Uniformity and representation of taxonomic and other characters and semi-automatic extraction using computer tools

Renaud FORTUNER *

La Cure, 86420 Verrue, France

Presented at
Colloque pour l'Étude de la Biodiversité des Nématodes et des Helminthes, Paris, France, 17-19 septembre 2001

Summary – A uniform decomposition of morphological-anatomical characters is proposed, based on results from the GENISYS project. Each character is decomposed into a 'structure' (organs and organ parts), a basic property of this structure, and possible states or values of the character. This decomposition can be applied to any character without enforcing any limitations to data entry by biologists. It is shown that a tool ('Terminator') based on this decomposition can be used for semi-automatic extraction of characters from published descriptions or from new data entry. If built, such a tool can be used to populate a database with the decomposed characters. This database could be used with existing computer identification or systematics tools. The same approach can be applied to the decomposition of other kinds of data, including molecular and physiological data. This would create sets of interrelated databases housing various types of knowledge on biodiversity.

Keywords – character representation, data extraction, databases, identification tools, schemas, systematics tools.

For the general public, biodiversity means diversity of the species that exist on Earth, but in fact it is a multi-faceted concept that relates to many scientific fields such as systematics, of course, but also ecology, genetics, embryology, development, physiology, biochemistry, and many more. It can be argued that a good way to organise and link together all these aspects of biodiversity would be to use a system based on the morphological-anatomical description of the species.

Fig. 1 presents the hierarchy typically used, *e.g.*, in biology textbooks, to present the morphological-anatomical description of various biological groups. This hierarchy includes the major systems such as the nervous system or the genital system, the organs that are included in these systems, the tissues and cells that are parts of these organs, and the intracellular components of the cells, down to genes, DNA and bases. As indicated in Fig. 1, the various scientific fields of interest to biodiversity can be related to the various levels of this hierarchy, including fields such as embryology, development, or paleontology that can be arranged along an additional time dimension.

Our knowledge about morphology and anatomy and about the various other scientific fields included in Fig. 1 represents an enormous amount of facts, which must be carefully classified and stored in a way that supports easy retrieval. This is particularly true if the various experts that are interested in biodiversity want to be able to access data in fields with which they are not familiar.

Only computer science offers some hope to put this huge mass of knowledge in order and store it in such a way that the data of interest can be retrieved easily. However, when biologists turn to computer science, they discover that computer scientists have their own needs that may differ from the needs of biologists.

The main object of this article is to describe how biological data can be represented so they can be used by computer scientists. It is based on the work done for NEMISYS (Nematode identification system), a project that later was enlarged to GENISYS (General identification system). The NEMISYS team was created in 1987 in California by the present author and two computer scientists from the University of California at Davis, Jim Diederich and Jack Milton. GENISYS is described on the Web at: http://www.math.ucdavis.edu/~milton/genisys.html, soon to be moved to http://www.genisys.prd.fr/genisys_home.html. Compared to existing computerised identification aids, NEMISYS/GENISYS is unique in that it does not rely on a single approach to identification but wants to include them all, from deterministic approaches such as multiple entry keys to probabilistic approaches based on

* E-mail: fortuner@wanadoo.fr

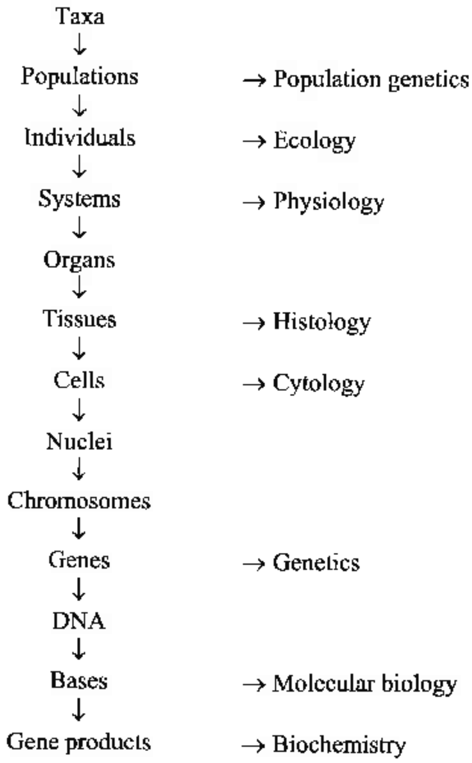


Fig. 1. *The classical hierarchy of morphological-anatomical characters arranged in major physiological systems with related fields of knowledge.*

similarity coefficients, and to approaches based on individual specimen data such as Discriminant Function Analyses (DFA) and other statistical methods.

A consequence of this philosophy was to reject the traditional practice of selecting a limited number of identification characters. This is because the character set that can be used for a particular approach is not necessarily the same as that needed for a different approach. In particular, it is impossible to use a limited set of characters to do both identification and systematics because, in any taxon, identification characters are not always the same as systematic characters. For this reason, the GENISYS database will be as comprehensive as possible, *i.e.*, it will include all possible characteristics in a biological group such as the tylenchids or the rhabditids.

Our attempts to create the schema of such a comprehensive database quickly ran into the problem of data uniformity. (In this article, schema is used in the restrictive sense of a list of taxonomic characters arranged according to a certain format, as described below.)

Representativity *vs* uniformity

UNIFORMITY

When computer scientists are asked to process data, their first demand is that these data are reasonably uniform, which means that similar data must always be represented in a similar way. This is necessary to control the complexity of computer applications, and is particularly important in biology because biological data are far more complex than, *e.g.*, business data. Data uniformity is also useful to biologists when they have to process more than a few characters. In fact, in the GENISYS project, our work on data uniformity originated from a request I made as the team biologist. The computer scientists had asked me to create a data matrix for them to use to test a few ideas. A first list was created with 130 characters. This number of characters was rather low, but it was enough to create problems with controlling data uniformity. I found it very difficult to keep track of the various ways used by different authors to describe, *e.g.*, shapes. For example, the same organ will be described in the literature as round, rounded, spheroid, spherical, or circular.

For a uniform database it is necessary to select one of these terms and to always use it to describe this type of shape in the organ in question but also in other organs that exhibit a similar form. This is a very difficult problem because there are far more than 130 characters and a comprehensive schema could include several thousand characters.

REPRESENTATIVITY

In theory, this problem can be solved by enforcing the use of one term for each shape in the database, *e.g.*, spheroidal in the above example. However, it is not always possible to reach a consensus on the term to use. Also, even if were possible to force currently active nematologists to agree on a list of terms (in itself a daunting task!), we could not retrospectively impose our selections on past nematologists. This is an important point as no study of biodiversity can afford to ignore previously published data because:

- some descriptions are irreplaceable (*e.g.*, due to destruction of the type locality of a species);
- it would be far too expensive to describe afresh all known species;
- overall quality of new descriptions might not reach the level of the works of past great nematologists.

A first way to use published descriptions would be to store the articles as published, after they are put into an electronic format. Characters would then be extracted from the stored texts as and when they are needed by successive users. Such an approach would raise several technical problems, in addition to legal problems concerning copyright. It is far better to extract the characters from the published texts and store them as individual characters in a database. In the database, the characters need to be in a format that is both uniform to satisfy computer scientists and representative to please biologists.

To meet these apparently contradictory requirements, we started from the classical decomposition in Entity/Attribute/Value. This was, in fact, the same decomposition as that used by Lebbe (1991). Then, we further detailed this representation and we defined:

- the entity as restricted to biological structures only, *i.e.*, the organs, grouped into the major classical systems, and the organ parts such as tissues, cells, and cell components;
- the attribute as any property that can be used to described a structure as defined above;
- the value as a qualitative state or a quantitative value of the character in a taxon, a population or an individual.

After we created a first list where the characters were represented by separate structures and properties, Diederich (1997) observed that most or all properties in the list belonged to a short list of what he called basic properties (Fig. 2). These basic properties have been classified into four major categories in Fig. 2. The list also includes 'presence', which is not really a property but which is often used as one in species descriptions.

Because of the fruitful interactions that existed in the GENISYS team between biologist and computer scientists, we quickly discovered that this decomposition alone was unable to maintain uniformity. Most difficulties stem from the fact that biologists tend to include all kinds of information in the name of characters. The first list of characters included, *e.g.*:

Structure:	Tail end indentation
Basic property:	Kind
States:	Shallow depression
	Depression
	Notch
	Indentation
	Groove

APPEARANCE		DIMENSION
posture		length
shape		height
kind		width
texture		diameter
arrangement		depth
symmetry		ratio of * to *
		size
PLACEMENT/LOCATION		QUANTITY
position relative to *	presence	
distance to *		quantity
orientation		number
angle		

Fig. 2. *List of basic properties for morphological-anatomical characters (Diederich, 1997).*

This character* includes information on the presence of an indentation at the tip of the tail, but also information on the shape of this indentation, from a depression to a notch, and on its size. Obviously, such a character is far from being decomposed into its most basic elements. If biologists were allowed to use such complex characters, they would be unable to maintain uniformity.

To enforce uniformity, we worked in two directions: *i*) we defined rules and concepts for better uniformity; *ii*) we designed a computer tool to apply these rules and concepts. This tool was defined as including three types of functions: *i*) accept any data, either as extracted from the literature or as entered by an active user; *ii*) decompose these data according to the rules and concepts we defined; *iii*) reassemble the stored decomposed elements into complex characters so as to present each user with the characters needed in a selected format.

This approach guarantees the freedom of biologists who can enter and extract data without any constraint, while storing these data in a uniform manner. One way freedom is maintained is through the addition of synonyms. Each structure in the schema has a preferred name and a list of synonym names.

GENISYS rules and concepts

The rules and concepts for better uniformity were designed in a very practical manner in the course of the NEMISYS/GENISYS projects, as each one originates from the discovery by the team computer scientists of a problem with some part of the schema as proposed by the biologist. A discussion generally followed that would be at times heated, always lengthy, until a solution was found that was acceptable by computer scientists and biologist alike. It was then defined as a new rule or a new concept by the computer scientists.

Among the 17 articles published on NEMISYS/GENISYS (see complete list at the end of this article), several described the rules and concepts for better uniformity:

– Diederich *et al.* (1989) list 12 types of metadata (*i.e.*, data about the data) that relate to each character or to the state or value of the characters. The metadata include, among others, the type of character (qualitative, integer, ordered, *etc.*), the unit of quantitative data, and fuzzy attributes such as 'rarely' or 'often' that are used to describe probabilistic data.

– Diederich (1997) proposed the concept of basic properties and many other concepts such as name extensions, implicit properties, general *vs* specific states, state-based relationships, dependant *vs* summary characters, redundant characters, fuzzy characters, *etc.*

– Diederich *et al.* (1997) gave a list of rules for creating schemas, including rules for the use of name extensions.

– Diederich *et al.* (2000) described various difficulties encountered during the creation of the first schema based on previously defined rules. Part of this schema was included in the article as a list of structures to which can be applied the basic properties defined by Diederich (1997).

Some problems remain to be solved and the GENISYS team continues its schema cleaning task.

Extraction of published data

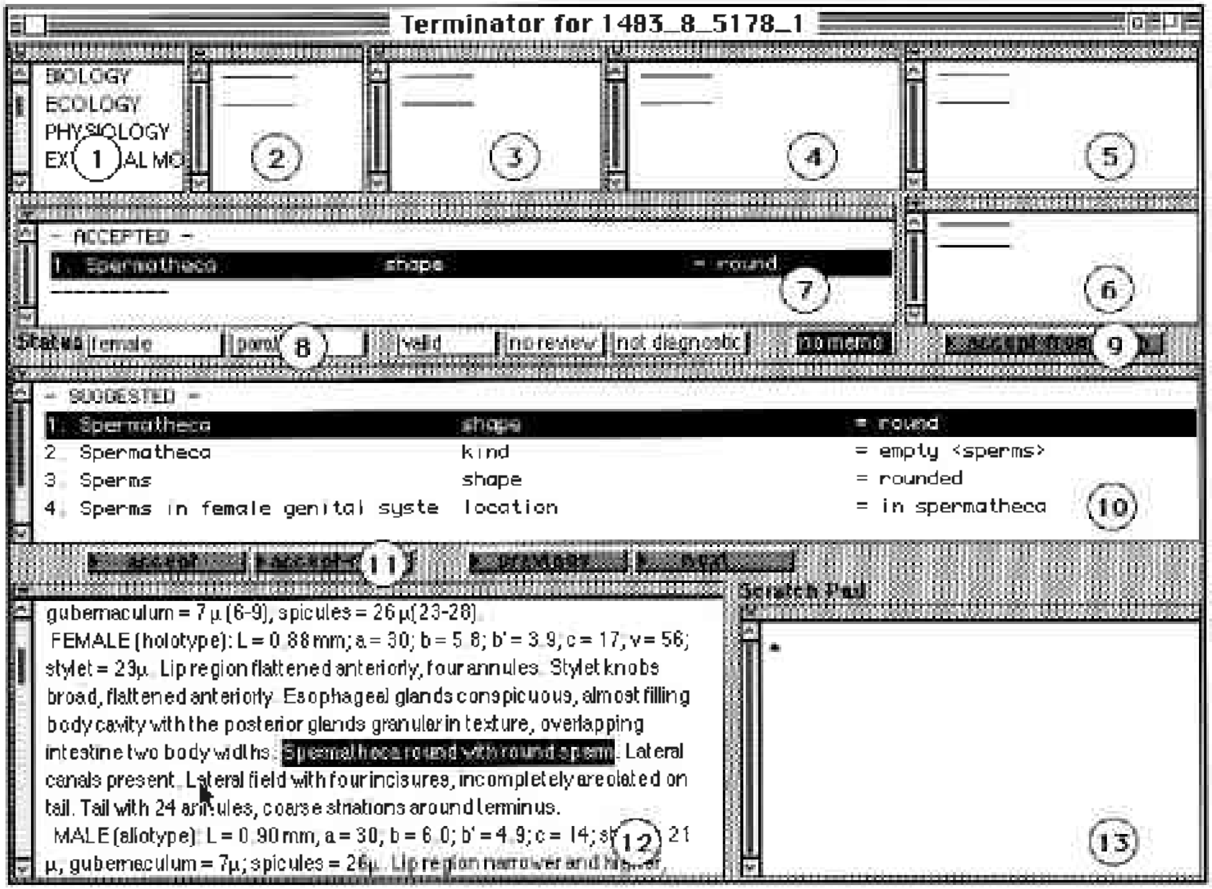
THE VARIOUS APPROACHES TO DATA EXTRACTION

When a first schema is complete or reasonably so, the morphological-anatomical database will have to be populated with data from published articles. There are several ways to do this, including manual data extraction, natural language recognition, and keyword-based approaches.

We immediately rejected manual data extraction and entry as being too long, too difficult and too error prone. It would not assure data uniformity as it is too difficult for a human operator to remember and properly apply all the rules and concepts we defined. In contrast, we could have made the computer able to understand the texts fed into it so the machine itself could be entrusted with data extraction and entry. This would have made it necessary for the project to enter into the field of natural language recognition. This would have required the creation of what computer scientists call a lexicon, which is basically a table housing all the terms in the domain together with the semantics and other properties needed by the computer to understand each term. Creating a lexicon is a very demanding task and a lexicon needs to be continuously updated as new terms are added to the schema. We decided that we did not have the resources in manpower and money to do it and we abandoned this option.

The intermediate option we selected was based on the schema, which is used as a list of keywords. Keywords alone are poorly understood by computers when they need

[*] The word 'character' is used in the sense of 'a characteristic that can be used to differentiate two objects', including or not the list of possible states of the character in a biological group.



- ACCEPTED -

1. Spermatheca shape = round

- SUGGESTED -

1. Spermatheca shape = round
2. Spermatheca kind = empty <sperms>
3. Sperms shape = rounded
4. Sperms in female genital syste location = in spermatheca

gubernaculum = 7 μ (6-9), spicules = 26 μ(23-28).
FEMALE (holotype): L = 0.88 mm; a = 30; b = 5.8; b' = 3.9; c = 17; v= 56; stylet = 23μ. Lip region flattened anteriorly, four annules. Stylet knobs broad, flattened anteriorly. Esophageal glands conspicuous, almost filling body cavity with the posterior glands granular in texture, overlapping intestine two body widths. Spermatheca round with round sperm. Lateral canals present. Lateral field with four incisures, incompletely areolated on tail. Tail with 24 annules, coarse striations around terminus.
MALE (allotype): L = 0.90 mm, a = 30; b = 6.0; b' = 4.9; c = 14; st 21 μ, gubernaculum = 7μ; spicules = 26μ. Lip region narrower and higher

Fig. 3. *The interface of the Terminator prototype of 1993 (from Diederich* et al., *1999).*

to distinguish between, *e.g.*, the ratio 'a' and the indefinite article 'a'. Without a lexicon, such a distinction is nearly impossible. In our solution, we used a human operator to complement the keywords.

THE TERMINATOR

Broadly speaking, the system uses keywords from the schema and some search algorithms to propose one or several possible decompositions for each successive character in the text. The human operator selects the correct decomposition and the computer completes the treatment of the character by formatting and storing it in a text file with delimiters for future loading into a database.

For this approach, the operator must rely on a tool with a well designed interface and we put a lot of work into the design of Terminator, as we called this tool, that relies on terms from the schema to recognise the data we are interested in.

Due to lack of funding, we have not been able to build this tool and it does not exist at this moment. However, we were able in 1993 to build a prototype that we used to demonstrate the feasibility and usefulness of the concept. This demonstration was described in a virtual publication by Diederich *et al.* (1999).

This is not the place to describe in detail the interface of the Terminator prototype, which does not exist any more. I only need to say that the operator saw the text of the original article in pane 12 (Fig. 3) where each character was highlighted in succession. In pane 10, the tool made one or several suggestions for the decomposition of the character currently highlighted. If the operator decided that one of these suggestions was the correct one, he clicked on button 11 and the character, correctly decomposed according to the GENISYS format, was added to a file, ready to be loaded into a database.

Table 1. *Processing 12 nematode descriptions using the Terminator prototype (from Diederich* et al., *1999a).*

Processing time (min)	Number of characters	Number of characters/min	Number of schema changes
27	84	3.1	1
58	89	1.5	5
24	56	2.3	5
21	47	2.2	4
38	88	2.3	6
23	88	3.8	5
31	90	2.9	7
25	78	3.1	5
19	79	4.2	3
14	76	5.4	2
22	58	2.6	11
14	52	3.7	5
26.3*	73.75*	3.09*	4.9*
(12.00)**	(15.98)**	(1.06)**	(2.5)**

* mean;
** standard variation.

If the operator decided that none of the suggestions from the tool were correct, he had the option of using panes 1 to 6 to navigate through the schema to the correct character. If the character was absent from the schema, the operator used another tool, the Schema tool, to add it to the schema.

The Terminator prototype was tested in 1993 with 12 descriptions. The results of this test are shown in Table 1. On average, the descriptions were 1.3 pages long and included 73 characters. Using the Terminator prototype, I was able to process three characters per minute, including time for updating the schema. This means that the average treatment time per description was a little over 26 min. The success, *i.e.*, the percentage of correctly processed characters, was 100% since I was able to process manually any character incorrectly recognised by the tool.

It must be noted that these results were obtained with a prototype and a schema that were far from being perfect. The schema had to be modified on average five times per description because this test was conducted before our major drive to achieve schema uniformity. Obviously, a new tool using more efficient search algorithms, a more user-friendly interface, and a more uniform schema would give far better results.

But, in spite of its limitations, the 1993 prototype was good enough to demonstrate that the concept is feasible and that Terminator, if built, can make it possible to create a comprehensive morphological-anatomical database, *i.e.*, one including all the characters described for all known species in a given group.

Using the GENISYS database

The GENISYS tools

Originally, the GENISYS database was to be used with special identification tools we intended to develop within the project. As stated above, GENISYS was intended to be a set of tools, each tool designed to help the user with one of several identification tasks during an identification session.

For example, an identification session can start with instant recognition of what I called a promorph (Fortuner, 1989), which is basically a form that can be recognised at first glance. Then an elimination tool can rely on a dichotomous key or a multiple entry key approach to get rid of all the species in the promorph that are obviously different from the specimen, based on their primary identification characters, *i.e.*, characters for which there is a very low risk of error in the species in question (Fortuner, 1989). Then the user may decide, *e.g.*, to compare the remaining species with the specimen using a similarity tool. Finally, a verification tool can be used to check whether the most similar species is actually the one to which the specimen belongs.

Of course, each identification session is unique and the user would have been free to select any tool in any order. The various tools would have been integrated and the results obtained with one tool would have been available with all the other tools. However, no GENISYS identification tool was ever built because we never managed to obtain the necessary funding. One of the reasons is that building computer tools requires a level of funding far above the typical costs of biology projects (see Appendix).

Existing identification tools created by other authors

No GENISYS tool exists but a GENISYS database can be used with identification tools developed by other authors. For example, botanists often rely on a system called DELTA (Descriptive Language for Taxonomy) developed by Dallwith (1980). In this system, selected characters are represented by a standardised code and generic tools use the resulting code for species identification. To use

DELTA tools with data from a GENISYS database, these data can be extracted from the database and transformed into DELTA codes. The rules for writing DELTA codes are fairly simple and any GENISYS database would be uniform by design, which means that it should not be too difficult to develop an automatic DELTA code writing function for user-selected characters. It might also be possible to use what computer scientists call a 'view', which is a kind of virtual representation. The GENISYS characters would remain as they are in the database but the user would 'see' them in the guise of DELTA codes or in any other form required by other tools.

ALPHA TAXONOMY AND SYSTEMATICS

In addition to identification with existing identification tools, a GENISYS database could be used for alpha-taxonomy studies. For example, a user would enter the description of some specimens into the base and check, using existing identification tools. whether they belong to a known or a new species. If they belong to a new species, the description of a new species in a journal would then be limited to a name and a few lines of diagnosis with reference to the actual description stored in the GENISYS database. This database would then function as an electronic journal accessible *via* Internet. Taxonomists would no longer have to spent an inordinate amount of time doing alpha-taxonomy and they would be free to concentrate on the kind of fundamental studies that are well accepted by journals with high impact factor. This would remove one of the stumbling blocks of hiring new taxonomists as described by Hugot (2002) during the meeting.

High level systematic studies also can use the data from a GENISYS database since it would be possible to set up an export function for reformatting the GENISYS data into a matrix format that could be loaded into cladistic tools such as PHYLLIP, PAUP, HENNIG86 or MacClade. As the database would include all possible characters and not only identification characters, it would include also systematic characters.

MOLECULAR, PHYSIOLOGICAL, AND OTHER DATA

Some molecular biologists argue that traditional identification is outdated and that identification (and systematics) can now be based on a molecular 'bar code' system described from ribosomal cistrons. Even if this approach were acceptable and feasible in practice for all existing species, a GENISYS database would still be useful because biodiversity is far more than the diversity of genetic sequences.

One of the lowest levels of the structure hierarchy in the GENISYS schema is that of DNA bases (Fig. 1). If the data structure at this level (properties and state/values) is the same as that of existing molecular databases with sequence data, then it will be easy to link these bases to GENISYS bases. This will make it possible to express in a computerised manner the relations that exist between a particular sequence, the gene that includes this sequence, the type of cell where this gene is expressed, the product of this expression and the physiological function of this product. Obviously, this function could then be linked to a physiological database.

I suggested in the Introduction that a morphological-anatomical database can be used to better organise the various kinds of knowledge relative to biodiversity. I have to admit that I do not know how a physiological database schema can be designed because I am not a specialist in nematode physiology but it should be possible to:

- apply to physiological and other kinds of data the classical decomposition of data into entity/attribute/value;
- define entity as one of the morphological-anatomical structures of the GENISYS database;
- define attributes as basic properties that would be different from the morphological-anatomical basic properties but that would be defined based on the same approach.

I will not attempt to go beyond these very general principles. For the actual definition of physiological (and other) data according to the principles above, the GENISYS team needs to enlarge and welcome specialists in the various fields of knowledge involved. These experts would define what data they are using and what is the current format of any existing database. This would make it possible to match the formats of these various databases.

Conclusion

In the present article, using past work by the GENISYS team, I tried to show that it is possible to define uniform data that can be used by computer scientists while responding to biologists' demands for representativity and freedom. I also tried to show that a uniform format cannot be defined from theoretical considerations alone and that many problems related to data decomposition can be uncovered only during the creation of a full-size schema. Biological data are so rich and complex that a usable solution can be proposed only when problems are discovered and solved by using theoretical concepts in real life applications. This means that only a team with both computer scientists and biologists can reach a workable data decomposition solution.

I believe that the concepts defined for morphological-anatomical data also apply to data from other fields of knowledge. A morphological-anatomical database can be used to organise other types of biological data, including all the kinds of data that describe biodiversity.

On the practical side, the NEMISYS/GENISYS projects demonstrated that a data extraction tool such as Terminator can be created and used for actual semi-automatic extraction and formatting of new or published data. The Terminator prototype demonstrated the feasibility of the concept in 1993. However, creating the final tools would be very expensive compared to the usual level of financing in systematics. Only a joint action by many labs would have a chance to obtain the necessary funds.

Finally, the GENISYS team needs to accept new members from other scientific fields, or to collaborate with such experts to enlarge the scope of the project to include other topics: links with molecular data, schema definition for other types of data, *etc*. The '*Réseau pour l'étude de la biodiversité des nématodes et des helminthes*' can play a major role in the future developments of the GENISYS project.

References

CITATIONS IN THE PRESENT ARTICLE

DALLWITZ, M.J. (1980). A general system for coding taxonomic descriptions. *Taxon* 29, 41-46.

DIEDERICH, J. (1997). Basic properties for biological databases: character development and support. *Journal of Mathematical and Computer Modelling* 25, 109-127.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (1989). Building a knowledge base for plant-parasitic nematodes: description and specification of metadata. In: Fortuner, R. (Ed.). *Nematode identification and expert-system technology.* New York, NY, USA, Plenum Publishing Corp., pp. 65-76.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (1997). Construction and integration of large character sets for nematode morpho-anatomical data. *Fundamental and Applied Nematology* 20, 409-424.

FORTUNER, R. (1989). A new description of the process of identification of plant-parasitic nematode genera. In: Fortuner, R. (Ed.). *Nematode identification and expert-system technology.* New York, NY, USA, Plenum Publishing Corp., pp. 35-44.

HUGOT, J.P. (2002). Proposal for a network devoted to the study of nematology and helminthology. *Nematology* 4, 563-565.

LEBBE, J. (1991). *Représentation des concepts en biologie et en médecine. Introduction à l'analyse des connaissances et à l'identification assistée par ordinateur.* Thèse de doctorat. Paris, France, Université Pierre et Marie Curie, Paris, 282 pp.

OTHER PUBLICATIONS ON THE NEMASYS/GENISYS PROJECT (IN CHRONOLOGICAL ORDER):

FORTUNER, R. (1989). *Nematode identification and expert-system technology.* New York, NY, USA, Plenum Publishing Corp., 386 pp.

DIEDERICH, J. & MILTON, J. (1989). NEMISYS, an expert system for nematode identification. In: Fortuner, R. (Ed.). *Nematode identification and expert-system technology.* New York, NY, USA, Plenum Publishing Corp., pp. 45-63.

DIEDERICH, J. & MILTON, J. (1991). Creating domain specific metadata for scientific data and knowledge bases. *Institute of Electrical and Electronic Engineers Transactions on Knowledge and Data Engineering* 3, 421-434.

FORTUNER, R. (Ed.) (1993). *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, MD, USA, The Johns Hopkins University Press, 560 pp.

FORTUNER, R. (1993). The NEMISYS solution to problems in nematode identification. In: Fortuner, R. (Ed.). *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, MD, USA, The Johns Hopkins University Press, pp. 137-164.

DIEDERICH, J. & MILTON, J. (1993a). Expert workstations: a toolbased approach. In: Fortuner, R. (Ed.). *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, MD. USA, The Johns Hopkins University Press, pp. 103-123.

DIEDERICH, J. & MILTON, J. (1993b). NEMISYS: an expert workstation, a computer science perspective. In: Fortuner, R. (Ed.). (1993). *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision.* Baltimore, MD, USA, The Johns Hopkins University Press, pp. 165-179.

DIEDERICH, J. & FORTUNER, R. (1996). Endorsement of observations in identification. *Institute of Electrical and Electronic Engineers International Conference on Fuzzy Systems, 8-11 September 1996, New Orleans, LO, USA,* pp. 175-179.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (1998). A general structure for biological databases. In: Bridge, P., Jeffries, P., Morse, D.R. & Scott, P.R. (Eds). *Information technology, plant pathology and biodiversity.* Wallingford, UK, CAB International, pp. 47-58.

DIEDERICH, J. & FORTUNER, R. (1998). Classification using small fuzzy biological data sets. *Institute of Electrical and Electronic Engineers International Conference on Fuzzy Systems, 4-9 May 1998, Anchorage, AL, USA*, pp. 1429-1434.

DIEDERICH, J. & FORTUNER, R. (1999). A fuzzy classifier using genetic algorithms for biological data. *18th International Conference of the North American Fuzzy Information Processing Society, 10-12 June 1999, New York, NY, USA*, pp. 680-684.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (1999). Computer-assisted data extraction from the taxonomical literature. Virtual publication URL, http://www.math.ucdavis.edu/~milton/genisys.html.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (1999). Concepts and approach for a General Identification System. In: Ryss, A.Y. & Smirnov, I.S. (Eds). *Information retrieval systems in biodiversity research. Proceedings of the Zoological Institute Russian Academy of Sciences* 278, 75-76.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (2000). GENISYS and computer-assisted identification of nematodes. *Nematology* 2, 17-30.

DIEDERICH, J., FORTUNER, R. & MILTON, J. (2000). A uniform representation for the plan of organisation of nematodes of the order Tylenchida. *Nematology* 2, 805-822.

Appendix

It was mentioned above that the cost of building computer tools requires a level of funding far above the typical costs of biological projects. According to an estimate from a small French company specialising in the development of scientific tools, *i.e.*, with a price list adapted to this low-budget market, the development of one tool such as Terminator or the Schema tool would cost about €25 000, to which must be added about €7500 for a one-time pre-development analysis, about €15 000 for the licences for development software and hardware, €7500 for purchase of a machine, and €7500 for hosting the project during the development phase (18 months). The total estimate for three tools (Terminator, Schema tool, and one identification tool) was about €112 500, before tax. This does not include funding for buying some time off for the GENISYS team members so they can work with the development company during the creation of the tools.