnett" and "Gregory", respectively, without the word "District".

Localised standards and styles

Modifications to an international standard such as HISPID are frequently introduced for a specific, albeit localised reason. Most of these are style format changes that do not impact on the standard; however, some changes do affect the ability to exchange freely data using HISPID. For example, *NSW Collections* capitalises the Family name of the plant record. Other databases (eg. *ANSHIR* of the Centre for Plant Biodiversity Research, Canberra – CANB) use normal sentence case for the plant family name of the record. Although NSW's use of upper-case is only for emphasis, especially when printed on herbarium labels, it means that data load scripts need to account for these data exchange variants.

The *ADHERB* database (AD) uses curly braces "{" and "}" in two different ways. These braces are used to convert the "+" symbol into "±" to represent 'approximately', as in "{+} 8 miles N of …". This would be printed as "± 8 miles N of …" on the herbarium label [descriptive locality field: HISPID – *loc*]. The other use of these braces is to enclose text that is to be emphasised by under-lining when printed, eg. "Open {Eucalyptus deglupta} Forest …" (= Open Eucalyptus deglupta Forest …") [habitat field: HISPID – *hab*]. The potential problem with the use of these curly braces for data load scripts is that they are reserved as special HISPID standard notation that defines the start of a record (namely, "{") and end of a record ("}"). Therefore, the load scripting has to be sufficiently robust to handle braces that occur within records.

In general, non-standardised abbreviations (with respect to national or international standards) have consequences on data quality within databases. Incoming data needs to be aligned with local database rules by data load scripting. However, inconsistency of data-entry frequently result in abbreviation variants, even though the local syntax is well-understood by the data-processors. Although the accepted protocol in *NSW Collections* is to minimise the use of abbreviations, variants are prevalent because of data-entry inconsistencies. Therefore, both approaches require comprehensive data load scripts to standardise the abbreviations according to the data standard of the local database. Common abbreviations and their variants include: TO, T/O (referring to 'turnoff', an Australian colloquialism for road junction; S, Stn ('station', referring to either an agricultural grazing lease or a railway station; HS (mostly referring to a 'homestead', including the land associated with the home, of a grazing lease, but sometimes referring to a High [Secondary Education level] School); R, Rd (the former usually referring to 'river', but both are used to mean 'road'); SF, S.F. (State Forest reserve); NP, N.P., Nat. Park, Natl Park (National Park reserve). More localised abbreviations that are often difficult to expand are used by particular disciplines TR, T.R. (Timber Reserve); LA, L.A. (logging area); Compart., Cpt. (Forestry compartment area of a timber reserve); dbh, d.b.h. (diameter [of tree trunk] at breast height).

Database structure

The length of database fields frequently causes problems when data are transferred from one institution to another. For example, *Texpress* (KE Software) databases as used by AD, MEL, PERTH and, in part, *EMu* (KE Software) databases (as used by NSW) have fixed length fields that frequently truncate incoming data because of longer field lengths. This is a widespread problem since fixed length fields are the norm for almost all database management systems (eg. also in ORACLE® database systems as used by CANB and DNA). Truncation can result in a corruption of data without appropriate data load scripting and possibly database restructure.

Data entry and data transfer scripting errors

Naturally, data transfer scripts are generally unable to automatically handle data-entry errors without a detailed examination of the data transfer file. This is less of a problem for databases that have highly atomised fields compared to those that group disparate information into single fields. As an example of a potential data load problem, the *MELSIR* database (MEL) is able to store the names of more than one collector in the database's collector's name field. If this information is not separated into the two transfer fields [HISPID – *cnam* and *cnam2*], when exported, then data load scripts will fail during the importing of these data. Likewise, some databases (*eg. MELSIR*) group the abbreviations of the institutions receiving replicate material of a particular record into a single string, according to a defined syntax. The transfer field [HISPID – *desrep*] expects the format to be 'organisation abbreviation1,[space]organisation abbreviation2,[space]organisation abbreviation3,[space] …'. If this format is inconsistently modified within a database during data-entry, then it is difficult for either of the data transfer scripts to correct automatically these modifications. Finally, errors in the export data transfer scripting, especially if related to records that are not in the HISPID-format, are difficult for data load scripts to correct reliably and automatically.

Sociological impediments to data interchange

The exchange of information between herbaria is as old as the exchange of herbarium material itself. However, the exchange of this information in an electronic format has been hampered by an inability of herbaria to donate and receive these data. Prior to the development of HISPID, this was certainly true for herbarium collection information. Indeed, it is still true for taxon-based information. For example, the database information model for plant names developed by Bisby (1995) does not provide an adequate framework for the transfer of botanical nomenclature. However, there are other than technical reasons why the interchange of data continues to be difficult. The reasons given are many; the need to protect the locality of sensitive taxa, regain the financial cost of generating these data, protection of intellectual property held within the data, and more generally, the protection of unpublished information. However, there are a few additional underlying fundamental reasons that impede the ready interchange of digital data. Two frequent reasons include the: (1) conflict that arises between data capture and data interpretation, and (2) status of published and unpublished data. Although these impediments are not directly related to conflicts between the application of data standards, they represent different philosophical approaches to the handling of data and information.

Data capture versus data interpretation

One of the most common conventions enforced on the data-entry process of herbarium label information is the almost universal requirement of capturing these data without modification or interpretation. There are two main reasons why this is regarded as important. Firstly, an electronic record of what the collector actually recorded is regarded as of historical importance. For example, the spelling of many place names has changed over time. Changing these spellings to modern equivalents looses the historical component of the data. Secondly, interpretation of the information is regarded as increasing the likelihood of introducing errors. Attempts to 'compartmentalise' text strings into separate data fields, particularly those describing habitat preferences and plant features, may be extremely difficult to do without changing or corrupting the intended meaning. The use of the square bracket notation (namely, "[" and "]" has been successfully used to identify text modified by the data processor. However, frequently, data interpretation is not done and this can significantly reduce the usefulness of this information. For example, there are currently 27 database records in *NSW Collections* database that record the collector as 'P.G. Wilson'. In Australia, there are two well-known collectors with these initials (namely, Paul G. Wilson – formerly of the Western Australian Herbarium, PERTH and Peter G. Wilson – National Herbarium of New South Wales, NSW). Although most, if not all of these records refer to collections made by the latter, the longer the collector's name is not fully identified, the more difficult it will become. The *MELSIR* database (MEL) has more than 200 database records that refer to the plant collector as 'P.G. Wilson' (P. Neish, *pers. comm*., 14 October 2002). In this case, most appear to refer to 'Paul G. Wilson'. All of these records are fully HISPID compliant with respect to this data field. Digital images of the botanical specimen, together with any label information and handwritten notes may provide an affordable solution that maximises both integrity of original data and usefulness of electronic data.

Published versus unpublished data

The traditional form by which we share botanical information has been, and continues to be, through publication in paper-based scientific journals and books. Unfortunately, this very limited definition of publication has not been readily expanded to include digital information. In the case of information held within botanical collections, it is assumed that the botanical object needs to be examined so that herbarium label information and other details can be accurately assessed by the user. Indeed, this is probably almost always true for certain applications of this information. Botanical specimens are readily shared between herbaria around the world for professional systematists to examine. This is regarded as an essential activity of all major herbaria. However, the sharing of the collector's information, contained on the herbarium label, in an electronic format raises concerns of ownership and ultimate usage. Although, the concerns appear to be greatest when data requests are made by the general community, commercial environmental consultants and other government agencies, these concerns also exist when electronic data are made available to colleagues at different agencies. Likewise, copyright concerns are actively and vigorously debated when these herbarium collections are presented as digital images, but not when available as physical collections *per se*. The reason for these differences of opinion, based on the nature of the information, is somewhat elusive. However, it may have something to do with our concerns about the lack of rigour in primary scientific data. I suggest that professional botanists are uncomfortable with exposing errors, from typographical to botanical identifications, to our colleagues and the broader community. In particular, the general community is probably unequipped to deal with, or allow for, these errors while using the data.

Paper-based publishing is a very structured (standardised) interactive process between writer, editor and publisher. Paper-printed scientific journals and books are remarkably uniform in structure (format and layout). However, in the botanical world, electronic publication of data (here, including the presentation of digital data as a form of publication) has not involved the general botanical community; but rather, the development has been handed to specialist technical computer experts and linkages to a few botanical specialists. Although it is far from true, there is a tendency for the general botanical community to assume that digital media are largely free-form, unstructured, rapid and not based on international standards. The lack of perceived standards ('controls') reduces electronic data to the status of 'work-in-progress'.

Finally, there is a strong concern that these data will be used by a third party to form the basis of other publications without due recognition of those involved with their creation. This has already happened on many occasions. Although, plagiarism has always been a concern, albeit minor, in all disciplines, measures are in place for controlling the inappropriate use of other workers' information. The more frequent use of joint publications is also providing a solution to this concern.

Conclusion

There is a need for the acceptance of digital data formats as an equally valid information medium for researchers and the general community alike. Data interchange standards are required as the framework for sharing these data. Conflict between available standards must be minimised so that information can be readily transferred between users. Since community groups, at least within Australia, are responsible for the health of the environment, they are demanding access to data and information that was previously only available to professional scientists. Technology is able to deliver these data and information rapidly and in many formats. The challenge for the professionals is to provide supporting explanation, documentation and caveats to assist the community to correctly derive and interpret this information, within the limits of these data sets. An additional challenge for those involved with the free transfer of electronic data is that although data interchange standards were first recognised as important 20 years ago (from 1973), the free exchange of these data has still not been fully realised.

Acknowledgements

I gratefully thank Peter Neish (MEL) for his useful comments on the draft manuscript.

References

Anonymous (1996) *Dublin Core Metadata Element Set: Reference Description* (OCLC Online Computer Library Center, Inc.) [http://purl.org/metadata/dublin_core_elements].

Anonymous (2002) *Spatial Data Transfer Standard (SDTS)* (United States Geological Survey: Rolla) [http://mcmcweb.er.usgs.gov/sdts/].

ANZLIC (2002) *The spatial information council* [http://www.anzlic.org.au/].

Barker, W.R. (1998) *The Virtual Australian Herbarium: a cooperative flora information system being developed by Australian herbaria* [http://www.rbgsyd.gov.au/HISCOM].

Bisby, F. (1995) *Plant Names in Botanical Databases. Plant Taxonomic Database Standards No. 3, International Working Group on Taxonomic Databases for Plant Sciences (TDWG)* (Hunt Institute for Botanical Documentation: Pittsburgh).

Bridson, G.D.R. & Smith, E.R. (1991) *Botanico-Periodicum-Huntianum/supplementum* (Hunt Institute for Botanical Documentation: Pittsburgh).

Brummit, R.K. & Davies, R.A. (ed.) (1998-) *Authors of plant names, pilot project for the Plant Name Project* (Royal Botanic Gardens Kew, Harvard University Herbaria & Australian National Herbarium) [http://www.ipni.org/ipni/query_author.html].

Brummit, R.K. & Powell, C.E. (1992) *Authors of plant names* (Royal Botanic Gardens Kew) [http://www.rbgkew.org.uk/web.dbs/authors.html].

Busby, J.R. (1979) *Australian Biotaxonomic Information System: introduction and data interchange standards* (Australian Government Publishing Service: Canberra).

Conn, B.J. (ed.)(1996) *HISPID3 – Herbarium Information Standards and Protocols for Interchange of Data*, version 3 (Royal Botanic Gardens Sydney).

Conn, B.J. (1998) *Sharing of Accession-based botanical information – Reduction of Costs in Herbarium Data-entry in Australia Using HISPID3* (Royal Botanic Gardens Sydney) [http://www.rbgsyd.gov.au/HISCOM].

Conn, B.J. (ed.)(2000) *HISPID4 – Herbarium Information Standards and Protocols for Interchange of Data*, version 4 (Royal Botanic Gardens Sydney) [http://www.rbgsyd.gov.au/HISCOM].

Croft, J.R. (ed.)(1989) *HISPID – Herbarium Information Standards and Protocols for Interchange of Data* (Australian National Botanic Gardens: Canberra).

Greuter, W., McNeill, J., Barrie, R., Burdet, H.-M., Demoulin, V., Filguerias, T.S., Nicolson, D.H., Silva, P.C., Skog, J.E., Trehane, P., Turland, N.J., Hawksworth, D.L. (eds & compilers)(2000) *International Code of Botanical Nomenclature (Saint Louis Code)*, adopted by the Sixteenth International Botanical Congress St. Louis, Missouri, July - August 1999. Regnum Vegetabile 138 (Koeltz Scientific Books, Königstein).

Holmgren, P.K., Holmgren, N.H. & Barnett, L.C. (1990) *Index Herbariorum, Pt. 1: The Herbaria of the World*, edn 8. Regnum Vegetabile 120. [http://www.nybg.org/bsci/ih/ih.html].

Hollis, S. & Brummitt, R. (1992) *World Geographical Scheme for Recording Plant Distributions*. Plant Taxonomic Database Standards No. 2, International Working Group on Taxonomic Databases for Plant Sciences (TDWG) (Hunt Institute for Botanical Documentation: Pittsburgh) [http://www.bgbm.org/TDWG/geo/default.htm].

ISO 8601 (2000) *International Standard Date and Time Notation*, version 2 (International Organization for Standardization: Genève) [http://www.iso.ch/iso/en/ISOOnline.frontpage].

ISO/IEC 8824 (1990) *Information technology –Open Systems Interconnection – Specification of Abstract Syntax Notation One (ASN.1)*, 2nd ed. (International Organization for Standardization: Genève) [http://www.iso.ch/iso/en/ISOOnline.frontpage].

Lawrence et al. (1968) *Botanico-periodicum-huntianum* (Hunt Institute for Botanical Documentation: Pittsburgh).

OGC (2002) *Open GIS Consortium, Inc*. [http://www.opengis.org/].

Stafleu, F.A. & Cowan, R.S. (1976-) *Taxonomic literature*, ed. 2 and Supplements. Regnum Vegetabile 125-.

Trehane, P., Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Spongberg, S.A. & Vrugtman, F. (1995) *International Code of Nomenclature for Cultivated Plants* (Quarterjack Publishing, Wimborne).

Underwood, J. (1975, unpublished mimeo) *Core data for regional data banks* (Australian Biological Resources Study: Canberra).

WCMC (1995) *Plant Occurrence and Status Scheme, a Standard for Recording the Relationship between a Plant and a Place*. POSS version 2.0 (WCMC, Cambridge) [http://plants.usda.gov/npdc/poss_standard.html].

Whalen, A. (ed.) (1993) *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*, Version August 1993 (National Herbarium of New South Wales: Sydney).

Wyse Jackson, D., Conn, B.J., Piacentini, R., Waldren, S. & Ward, C. (eds)(1997) *International Transfer Format for Botanic Garden Plant Records*, Version 2, Draft 3.2 (Botanic Gardens Conservation International: Kew) [http://www.rbgkew.org.uk/BGCI/news.htm].