Zeitschrift: Boissiera : mémoires de botanique systématique
Herausgeber: Conservatoire et Jardin Botaniques de la Ville de Genève
Band: 24 (1975-1976)
Heft: 1

Artikel: Some thoughts on computerisation of label data in major herbaria : with particular reference to type specimens
Autor: Brummitt, R.K.
DOI: https://doi.org/10.5169/seals-895533

Boissiera 24: 403-410. 1975.

Some thoughts on computerisation of label data in major herbaria, with particular reference to type specimens

R. K. BRUMMITT

SUMMARY

Some discussion is presented of possible ways in which AETFAT might promote use of EDP methods in the study of the African flora. A scheme involving recording minimal data, consisting of only four items of information for each herbarium specimen, might possibly be desirable and practicable for application throughout major herbarium collections, while recording of more detailed information seems to be possible in the foreseeable future only for relatively small areas under particularly intensive study. If any effort is to be made to provide a comprehensive guide to type specimens this should be done by photographing the specimens and pooling photographs from different institutions in one central depository.

RÉSUMÉ

Les différentes possibilités que l'AETFAT pourrait adopter pour utiliser les méthodes EDP (electronic data processing) dans son étude de la flore africaine sont présentées. Un schéma ne nécessitant qu'un minimum de critères est proposé; quatre informations pour chaque échantillon d'herbier pourraient être appliquées à tous les herbiers d'importance mondiale en donnant des résultats utiles et pratiques. L'utilisation de critères plus détaillés ne pourrait être prévue dans l'avenir que pour des régions restreintes, en cas d'études approfondies. S'il y avait lieu de fournir un guide compréhensible des échantillons types, ceux-ci devraient être photographiés. Ces photographies, provenant de différents instituts, devraient être groupées dans un seul dépôt central.

In October 1973 representatives of many European herbaria met at Kew to discuss the use of electronic data processing (EDP) methods in herbaria, and among them happened to be a number of members of AETFAT. Inevitably there was some informal discussion as to whether AETFAT might initiate any projects on the African flora, and it turned out that Dr Schreiber had already taken a lead in suggesting this in her letter to the *AETFAT Bulletin* (Schreiber, 1974). Any such action would apparently involve a great effort by the major herbaria, and with this in mind I agreed to throw out some suggestions at this meeting as to what this Association might or might not consider doing. As may well become apparent, I am a mere botanist and no computer expert, and I am not trying to sell any ideas. Any views expressed here are personal opinions and not necessarily those of my institution.

In recent years a number of papers have been published describing floristic projects in which computers play a major part in storage and sorting of data. Of such projects, those which seem so far to have proved successful have, as far as I

can ascertain, dealt with data from not more than about 50 000 herbarium specimens, often much less, which have usually been entered over quite a number of years. Furthermore, they have been clearly defined projects covering relatively small geographical areas which were usually fairly intimately known to the compilers, who may also have had considerable knowledge of collectors' names, handwriting and itineraries, and other background knowledge perhaps even including the plants themselves. These projects may be excellent and provide valuable results, but they do not fall within my terms of reference here, for I must emphasise that I am concerned with problems facing major herbaria which may hold between one and five million specimens from a wide geographical area. Here the sheer number of specimens and diversity of information are so great that they become special problems of their own. The work recently started at Pretoria, in a fairly large herbarium but concentrating on a fairly restricted area, will perhaps to some extent bridge the gap between the smaller projects and the larger herbaria, and we look forward with interest to hearing about this at this meeting (cf. Morris & Leistner, 1975).

General problems of computerising herbarium specimen data

A major herbarium is a vast store of various kinds of botanical information Some of this is readily available to users, while other information, such as a simple species list for any one country, may be very difficult or laborious to obtain. By judicious use of EDP methods it might be possible to make a lot more information readily available, possibly with relatively small output of money and man-power. On the other hand, I suspect that injudicious use of EDP methods, involving input of vastly more information than can be justified by the required end product, could result in colossal waste of time, energy and money and seriously hinder proper botanical work. The main problem seems to be to strike a correct balance between the difficulty of putting information into the system and the usefulness of the information to be taken out. There are, of course, other problems, such as the unreliability of information extracted uncritically from herbarium collections and the difficulties of keeping records up-dated, but botanists are well aware of these and it is not my intention to discuss them here.

The idea of having an inter-institutional data-bank from which we could extract all available information on diverse aspects of the flora of all Africa at the press of the proverbial button may sound attractive, but it would involve tremendous expenditure of time and money by perhaps a dozen or more major herbaria of Europe and Africa. I have done some simple and rough calculations for my own institution, Kew, producing figures for the entire herbarium since our work is not confined to Africa. The most optimistic forecast of the average length of time to record the usual label information (specific name, country, locality and date of collection, altitude, habitat, collector's name and collecting number, and perhaps some grid system reference) is said to be about three minutes per specimen. This would be 20 specimens per man-machine per hour, or 150 per day, or 33 000 per year, which would mean 150 man-years for our estimated 5 million specimens. Allowing two man-machines for our annual intake of over 60 000 specimens per year and one more to deal with up-dating

of information already processed, we would need an absolute minimum of 8 man-machines for 30 years, together with a very considerable amount of expert advice from botanists, before the back-log were cleared off. This figure may not seem excessive to some, and it would of course be considerably less if we confined ourselves to our African collections, but I suspect that in practice it would prove to be far too optimistic. The cost would certainly be several million pounds. And then there is even doubt as to whether the resulting data-bank could operate practically and economically, for it has been suggested to Kew that the amount of information might be so colossal that a single search of the files by the computer would take several hours. I myself do not consider that such expense of effort and money would ever be anything like justified by the benefits achieved. I do, however, wonder if a simplified system might be more feasible.

Theoretically there is no limit to the information about herbarium specimens which could be put on a computer. One could, for example, record which specimens have labels written in pencil, which specimens weigh more than ½ kilogram, or which are housed in cupboards above head-height, but such information would serve little or no purpose. Other items of information, such as the date of collection or whether the specimen is in flower or fruit, *might* prove useful on some occasions, but one must always weigh the likelihood of their being needed in out-put of *useful* information against the expense of time and energy required in putting them into the system for all specimens. We must get away from the idea that *all* information on a label must be processed just because it is there, and also that every scrap of information must be captured on the first run through the herbarium or else not at all. I am convinced that if EDP methods are to be applied throughout major herbaria in the foreseeable future it will be necessary to be ruthlessly selective about which items of information are to be recorded. Fortunately, it seems to me that one can achieve a very high return in terms of usefulness by employing only very few, and very simple, descriptors. If we can answer perhaps 80% of the serious questions likely to be asked by doing only 10% of the in-put work we must think very carefully before doing the other 90% of the work in order to achieve the other 20% out-put.

The first type of information to be eliminated would be anything of free format which cannot be selectively recalled, and this would include exact localities. When economy of time and effort, and storage space in a computer, are critical factors there seems to be little or no justification for recording information such as "ten kilometres south-east of Mbala on the road to Mpika", or even simply "near Mbala". Even if one should ever need to search for all specimens bearing such a locality, it would be either extremely difficult or extremely expensive to do so on a computer. On rare occasions when lists of localities are needed they can be copied by hand from the actual specimens. For selective recall the origin of specimens should be recorded only according to larger areas, preferably political or politico-geographical areas since grid systems such as latitude and longitude or the Universal Transverse Mercator system or the Geocode are extremely rarely or never recorded on specimens. Ecological data as they appear on specimen labels must also be regarded as free-format, and although this *might* be converted into a very much simplified system of single words for computer storage and recall this would involve so much time-consuming botanical editing that it appears to have no practical relevance. Altitudes are so

seldom recorded on specimens that anything like completeness could never be achieved.

What I suggest *might* be a practical proposition, if major herbaria wish to record label data for all specimens, would be basically to record only four items: the name of the species, the country of origin, the collector's name, and the collecting number. In addition it might be desirable to show which specimens are known or thought to be types. I am aware of other complications which could arise, such as specimens lacking collector's name or number, specimens distributed under institutional numbers rather than collector's numbers, specimens of uncertain determination, and so on, but these are minor details which need not concern us here.

At Kew, and no doubt in other herbaria too, this basic recording of minimal data could be done very simply, the species name and the country of origin (except in poorly curated parts of the herbarium) being read from the species covers and only the collector's name and number being read from the actual specimens, so reducing interpretation of collector's handwriting to the minimum of reading his or her name. In practice, the species name can be written on a sheet of paper as a major heading, with countries as subordinate headings, and beneath these it is only necessary to list collector's name and number for each specimen. If a specimen is in a red folder, and so is suspected of being a type, a simple symbol can be given after the number. When the labels are all in the bottom right hand corner of the sheet the information can be read off and recorded simply by lifting up the corner of each sheet successively without even opening the folder (but one is rather lucky if this happens in practice). Perhaps up to a hundred specimens could be recorded on one sheet of paper, which would then go to a typist for punch-typing.

This might be practicable, but would it be useful? I think it would, and it would permit a surprising range of potential queries to be answered. By recording only these minimal data one could still satisfy most of the serious demands which would be made even if one had a fully comprehensive data-bank. For example, the suggested minimal data would: i) afford species lists of the holdings within the herbarium for any country in the world, with indication of relevant specimens if desired; ii) allow print-outs of distributions country by country, again with indication of all relevant specimens if desired, of any species or genus (also for higher taxa by means of a simple thesaurus); iii) enable any specimen identified by collector's name and number to be located in the herbarium; iv) allow print-outs of determination lists for collectors; v) facilitate investigation of collectors' itineraries; vi) give list of all collectors whose specimens are held for any specified country or other region; vii) rapidly provide much information on what types, or possible types, are held; viii) facilitate interchange of determinations of duplicated collections between institutions, and ix) provide statistics on the holdings of the herbarium which might be helpful in curation. It would no doubt even be possible to develop a simple distribution map print-out showing presence or absence in each country, rather like the vice-country distribution maps which were popular in the British Isles before the more sophisticated *Atlas of the British Flora* based on grid squares was produced. What the minimal data would not do would be to provide lists of species according to altitude or ecological habitat, but it is very doubtful if such information could ever be satisfactorily processed in this way on a large scale without extensive editing, and the demand for it would

certainly not justify the expense of energy and money needed to in-put such information for all specimens in a major collection.

However, I do not wish to argue that it would never be desirable to have additional more detailed information available in a data-bank. Somebody working intensively on one country or one family might well consider it worth-while to make the effort to record precise localities, dates of collection of specimens, ecological information, and so on, and it would be desirable to allow the facility for this to be done by anybody so interested. What I envisage, in fact, is a two-level recording system, one level organised on an institutional basis recording only minimal data but being applied throughout the herbarium, and the other for more personal smaller projects in which more comprehensive data are added by botanists themselves for those countries or other areas, or alternatively for taxonomic groups, in which they are interested and in which the identification of the specimens has been verified. In time, of course, if it were considered desirable it might be possible to extend the more comprehensive coverage to more countries until all Africa were covered, or in the long term the whole herbarium was included.

I have myself done some small trial runs in order to compare rates at which the different levels of recording can be done. Using only the minimal data I was able to record 100 specimens in 22 minutes, using just over one sheet of paper. Then, recording comprehensive data, including localities, ecological data and so on, on a *pro forma* designed for this job, I found that it took 205 minutes to deal with the same 100 specimens, using 100 sheets of paper. Recording minimal data thus seems to be about 9 times quicker than recording comprehensive data. These samples were done in African *Acanthaceae* on material already familiar to me, without checking to see whether my interpretation of handwriting of localities or collectors' names was likely to be correct or checking which provinces the localities should be referred to, and done in several short spells of up to an hour. For a non-botanist working on unfamiliar material, attempting to be as accurate as possible, and doing the same job all day and every day, I have no doubt that these rates, particularly for the comprehensive data, would be much slower. Although it is not at all realistic to extrapolate from these figures to consideration of the whole collection, for those who wish to do so it seems that at these rates it would take about 12 man-years to do 5 million specimens by minimal data, as against about 110 man-years using comprehensive data. Current intake of specimens and up-dating of information would be additional.

We are hoping at Kew to produce more accurate and realistic estimates of the problems and time involved in such work by running three pilot schemes. One of these would record minimum data for all specimens, world-wide, of one family; the second would record minimum data for all specimens from one area of Africa in which we have a particular interest; and the third would record comprehensive data for all material from a smaller area, again in Africa. Unfortunately, owing to financial restrictions we have not yet[1] been able to get even these pilot schemes under way!

---

[1] While this paper was in press one pilot project, covering Kew specimens from Aldabra, has been carried out.

The question of type specimens

Type specimens seem to be a rather obvious target for EDP methods, and a register of such for tropical Africa might well sound like a project which would appeal to AETFAT members. An elaborate program for creating a type specimen register has already been developed at the Smithsonian Institution in America and is now operational (Shetler & al., 1973), and Dr J. Cullen at Edinburgh has recently initiated a survey of the pros and cons of such a scheme in European herbaria in connection with a working-party set up at the meeting held at Kew just a year ago. Again, however, the idea of recording detailed information, including what name each specimen is the type of, the bibliographic reference to the name, what sort of type it is, and so on, for all type specimens in a major herbarium (we are thought to have about a quarter of a million types at Kew) is a daunting prospect to any curator.

I suggest that the need for such a type register has been over-stated on occasions. The main service it would provide would be to tell taxonomists where to look for type specimens, but those who claim that no indices exist at present to guide them in this overlook a considerable literature on the subject. Since De Candolle's *Phytographie*, and perhaps before, a great body of information has been built up about the location of types, and thanks to the past activities of AETFAT we who work on African plants are very well-placed to go straight to the type of perhaps the great majority of names we encounter after consulting one or two standard reference works. In fact most holotypes can be located simply by knowing that the collections worked on by botanist X are now housed in herbarium Y, and it does not need a computer to make this information easily available. What is needed is more information simply connecting X with Y, made available in an alphabetical published list. But even if it was not stated in the original publication, it is common knowledge that, for example, the types of names published by Bentham, Hooker, Oliver, N. E. Brown, Bullock, Brenan, etc. etc., are almost certain to be at Kew. Why on earth should we go to the lengths of recording the full locality, date and collector and so on of all such types, to say nothing of the full bibliographic reference to the name, in order to tell botanists something which they already know? The single item of knowledge that Bentham's types are at Kew makes superfluous the recording of perhaps 20 fields of information for each of several thousand type specimens if locating holotypes is the object of the type register. Finding isotypes may be rather more difficult, but to the taxonomist this is not of any fundamental importance anyway unless the holotype is lost. The spin-off benefits from a comprehensive register, such as the ability to list all the types collected from one country before a certain date, are largely unnecessary luxuries, and if the price of buying these is high we should not fall into temptation, even if we should ever have the money available.

I myself can, in fact, see little justification for the massive undertaking of a comprehensive type register. As a purely internal project within an institution it would be quite superfluous, as it would almost certainly prove as easy to go to the actual specimens for information as to a computer, and the idea can only be justified, if at all, as an inter-institutional service. An institution taking the trouble to record all its own types would only benefit if other institutions did the same. But while some herbaria

are reasonably fortunate in having their known or suspected types filed in red folders, many others are not, and for these the project might seem to be impossible from the start. Many others which now house large numbers of types have virtually no staff to do the work. Even where staff could possibly be made available and red folders are already in use, it is often very difficult to find what name a specimen is the type of, and the problems of compiling a register would be enormous.

However, if the idea of recording minimal data for all specimens, as outlined above, should ever prove operational, it would in fact provide merely as a by-product, all the essential information necessary to help anybody locate types if the normal line of approach has failed. It is for a botanist to decide what is the type of a name, and he must then ask for a particular specimen, or specimens, usually quoted by the collector's name and collecting number if there is one. In cases of difficulty it would be possible to supply lists of all suspected types (or other material) from a particular country according to the genus and species under which they are now filed, though of course for most requests for information on types in a smallish number of species it would usually be easier to go direct to the herbarium. I feel sure that it would be possible to record minimal data for all specimens in a major herbarium in far less time than it would take to complete a comprehensive register of the types alone.

For a final point I would like to take up a comment made by Dr F. H. Perring at the meeting held at Kew last year. As he pointed out, the most important thing about a type is what the plant itself actually is, and no matter how many label data you record, EDP methods do not preserve this information. His suggestion that the best way to produce a catalogue of types would be to photograph them is, I think, an idea deserving very serious consideration. Members of AETFAT are well aware of the difficulties arising from the destruction during the last war of the Berlin herbarium, including the types of names of many African plants. If I may refer again to my own institution, Kew, I would point out that our herbarium building lies almost directly under the incoming flight path for London airport and is regarded as the greatest fire risk in the county of Surrey. No herbarium is exempt from the threat of destruction, and, quite apart from the information retrieval viewpoint, I think we should be giving thought to the desirability of making a photographic record of suspected types in as many herbaria as possible.

Photographing specimens is very quick, involves no interpretation at all of handwriting, and these days can be done surprisingly cheaply. When the Wallich herbarium was photographed at Kew recently by a professional firm they did 20 000 specimens in four months using one camera. At that rate, with some help from herbarium assistants, we could get through all our types at Kew in four years, or perhaps in two years using two cameras. Compared with the costs likely to be incurred in computer work, the expense of this operation might be quite modest.

If different herbaria could co-operate in this way the negatives could all be pooled and interfiled at one central point. Before photographing each specimen it would be necessary to place a card printed clearly with the appropriate abbreviation for the herbarium on the sheet, and to write on the same card the name under which the specimen is currently filed. To trace a type botanists would be obliged to use only the same degree of initiative as they need to look for a type in an actual collection of specimens. Dr Perring's suggestion was that one set of negatives should be filed

systematically while a duplicate set should be filed according to collector's name and number. Alternatively, it might be preferable to have only one set of negatives and use a simple card index cross-reference system to other items of information. A data-bank for type specimens could thus be set up providing a postal service which would not only tell us where a type specimen is housed but also, if required, produce a photograph of the specimen at an appropriate fee. Apart from the problem of identifying what is a type in herbaria not employing red folders, the major problems would be of finance and of finding a centre willing to house the negatives and supply photographs on request, but possibly a commercial firm might be interested.

LITERATURE CITED

Morris, J. W. & O. A. Leistner (1975) Progress with computerization of the National Herbarium, Pretoria. *Boissiera* 24: 411-413.

Schreiber, Annelis (1974) Electronic data storage for Flora of Africa? *Assoc. Etude Taxon. Fl. Afrique Trop. Bull.* 24: 46.

Shetler, S. G. (1973) An introduction to the botanical type specimen register. *Smithsonian Contr. Bot.* 12: 1-186.