KNOWLEDGE MODELLING FOR A WOOD IDENTIFICATION SYSTEM

K. VANDER MIJNSBRUGGE* & H. BEECKMAN**

* Laboratory of Genetics, University of Ghent, Ledeganckstraat 35, B - 9000 Gent

** Section of Agriculture and Forestry Economics, Royal Museum for Central Africa, Leuvensesteenweg 13, B - 3080 Tervuren

1. INTRODUCTION

Because science and philosophy were split from each other since Immanuel Kant, the phenomenon of an all-round scientist, possessing all the knowledge and experience in all existing scientific disciplines, became historical. Anyhow, the growing importance of ecology gives nowadays rise to a strong demand for multi- and interdisciplinary approach and to a wakening up of scientists working in a too narrowed field of research. Whenever reached higher levels of system integration, the emphasis on concepts of time and complexity will be decisive for the modern scientific enterprise. To tackle these kind of problems, appropriate tools should be developed. These techniques should always contribute to a scientific service to rationality and not to a replacing of rationality by computer modelling. In forestry for example efforts could be made to look for suitable theories, methods and techniques with the help of rapid growing computer technologies. However, foresters should be aware not to try to displace every judgement by computer models. Human expertise will stay important. This is also true even on the lower levels of forest ecosystem integration: the cells and tissues.

The wood collection of the Royal Museum of Central Africa possesses more than 12000 different species of wood and to depict the ecological, taxonomic or regional variability within species, more than 52000 wood samples. Given such a vast number of species of woods, identification of an unknown wood sample can be a difficult task for wood anatomists, especially when the geographic source of the wood is not known. Many species have already been described in literature, but it is impractical to survey it all, so dichotomous or multiple entry identification keys are used. Computer-aided wood identification has many advantages over mechanical sorting or traditional identification keys (Wheeler *et al.*, 1986). Speed is an obvious advantage, but the most important advantage is flexibility. A specified number of mismatches in feature descriptors between an unknown sample and taxa in the database can be allowed, the presence or absence of certain features can be "required" of any proposed match (Wheeler *et al.*, 1986).
Several attempts to develop computer-aided wood identification systems have already been made. The most powerful programma is GUESS (General Unknown Entry and Search System), developed at the North Carolina Agricultural Research

Service of the North Carolina State University (Wheeler *et al.*, 1986). This program makes use of several databases, but no database exists that incorporates as much as the number of species of the Tervuren wood collection.
The subject of this study is the conceptual modelling process of a wood identification system. This stage of development of an identification package is essential as a first step.

Knowledge structures are analyzed on an epistemological level, independent of the future programming tool in which this model of a wood identification process could be implemented. The idea is that during the development of a computer system this conceptual modelling step should store all knowledge about the subject, without taking into account the structures and logic of the programming tool to be used. This paper wants to show that the conceptual modelling of a problem is a very important phase often omitted. The resulting model should describe the problem as a whole in clear and unambiguous descriptions, so that it can easily be interpreted by anyone unfamiliar with the project. Many authors wrote about the need of a modelling methodology (e.g. Breuker & Wielinga, 1988; Lenferink, 1991).

The explicitation of the precise objectives for a certain project, which are the reasons why a solution for the stated problem is desirable and/or necessary, is an important step towards a successful project development. The main purpose in elaborating the wood identification process is the search for relevant techniques and formalisms for modelling knowledge at a conceptual level. From this perspective the attention is given to the possible ways in which the identification problem can be represented, which difficulties can occur and which solutions can be considered.

2. THE MODELLING PROCESS OF A WOOD IDENTIFICATION SYSTEM

When solving a complex problem in a certain domain, one should be able to understand the principles in the domain and the typical language used in that particular domain. For the wood identification on a microscopic level it is essential to know the differences between the four major types of cells wood construction is based on: wood fibres, tracheids, vessels, and wood parenchyme. The presence of these types and the way they are organized within the wood structures are crucial for the determination of the exact wood species.

The knowledge model, which is the result of a knowledge modelling, can be regarded as a black box operating within a well defined environment, using a certain input and producing a certain output, dependant on predefined conditions inside and on external biasing influences (fig 1). In the wood identification system the black box can be supposed to be able to create a (heuristic) relation between the input, that consists of thin wood sections in radial, tangential and transversal

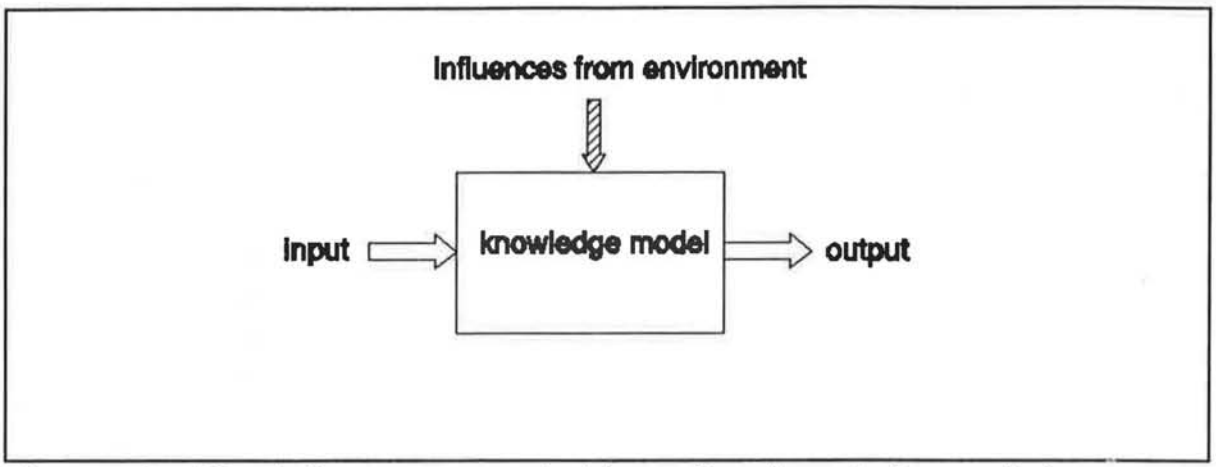


Figure 1. Black box representation of a knowledge model

direction, and the output, which is a wood species. It is important to situate this black box within its environment, to locate the place where it fits in the processes of the domain, and to identify the possible influences from this environment (fig 2). The specific purposes of the project will play a key role in this step. The possible external influences must be kept in mind while actually modelling the specific problem (the inside of the black box). This way of looking at the tackled problem should make it possible to delineate what knowledge should be enclosed in the black box, and what knowledge should not.

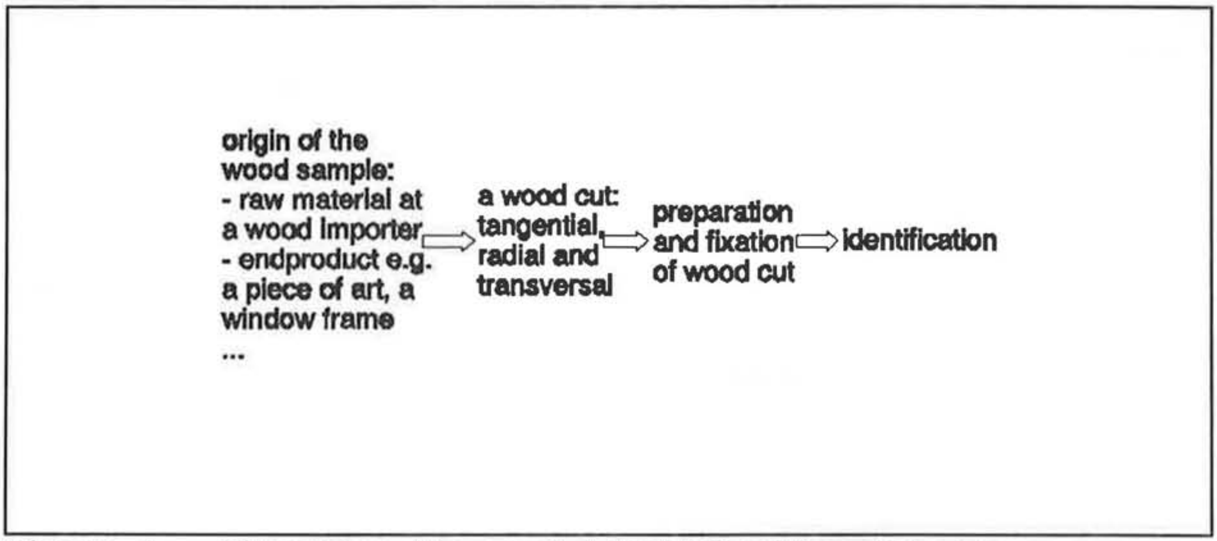

origin of the wood sample:
- raw material at a wood importer
- endproduct e.g. a piece of art, a window frame
...
a wood cut: tangential, radial and transversal
preparation and fixation of wood cut
identification

Figure 2. Situating the wood identification process.

Another aspect of the knowledge model that should be considered beforehand is the description language that will be used. The user of the system has to understand the language the system uses, so as to be able to communicate with the

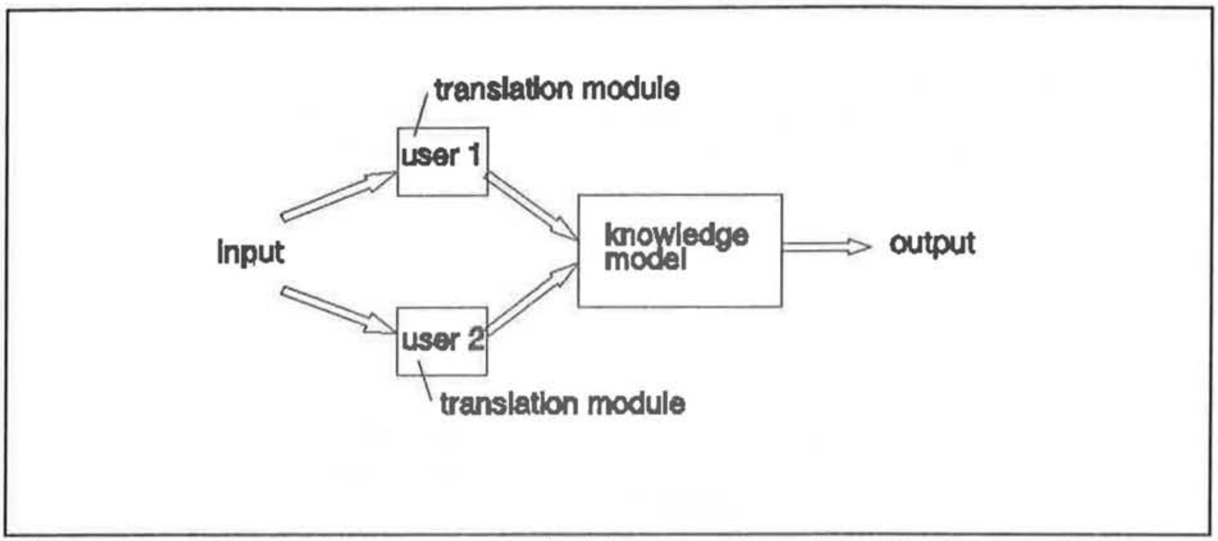


Figure 3. Black box representation of the knowledge model containing two translation modules.

system. One can think of an ideal system using domain specific terminologies and that consists of one or more translation modules dependant on different user models to guarantee an efficient man-machine dialogue (fig 3). The unfamiliar user can for instance switch to the translation module containing elaborate explanations to guide him/her through the system.

2.1. Choosing a representation paradigm

We know that our problem is an identification problem. The difficulty is to determine the representation of this problem closest to reality. We selected the formalism of a decision tree. The static knowledge is represented by a hierarchical classification (Chandrasekaran, 1988). A scheme of the overall representation of the decision tree is shown in figure 4. Every step through the decision tree is guided by a query for specific knowledge necessary to make a distinction between different possibilities. By means of a question-answer (machine-user) dialogue, one can reach a solution.

Control information includes all features with distinguishing characteristics, but not as strong as the feature(s) used in the questionnaire.

Although it is not our intention to describe user modelling one should realize that it influences the knowledge model. The user model defines the form of the man-machine dialogue, the knowledge model defines the content. The two disciplines have a lot of interaction, as will be indicated further on.

2.1.1. Hierarchical classification

How is a hierarchical classification constructed ? In a standard hierarchical classifi cation the general categories are situated in the upper parts of the tree. The more specific cases are situated near the branches. Wood for instance can be divided into hardwood (from deciduous trees) and softwood (from coniferous trees). These categories can be split up in further groups until the botanic species are reached (fig 5). This can be a typical biologic classification. Features should be isolated, which contain separating information. For instance, when studying the domain of wood anatomy, one can isolate very quickly the following elements: wood, hardwood, softwood and vessel. The static representation is as follows: 'wood can be divided into softwood and hardwood'. The search strategy which will be placed above is the following: 'if vessels are present in the wood, it is hardwood'. When one uses a standard biological hierarchical classification, difficulties might appear when trying to find a suitable search strategy. It seems more advisable to build up the two (hierarchical classification and search strategy) simultaneously.

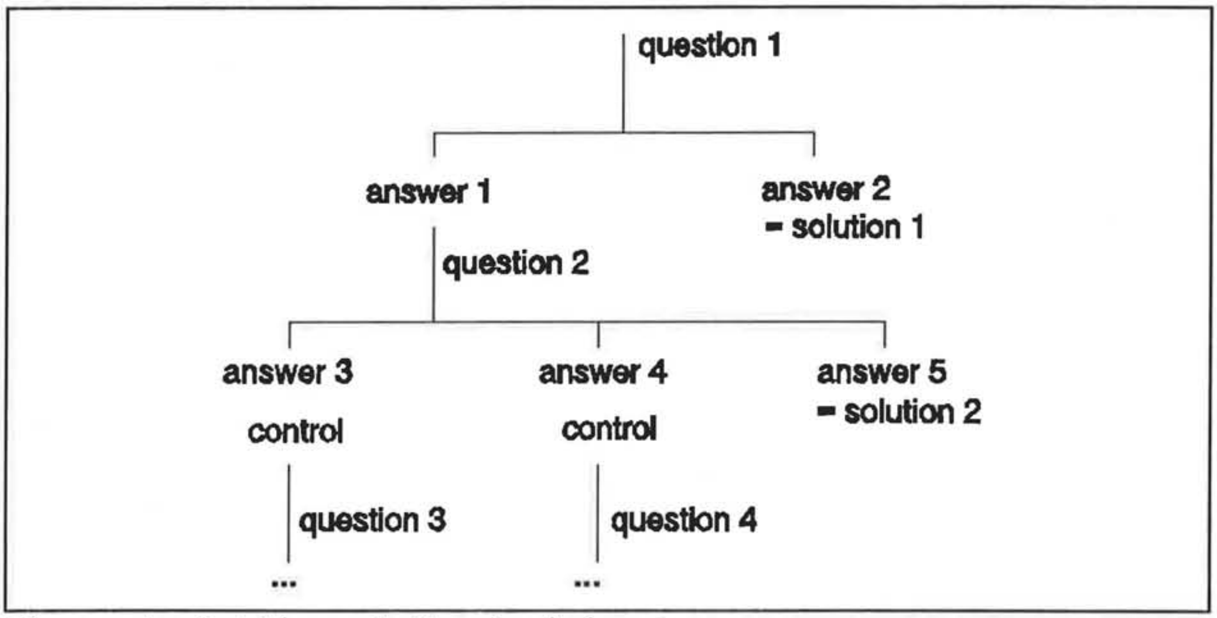


Figure 4. Outline of the decision tree.

Another point which is interesting to consider is the delineation of all possible solutions. Studying all possible features of these solutions might give interesting insights in the way the hierarchical classification can be build up. This delineation is also one of the answers to the question which knowledge the model should contain, and which not.

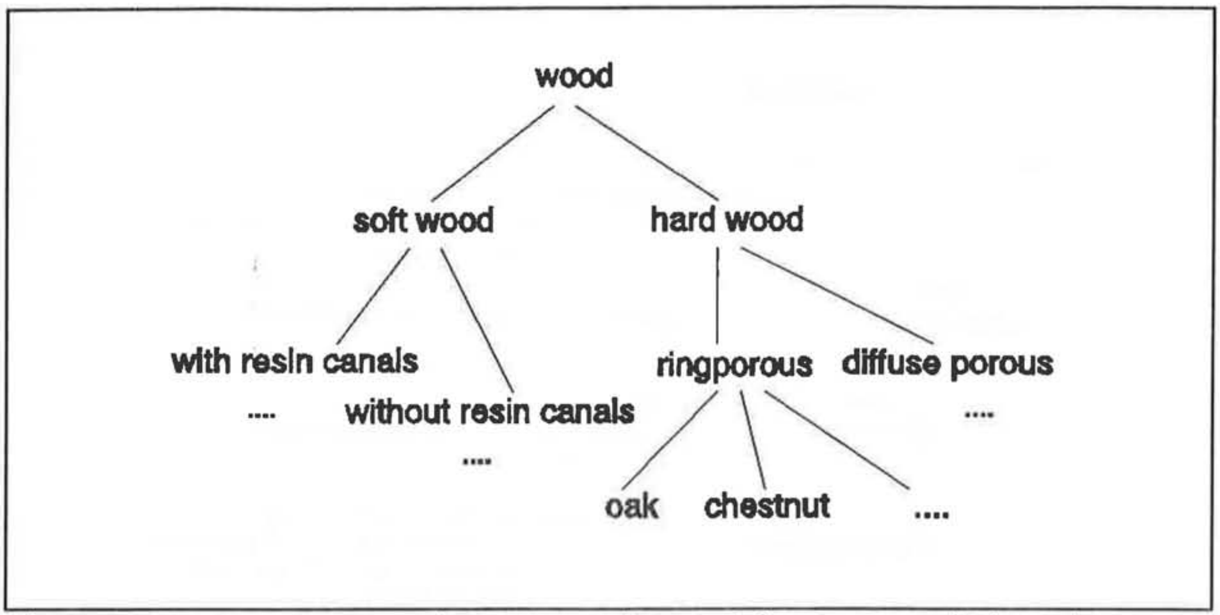


Figure 5. Representation of a hierarchical classification.

2.1.2. Search strategy

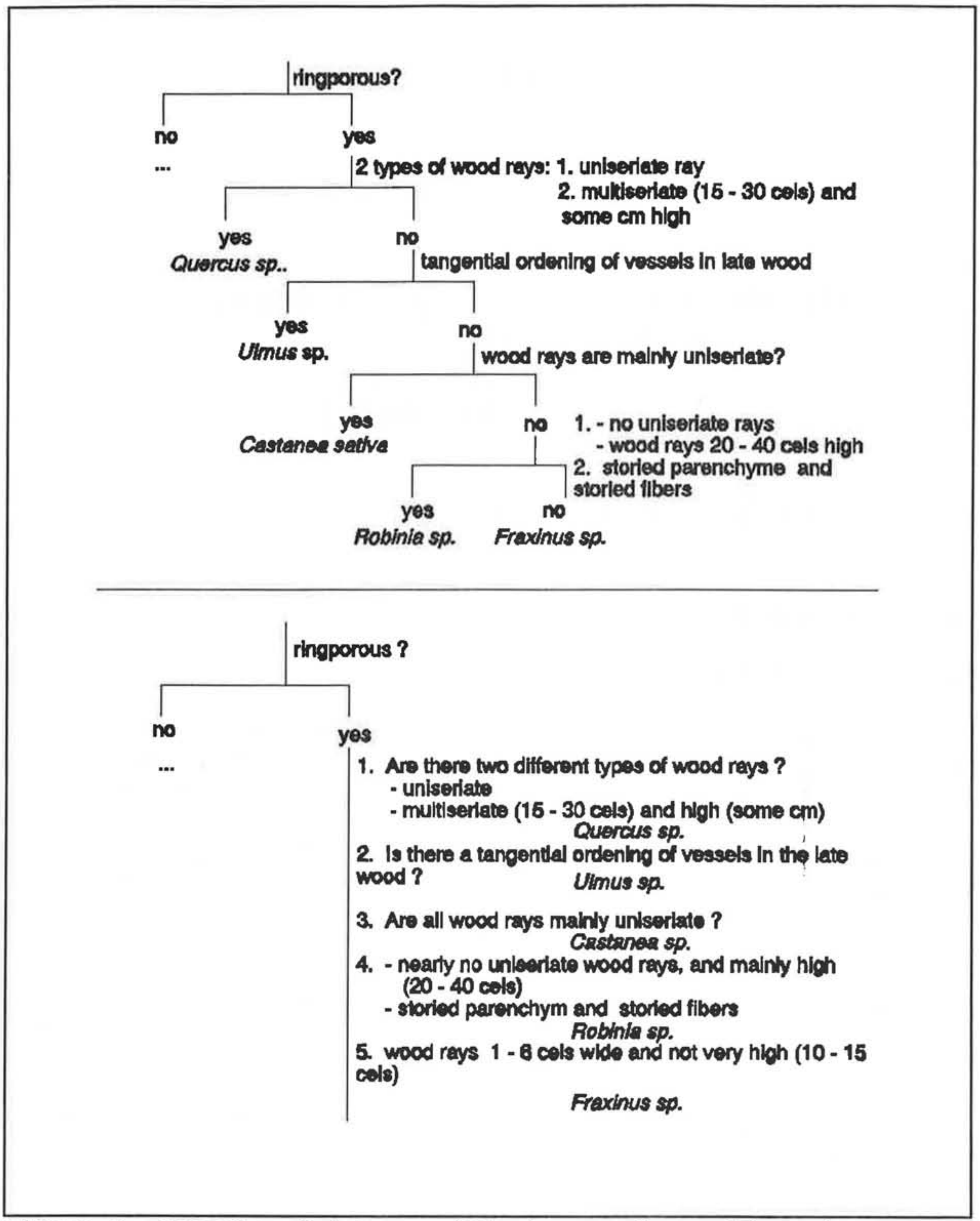
2.1.2.1. Content of questions

Questions are declared and attached to the hierarchical classification so as to be able to reach all possible solutions, and only those solutions. There are two main possibilities to formulate these questions (an example is shown in figure 6):

- questions with only yes/no answers
- questions with a number of possible answers among which the user makes a choice.

Preferring one of the two depends on different aspects. Questions with a number of possible answers might give rise to some malfunctions because of gaps in the search strategy. It is possible that a certain combination of features is not included (when there is no default or else clause). Another potential mistake occurs when an observation can be classified into two answers. These problems do not happen when using questions with yes/no answers only. The user model can also influence the choice of the type of questioning. For non-experts in the particular domain, it is advisable that he/she should not be confronted with complex questions and a lot of possible answers containing an abundance of information.

Wood identification is strongly based on the interpretation of visual information (e.g. recognition of microscopic structures). The knowledge engineer, while modelling the knowledge, should be aware of the fact that the end user has to


ringporous?
no
...
yes
2 types of wood rays: 1. uniseriate ray
2. multiseriate (15 - 30 cels) and some cm high
yes
Quercus sp..
no
tangential ordening of vessels in late wood
yes
Ulmus sp.
no
wood rays are mainly uniseriate?
yes
Castanea sativa
no
1. - no uniseriate rays
- wood rays 20 - 40 cels high
2. storied parenchyme and storied fibers
yes
Robinia sp.
no
Fraxinus sp.
ringporous ?
no
...
yes

1. Are there two different types of wood rays ?
   - uniseriate
   - multiseriate (15 - 30 cels) and high (some cm)
   Quercus sp.
2. Is there a tangential ordening of vessels in the late wood ? *Ulmus sp.*
3. Are all wood rays mainly uniseriate ?
   Castanea sp.
4. - nearly no uniseriate wood rays, and mainly high (20 - 40 cels)
   - storied parenchym and storied fibers
   Robinia sp.
5. wood rays 1 - 6 cels wide and not very high (10 - 15 cels)
   Fraxinus sp.

Figure 6. Example of two possible types of questionnaire.

translate visual structures in language concepts and vice versa. One can consider a translation module, being a component of the knowledge model as configured in fig 3, in which not only the level of expertise of the enduser is incorporated (complexity of the communication language), but also an adaption of the language for easy and clear translation to visual structures. This might for instance include hypertext and hypergraphics methods.

Another interesting remark for the system developer is that he/she should be cautious in choosing the distinguishing information to be processed in a questionnaire.

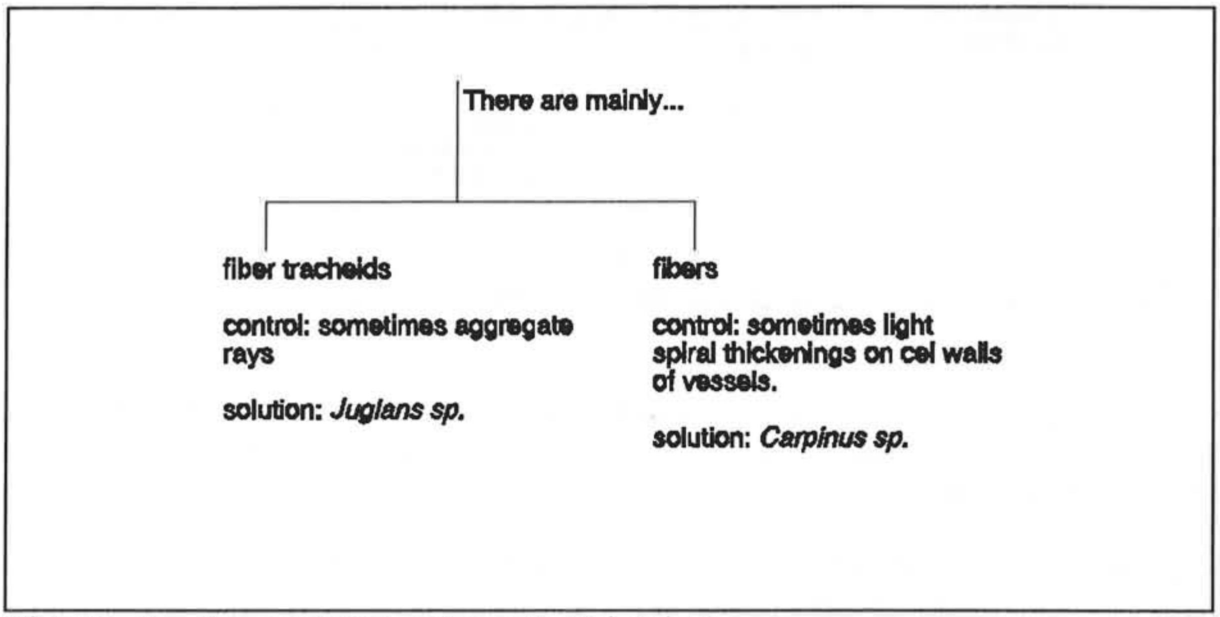

There are mainly...
fiber tracheids
control: sometimes aggregate rays
solution: *Juglans sp.*
fibers
control: sometimes light spiral thickenings on cel walls of vessels.
solution: *Carpinus sp.*

Figure 7. An example of a questioning.

When a certain feature appears not to be present in all cases, it is advisable to keep this information as control information and to choose another feature to distinguish between the different categories (fig 7).

2.1.2.2. The most efficient questioning

Which is the most efficient sequence of questions ? This order is dependant on the structure of the hierarchical classification which implies a kind of top-down specification. Although it seems quite reasonable to stick to a biological hierarchical classification, there are some non-biological factors influencing the sequence of questioning.

a) Visible features that are easy to detect should get a higher position in the decision tree. In general one can state that questions that have an answer that is easier to find, should be posed prior to the others. An example is shown in figure 8. At a certain moment during the modelling process one arrives at following distinguishing features:

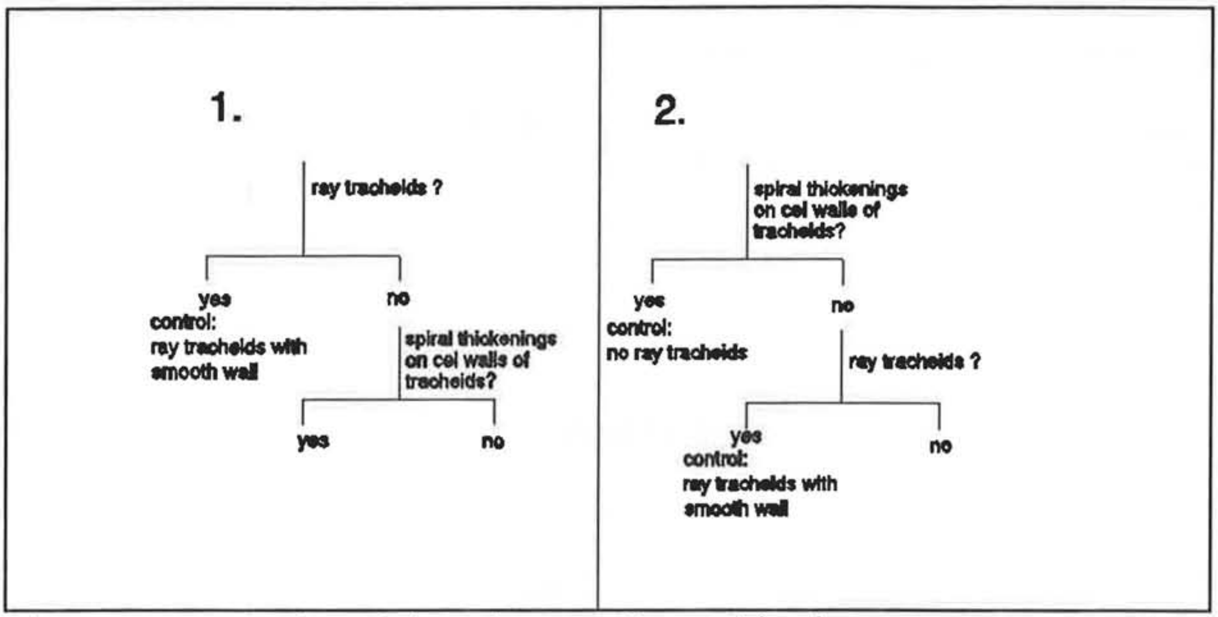


Figure 8. Example of two types of questioning.

- spiral thickenings on cell walls of tracheids and no ray tracheids
- ray tracheids with smooth cell walls
- no ray tracheids

It seems more advisable to formulate the questioning as designed in option number two, since spiral thickenings are visually easier observable than ray tracheids.

b) The frequency of the solution might also play a role in the sequence of questions. When for instance two distinguishing features (for two questions) have the same degree of easiness for detection, one can favour the question which includes the feature of a certain category that has a higher frequency of presence in reality, so that there is relatively more chance of that category to be the case. An example is shown in the upper part of figure 6, where oak (*Quercus sp.*) is first distinguished since the wood of oak is the most commonly used wood type in comparison to the other types mentioned.

2.1.2.3. What if the user does not know the answer ?

One possibility is to incorporate as much control information as possible. It is advisable to include all information available in the decision tree. When the enduser is not quite sure about his/her given answer, he/she can check whether the control information still fits the observations. When this is not the case, there should be procedures included in the system to revoke given answers and to return higher up in the decision tree.

What certainly should be included in the system is the possibility to give at any moment an "unknown" answer. The different categories that should have been distinguished at that point are then put together. In this way the possibility of ending up with different solutions gets higher.

2.2. Knowledge modelling and user modelling

The implementation of a system in a programming tool is normally based on a document which contains the combination of the knowledge model and the user model. As already stated, it is not very easy to separate knowledge modelling from user modelling.

For a wood identification system one can for instance decide that the endsystem should incorporate two modes (user modelling):

- the search strategy through the decision tree is performed as it is constructed in the model. The questions are asked in the sequence as defined in the tree.

- the user gives in all the knowledge he/she has about vessels, fibres, tracheids, wood rays, wood parenchyme, inclusions in cells, ... As information is still missing to reach a conclusion, the system will query the user for more.

The implementation of this second mode requires that the decision tree contains all knowledge available, so that it can easily be translated in another representation paradigm, to fulfil the requirements of the user model.

3. CONCLUSION

The main purpose of this text was to observe the knowledge modelling of a wood identification system. We illustrated the different observations with some examples as to make the reader aware of the pitfalls while modelling this identification problem.

4. REFERENCES

Breuker J. & Wielinga B. (1988). *Models of expertise in Knowledge Acquisition,* Memorandum 103 of the VF-project "Acquisition of Expertise".

Chandrasekaran B. (1988). *Generic tasks as building blocks for knowledge-based systems: the diagnosis and routine design examples.* The knowledge Engineering Review, september 1988 : 186 - 210.

Keravnou E.T. & Washbrook J. (1989). *What is a deep expert system? An analysis of the architectural requirements of second-generation expert systems.* The knowledge Engineering Review 4 (3) : 205 - 233.

Lenferink H. (1991). *AIDA, Leidraad bij de ontwikkeling van kennis(deel)systemen.* Kennissystemen 5 (5) : 7 - 11.

Wheeler, E.A., Pearson, R.G., LaPasha, C.A., Hatley, W., Zack, T. (1986). Computer Aided Wood Identification. General Unknown Entry and Search System. The North Carolina Agricultural Research Service. North Carolina State University, Raleigh, North Carolina. Bulletin 474A.