VegItaly: Technical features, crucial issues and some solutions

D. Gigante1, A.T.R. Acosta2, E. Agrillo3, F. Attorre3, V.E. Cambria1, S. Casavecchia1, A. Chiarucci2, E. Del Vico3, M. De Sanctis2, L. Facioni2, F. Geri2, R. Guarino6, S. Landi2, F. Landucci1, D. Lucarini1, E. Panfili2, S. Pesaresi4, I. Prisco2, L. Rosati2, F. Spada3, R. Venanzoni1

1Department of Applied Biology, University of Perugia, 06121 Perugia, Italy
2Department of Environmental Biology, University of Rome 3, 00145 Roma, Italy
3Department of Environmental Biology, Sapienza University of Rome, 00185 Roma, Italy
4Department of Agriculture, Food and Environmental Sciences, Marche Polytechnic University, 60131 Ancona, Italy
5Department of Environmental Science ‘G. Sarfatti’, University of Siena, 53100 Siena, Italy
6Department of Environmental Biology and Biodiversity, University of Palermo, 90123 Palermo, Italy
7Department of Biology DBAF, University of Basilicata, 85100 Potenza, Italy
8Museum-Botanical Garden, University of Camerino, 62032 Camerino, MC, Italy
9Aspix Consultancy Company, 06023 Gualdo Tadino, PG, Italy

Abstract

VegItaly is at present the largest Italian vegetation database. It is the result of a collaborative project aspiring to represent a major reference for the Italian vegetation scientists. The paper emphasizes its benefits for phytosociological data management and describes the solutions adopted to solve several technical problems, like the treatment of different vegetation stratification systems, the conversion of vegetation cover values, taxonomic and syntaxonomic issues, data import and access. The structure of the taxonomic list produced to support the storing of data is described. It allows an easy management of synonymic relationships and is constantly updated according to new publications and revisions. Issues related to data import from different formats have been solved by developing assistant software VegImport and TabImport, which are based on the most used formats in vegetation plot archiving. Bibliographic sources are managed according to the LISY standard and include descriptive geographic information, bibliographic and syntaxonomic reference. Distinct data access regimes can be selected by VegItaly’s users: visible, partly visible, invisible. Compared with the original project outline, many fundamental parts of the database structure have been fully built up, although several utilities still have to be developed or improved.

Key words: data sharing, ecoinformatics, Italian vegetation database, phytosociology, syntaxonomy, taxonomic list, vegetation plot

Introduction

Data sharing according to standard protocols is one of the most prominent topics in ecological studies (Bekker et al., 2007; Michener & Jones, 2012). This is particularly true for vegetation data, often collected for different purposes and across a wide range of scales, both by individual researchers and large teams, producing a massive amount of information often hidden and scattered in a huge number of private archives. In the last decade, a number of national and international projects rapidly stimulated a collaborative approach to vegetation data sharing, all over the world and especially in Europe. Among all, GIVD can be mentioned as one of the most recent and wide-scale coordinated initiatives related to this developing issue (Dengler et al., 2011).

In this framework, VegItaly stands as the largest rising initiative in Italy, started as a collaborative project and aspiring to represent a major reference for the Italian vegetation scientists, offering adequate tools cut to user’s measure and a scientific support to data storage.

In spite of a large territorial and social heterogeneity and an initial reluctance of potential contributors to deliver their data, VegItaly became in the last years more and more reliable and its users more and more persuaded to cooperate to the development of proper tools and solutions for data archiving and processing by a constant implementer-user feedback.

After an overview about VegItaly’s history, main aims and basic structure (Venanzoni et al., 2012) and a first report about the content of the database after two years of activity (Landucci et al., 2012), some practical aspects of this Italian vegetation database are here presented, with special focus on its comprehensive benefits for phytosociological data management.

Structural features

VegItaly is a web geodatabase built by using open-source software, designed to archive, retrieve and analyze vegetation data (Venanzoni et al., 2012). The data input works through plastic template, where users have to fill preset fields. In order to keep the system...
constantly efficient and suited to store such a peculiar type of information as vegetation data, users can also improve the template by asking for a supervised introduction of new fields. This type of feedback is the basic ingredient to obtain an adequate data harmonization and to avoid loss of information.

The developed system is setted in such a way that, when uploading data, some fields must obligatorily be filled. This is the case of data based on a dictionary, e.g. "data access type" or "cover scale", whose definition is mandatory, being essential for data processing. In case the user leaves these fields empty, they will be filled by default. There are also other types of mandatory data, which can be inserted as a free text, e.g. name, date, locality. They link the vegetation information to a number of details, mainly included among the header data, which allow geographic location, ecological comparison, monitoring in time and many other possible analyses.

As concerns cover values, different scales can be used. Currently the accepted ones are the original Braun-Blanquet scale (1928, 1964), Pignatti and Mengardo scale (1962), Braun-Blanquet scale modified by Barkman et al. (1964), van der Maarel ordinal scale (1979), percentage cover and presence/absence values. A set of mathematical functions have been developed ad hoc in order to convert one scale into another, even in the export stage (Fig. 1): this is a necessary step in order to process data and a well known problem in vegetation survey (Londo, 1976).

Also all the taxonomic entries are univocally treated by a proper dictionary, represented by the system taxonomic “master” list; this is a peculiarity of VegItaly, that makes it different from most of the other databases, where each project can manage its own check-list (e.g. TURBOVEG, Hennekens & Schaminée, 2001). Nevertheless, although one single database holds the whole set of data, specific projects must be created to manage data separately, on a geographical, thematic or logistic base, representing another structural peculiarity of VegItaly.

The system allows (and this is strongly recommended) to store data in their original form, even when this implies the use of obsolete taxonomic synonyms; valid names can always be put beside and kept visible for search engines, or even added in the export stage. This is a very sensitive and important point, related to the historical value of data and to the opportunity to keep a robust link to the original documentary sources, trying to contrast the so called “cultural” filter that can affect data availability, completeness and reliability as indicated by Berendsohn (1995) and Swetnam et al. (1999).

Differently from a classical taxonomic system, the “basic unit” of the taxonomic list is represented by the complete name of a taxon, formed by its genus plus all the lower taxonomic epithets available for that taxon (species, subspecies, variety, subvariety, form, race, subsinus, cultivar, when present). The only hierarchical relation is between each “basic unit” and the family it belongs to. At computer level, this unconventional structure allows to handle synonymic relationships and to shift from one family to another. The taxonomical complexity has been managed by keeping the link between each synonym and the corresponding valid name, still far from the complications deriving from a hierarchical management of, e.g., family synonymy and taxonomic splitting. Although rather simplified, this structure allows the possibility to have a flexible system, susceptible of modifications according to nomenclatural and taxonomical revisions. The synonymic link is maintained also in the import stage and is revealed by a list of suggested names.

As concerns the syntaxonomical classification, this is managed by the system in a not-hierarchical way. The complete information about all the phytosociological ranks, available in the original documentary source, can be uploaded during data entering or added afterwards. Short abbreviations followed by a colon are used to give a simple, basic structure to this field: “AS:” stands for association, “SA:” for subassociation, “AG:” for vegetation group, “VA:” for variant, “FA:” for facies, “AL:” for alliance, “SL:” for suballiance, “OR:” for order “SO:” for suborder, “CL:” for class, “SC:” for subclass, “SS:” for superclass.

**Taxonomical and nomenclatural issues**

The taxonomic list used by VegItaly was created on purpose for the database system anArchive (www.anarchive.it). This list was developed to support the storing of botanical data archiving and to facilitate any taxonomic and nomenclatural management in both herbaria samples and vegetation plots, including historical data (Venanzoni et al. 2012).

The master list includes 31,963 names of specific and infraspecific taxa (counting both valid names and synonyms), referred to 313 families. It is composed by five thematic sub-lists:

- **Cormophytes**, currently including 11,718 valid names and 11,528 synonyms; this sub-list has been compiled taking into account a large number of bibliographic sources: the most acknowledged floras of Italy and Europe (Pignatti, 1982; Tutin et al., 1968, 1972, 1976, 1980, 1993), the most recent Italian taxonomic check-list (Conti et al., 2005, 2007), the inventory of non-native flora of Italy (Celesti-Grapow et al., 2009), the Euro+Med PlantBase (Euro+Med 2006), several regional contributions and new records (e.g. Aeschimann et al., 2004; Arrigoni, 2006-2010; Bacchetta et al., 2009; Banfi & Galasso, 2010; Raimondo et al., 2010; Arrigoni & Viegi, 2011).
- **Alien cryptophytes**, currently including 1,373 valid names and 789 synonyms, restricted to exotic taxa not included in the Italian flora; main sources for their names are GRIN (USDA Germplasm Resources Information Network - http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl), the International Plant Names Index (IPNI, 2012) and the World Checklist of Selected Plant Families (WCSP, 2012; http://apps.kew.org/wcsp/home.do)

- **Bryophytes**, currently including 1,198 valid names and 2,983 synonyms of hornworts, liverworts and mosses of Italy according to Aleffi et al. (2008), Hill et al. (2006) and Grolle & Long (2000)

- **Algae**, currently including 34 valid names and 30 synonyms of macroscopic algae occurring in fresh and brackish water (Characeae and Cladophoraceae), mainly based on the recently published Italian flora of stonewort (Bazzichelli & Abdelahad, 2009).

- **Fungi**, currently including 1,493 valid names and 871 synonyms, based on Onofri et al. (2005).

The last four sub-lists are still in progress, while a lichen sub-list is under construction. Taxonomic names of families and their relationships with genera are in accordance with Smith et al. (2006), APG III (2009), Chase & Reveal (2009), Peruzzi (2010), Christenhusz et al. (2011a, 2011b) and Reveal & Chase (2011). The complete master list is available online at http://www.anarchive.it/anArchive/specie/browser.jsp. As explained before, each “basic unit” is linked to its genus and family. Further details associated to each taxon, such as nomenclature references, family, chorological type, lifeform and synonyms are also available. Hybrids are included as well as, in this case the parent species are also indicated.

During data entering, the system automatically suggests possible correct names, comparing the typed letters with the names included in the taxonomic list. Even taxa with uncertain determination can be managed by using provisional attribution (e.g. *Ephedra cfr. fragilis* Desf., *Sanguisorba minor* Scop. subsp. cfr. minor, *Festuca* sp.), and taxonomic aggregates (e.g. *Hyeracium murorum* L. aggr.).

The taxonomic list is continuously improved and updated according to new publications, revisions and changed taxonomic visions. This dynamic approach, also adopted by several other databases such as Euro+Med Plant Base and GRIN taxonomy, is widely appreciated by most taxonomists. On the other side, it has been also criticized by vegetation scientists since it can give rise to ambiguous references for taxa names in phytosociologic literature (Jansen & Dengler, 2010; Dengler et al., 2012). To avoid misunderstandings, we consider a good practice to cite web address and accession date as nomenclature reference, e.g. “www.anarchive.it. Taxonomic reference list [accessed on year-month-day]”. It is also possible to download a complete master list (as well as a selection of taxa) including the last modification date.

Such a taxonomic list directly associated to a vegetation database can be considered as a pioneer in vegetation science, also in the framework of the recently proposed European taxonomic Standard List (EuroSL - Dengler et al., 2012). Although not so refined as GermanSL, which has a hierarchical structure and considers all taxonomic concepts adopted in Germany (Japan...
sen & Dengler, 2008), VegItaly’s taxonomic list is not a mere index of names since it includes some essential attributes associated to each taxon and unequivocal links between taxa and synonyms. It should be stressed that most frequent mistakes, due to misapplication of names, should be necessarily treated case by case based on users’ personal expertise.

Data import and harmonization among different data banks

Import functionalities have been developed taking into account the two currently most used formats in vegetation plot archiving: TURBOVEG and spreadsheet (Landucci et al., 2012). To do so, import assistants have been developed ad hoc, named respectively VegImport and TabImport. The former has been specifically designed for TURBOVEG standard .xml format, while the latter allows the parsing and import of .odt, .xls or .xlsx files. Both have the classical structure of assistant applications, which guide the user step by step during data import. In case of repetitive header data, both the applications allow to introduce them only once, by using joint fields. Geographical coordinates can be introduced only when the reference system is also reported; they will be converted to WGS84 geodetic reference datum, the latest version of the World Geodetic System, defined and maintained by the United States National Geospatial-Intelligence Agency (NGA), which represents a standard in cartography.

A crucial issue is represented by the vegetation stratification, which can be expressed by using several different systems according to the user’s approach and the type of relevé. For instance, TURBOVEG adopts a 10-levels scale (0 to 9), where “0” corresponds to a unique, undefined layer (Hennekens & Schaminée, 2001). Similarly, in VegItaly the stratification system must be indicated during the upload stages, in order to build consistent comparative tables; there are no limits for the number of layers.

The process of data import is segmented in stages, during which the individual data are progressively sent to the server to be verified. A list of errors is then compiled by the system, which detects all the incidental inconsistencies, such as fields which are not automatically recognized. This type of problem can derive from formal discrepancies between imported data and preset formats (e.g. same content with different field name); in these cases the user can directly operate in order to obtain consistency. Both the import assistants work on an “all-or-nothing” approach: to avoid data loss, they import the whole set of data, if this is not possible no import is done at all. When a content to be imported is new, meaning that there is no corresponding preset format in the system, an integration is possible only in accordance with the developers, following a developing model close to incremental programming (Pressman, 2010), used both for the assistants and for the whole system.

Management of bibliographic references

Data deriving from bibliographic sources are provided with detailed information according to the LISY standard. This tool, developed in the 1990s, is a national bibliographic archive online, including all the syntaxonomic units of the Italian vegetation and a list of the related published sources (Bracco & Nola, 1995; Biondi et al., 1996, 1997; Bracco, 2001; Bracco et al., 2007). Currently it counts almost 3,400 bibliographic entries and more than 33,200 syntaxa. As concerns the syntaxonomic information, LISY is conceived to report data in their original form without any updating or critical review, even when they are notoriously obsolete or wrong. The inspiring principle, as for VegItaly’s taxonomic list, is to preserve the original documentary sources without introducing any bias and keeping a clear distinction between data storing and critical syntaxonomical revision (Bracco & Nola 1995).

Data stored in LISY include, for each bibliographic source, descriptive geographic information, bibliographic reference, all the mentioned syntaxa with authors and synonyms, and indications of the type of phytosociologic data. An hypertextual link between vegetation plot and the corresponding bibliographic/syntaxonomic LISY record is among VegItaly’s functionalities currently under development.

Data property and ethical treatment

The access regime to data stored in a data bank generally implies a serious debate on data property and ethical principles, as already pointed out by several Authors (Dittert et al., 2001; Salzberg et al., 2003; Arzberger et al., 2004; Baker & Bowker, 2007; Dengler et al. 2011). In fact, data ownership and authorship are generally values worthy of care in the scientific community. It should be considered that data sharing, especially for individual researchers, can be an unrewarded, time-consuming activity without any financial support or, at least, any tangible acknowledgment of its value in the scientific community (Arzberger et al., 2004). At the same time, restricted choices about data availability can affect spread of knowledge and seriously limit any intention of contributing to knowledge sharing. This issue becomes especially noticeable when different types of records are stored in an archive, including published, unpublished, private and public data.

Even if this is difficult to solve from a general/philosophical point of view, in practical terms this problem
Fig. 2 - Concept diagram representing VegItaly’s structure and metadata, re-drawn from Venanzoni et al. (2012) and updated on the ground of the most recent system developments. Dashed line: under construction; solid line: operative parts; the symbol “*” indicates additional tools, not planned in the beginning; the grey dashed line with grey text indicates collateral projects.
can be easily managed by developing separate access for data subsets, allowing the development of projects with differentiated access regimes. In order to respect both data contributor and data user rights, it is generally highly recommendable to draw up an agreement in which the involved parts share a number of basic rules about data management.

VegItaly data access regime is structured in 3 main types: a) visible, b) partly visible, c) invisible data. Type a) refers to data which are visible to all the users and web search engines, available for download without any special permission. Type b) includes data which are partially visible to users and search engines, whose use is subject to an explicit consent by data owners. This partial visibility generally provides a free access to a part of header data. Type c) includes data which are kept, for several reasons, invisible both for users and search engines. Data referred to types b) and c) can always be turned to a) and this is indeed the access type VegItaly firmly recommends, at least on the long term.

Open issues and future development

As reported in Landucci et al. (2012), the number of vegetation plots stored in VegItaly rapidly increased during the last two years, their geographic origin became larger and larger and the number of contributors started to grow significantly, indicating a clear will to convey energy towards the development of a shared archive for vegetation data, in the wake of other European countries (e.g. Chytrý & Rafajová, 2003; Schaminée et al., 2012a, 2012b; Apostolova et al., 2012) and in line with other national initiatives aiming to facilitate the access to biodiversity information (Martellos et al., 2011; Guarino et al., 2010) or offering a shared vision of large-scale topics (Biondi et al., 2012). Compared with the original scheme of VegItaly, synthesized in Venanzoni et al. (2012), many fundamental parts of its structure have been fully built up, although several utilities still have to be developed (Fig. 2). This is particularly true for management of synoptic table, exporting formats and, in part, GIS and Web cartographic tools.

Data syntaxonomic characterization is another important issue to address new efforts to. On one side, the development of a national syntaxonomic list is on the run and will be more robust if grounded on a large, national-scale data-set. On the other side, the syntaxonomic information management in VegItaly system currently lacks a hierarchical structure. More efforts will be spent in this direction, towards the development of tools able to manage a Syntaxonomic list based on a proper dictionary including all the hierarchic levels, typologic and synonymic details.

Conclusive remarks

As tools specifically designed for large-scale data sharing, vegetation databanks are increasingly showing new possibilities for vegetation science and phytosociology. With the introduction of some new implements able to solve well-known problems in vegetation data management, VegItaly stands as a rather innovative system. Its user-friendly interface and its structural plasticity make it easily accessible to a large number of scientists and may contribute to a broader view on vegetation issues. The described technical solutions currently adopted in VegItaly represent the achievement of several important goals and allow the scientific community to benefit from a number of free, useful services in data archiving and retrieving. In future, the users themselves might contribute towards the solution of open issues, by a constant user-developer mutual support.

Networking is the basis of knowledge sharing and management. It encourages scientists to interact, to share ideas and resources, to implement solutions in their work. Large-scale databases can therefore represent a further important step in the development and strengthening of the phytosociological approach, also in the light of the most recent methodological and conceptual improvements (Biondi, 2011; Blasi & Frondoni, 2011). Built up on the contribution of individual relevés made by different researchers, databanks may offer a robust tool for data validation, and represent a fruitful space where scientific relations and comparisons can be developed. Furthermore, phytosociological relevés can be considered also as important sources of biodiversity matrices (Feoli et al., 2011).

The large-scale data sharing is a basic tool in comparative sciences and represents a solid ground for syntaxonomic analysis, whose ecological interpretation can be strongly supported by the analysis of huge data sets. This approach to vegetation studies can lead to important, practical and ethical outcomes, and will certainly provide further advances for the phytosociological method. Each individual plot is a small brick - small, but essential in the bulk of knowledge of vegetation science. Brick after brick, the scientists who upload relevés into the virtually unlimited space of a shared database are contributing to build two precious elements: self-learning and emancipation from a private management of knowledge.

References


Check-list of the hornworts, liverworts and mosses of Italy. Bocconea, 22: 1-256.


Bianchi Regione Toscana.


Jansen F. & Dengler J., 2008. GermanSL - eine univer-