Associative Multilingual Classification Architecture for historical artefact

DAMIANO BOLLA

Computing Laboratory, University of Kent, Damiano.Bolla@iee.org

ABSTRACT

An artefact may be classified according to a conventional typology, typically based on some aspect of its appearance such as shape or decoration, such a typology classification is normally different from person to person and even the same person does change it over the time. Equally, it may be related to other artefacts by its presumed function or mode of manufacture. Other key variables may include the location and chronology of manufacture and of discovery. However, typology alone is of limited value and cannot fully document the processes of classification undertaken by the individual expert. It is even less useful to others who wish to understand the full range of associations between artefacts and their contexts of design, manufacture, use and deposition, much of which may remain hidden in the notebooks of experts. The situation is even further complicated by the fact that both the accessible and hidden documentation may exist in multiple languages that need to fit with the classification framework that is currently being used. In this paper we present a classification architecture that can be used to capture and document the knowledge and understanding of individual experts. The model is able to capture the need of classification and reclassification of artefacts, the need to associate semantic knowledge with an artefact and the need to be able to deal with knowledge in different languages. The approach is illustrated by an extensive case study of the classification of a collection of Japanese swords. Classification of Japanese swords does pose a particular challenge in terms of complexity of the subject and the peculiarity of the language. The case study demonstrates how technological, cultural, temporal and bibliographic associations are built up over time as the system is used to document these artefacts. The unique capability of the system to support and enhance the learning process of the researcher is shown by the refinement of the details embedded in the classification. The knowledge that is then embedded in the resulting web of relationship is further enhanced by the attached images and by the embedded support for different language views. Translation of words from one view to the other is aided by the web of knowledge that is bound to a word and even after the word is translated to another language the relationships are not lost but form an essential part of the translation. The formally strict definition of the architecture allows for the creation of a web of knowledge that retains in itself the semantic meaning of the classification and is therefore logically searchable.

1. INTRODUCTION

When the word Archaeology is mentioned in a conversation one usually thinks of it as brave people that work hard to hunt for hidden and mysterious treasures. This is surely one side of the story but another side, much less glamorous but maybe even more important is the long and tedious work of classification of the items that have been found.
Classification is really something that is embedded into our brain [REF] since no information sit alone and unconnected when it enters our knowledge. Being able to correlate information that is stored into our memory makes the brain the most wonderful searching machine.
It is a pity that the knowledge of one person classification effort is lost when he dies and in a way this knowledge is something that is forever lost if it is not put into a form that can survive the time.
It may be said that writing alone is enough to capture the knowledge of one's man but although this surely can be done for some logical thought or formal theory it is much more difficult to capture the wonderful correlation network of the mind.

2. OBJECTIVE

We attempt to study a possible pattern that the mind may follow when it classify historical artefacts and from this study derive a logical flow of concepts that can be incorporated into a software algorithm. This algorithm is then implemented and a real effort to use it is made.
It should be said that the study of how the mind classify data is not a rigorous one but it is done in an informal way by following the work of an expert of Japanese swords while he study his subject. It can be said that the classification of Japanese swords is in itself an art.

3. ANALYSIS

In this part we are investigating the work of a person that has the task of classifying a Japanese sword, the study is as accurate as possible and for every stage of work we try to deduce the logical step involved.

3.1 BUILDING THE BASIC KNOWLEDGE

No man is born expert and so all of us have been building their knowledge over the years. The same apply of our sword expert. What we are looking at is what kind of information is collected and how the mind sees them.
The basic information are related the world around us and they are conveyed by words, but they are really atoms that are not bound any particular word, the same atom of information can be described in different language but it still retain the same meaning.

3.1.1 PHYSICAL LOCATION

While studying historical artefacts one of the key information is where the artefact has been found, where it is stored, where it has been produced and so on. All of this kind of information is a physical location. An interesting thing to notice is that at first the location is not immediately classified into a structure of locations. At the beginning our expert just note that a particular sword has been produced in Tokio and Tokio becomes a physical location that is just that.
While studying swords it appear that more than one location should be grouped into a region and so mentally there is some sort of rearranging of the Tokio location to include that Tokio is bound to a particular Japanese region.
The process of shifting a particular atom from one more precise type classification does happen all the time in our mind and it is not uncommon to reclassify a particular atom in different ways while the knowledge increases. This kind of functionality should therefore be present in the proposed system.

3.1.2 TIME

Another key aspect of manufact classification is the time when the manufact was built, when it has been found, when it has been buried, etc. In this case it may be argued that just typing the year of the event may be enough to be able to specify when something happened to the item. But the reality is that our sword expert would like to have a finer grouping of time definitions. He may want to store a group of years under an era or may want to reclassify an era in a more precise way.
Time classification is then not just a year but a more complex grouping of events. It is somewhat strange at the beginning that it is not immediately clear if some "event" happens before or after another event. The point here is that often our expert reads that a certain sword has been built "during an emperor era" and this does not immediately translates into a well defined year. Only after our expert build enough knowledge this map of events becomes available.
It should therefore be possible to build an unordered list of events and be able, later, to order it in the desired way.

3.1.3 MATTER

Our sword expert is now trying to classify what the sword is made of. The same process of the two previous sections is repeated but in a different field. In this case the fact that a sword is made of particular steel or made in a particular way does not convey initially any broad knowledge. This means that only after a few swords with similar attributes are "learnt" then it is possible to see a common denominator and therefore be able to place the knowledge that a sword is made of a particular matter in a well specified area.
It seems to the writer that this process of knowledge refinement is common for all classification tasks in the field of archaeology, in other words classifying Japanese swords is not conceptually different than classifying roman vases or paintings.

3.2 ASSOCIATING KNOWLE2DGE

In the previous our expert has begun learning its basic building blocks for classification but it is immediately apparent that a full description of an object is the sum of apparently unrelated definitions. A sword may be made in one era and is built using a particular kind of steel and some part of it may be designed by a particular person. Our expert store in its mind the relationship between the above and it is important to note that the type of relationship between the associations is not a pure subset as it was in the type definition; in this case it is purely associative.

3.2.1 UNSPECIFIED ASSOCIATIONS

At the early stages it seems that associations can be unspecified. An unspecified association is one that declares two terms as being correlated but do not say what kind of relation there is. As an example we may look at the following association.
The tree being observed ends with AA-Unknown [Old Provinces] meaning that AA-Unknown is a subtype of "Old Provinces", it can be guessed that "Yamagata" that is a subtype of "Cities" has an association with AA-Unknown, furthermore it can be guessed that "Yamagata" is part of the "Old Provinces" but the expert does not know yet which one.

There are too many guesses that need to be done on the relationship of attributes, it is fundamental that the system is able to record not only the fact that some attributes are associated but also what kind of relationship there is.

3.2.2 SPECIFIED ASSOCIATION

Having noted that the children of a node better describe the node itself we note now that it is important to describe what kind of relationship there are between the node and its children. The kind of relationship we would like to model are the ones that English language allows. As an example let's assume that the node is a "sword" and one of the children is "date 1200" what we would like to model is that the "sword" has been build before "date 1200", we need a simple way to define the relationships.

3.2.3 VERBS

In the previous example the language being used makes it clear what we try to do. Since we would not like to invent new paradigms we are just proposing to use the same rules as the spoken language to capture the relationship between a node and its children. For each relation node – child it will be possible to specify a verb that defines the relation with the node. As an example let's take node "Jimmu" and a child node "Emperor" we may wish to define a relation that using the verb "is" that says that Jimmu is an Emperor. This is good in most cases but to better describe the relationship we really need something else to better qualify the verb.

3.2.4 ADVERBS

It is important to note that we use adverb not in a pure grammatical sense but simply as a way to define "a word that better specify the verb". As an example let's assume we would like to better define the relationship between the node "Tokio" and "year 500" and we would like to define that it was built in that year. We could use the verb "built" and the adverb "during". It is important to note that we the verb "built" may have different adverbs like "before" and "after".

3.3 MULTILINGUAL CONTENTS

A subject like "Japanese swords" is the most interesting for multilingual issues. Some objects are best described in a native language word but nonetheless we may aim to try to make a translation for it. The classification system proposed makes the job of translating a word simpler by providing additional information on what is the meaning of the word by following its position in the type tree and in the association panel. It is simple to accommodate objects from different cultures that do not exist in our own; we just need to classify them as best as we can and if they are not in the correct location the association with the parent type can be changed later.
One aspect that has become clear while researching is that also images are multilingual. Different cultures may have different images to represent the same concept and the system must accommodate this too.

4. SYSTEM DESIGN

Having broadly looked at how the knowledge of an expert build up it is now my goal to provide an instrument that allows the same logic to be extracted and manipulated upon. Such an instrument should be able to perform all the activities that out sword expert do while increasing his knowledge.
In the next part we try do define more precisely what should be provided and later in the document we will see how this can be provided.
A key point that is spread trough all the document is that the information being manipulated is somehow organised in a tree like structure. Note however that this is not the kind of strict tree that a computer scientist is used to deal with it is more a loose tree concept that is better expressed in terms of knowledge not being a flat, unorganised mass of items but instead having some sort of structure.
Even if at some time we use the tree paradigm it will be clear that this is just a starting point and by no means the end.

4.1 THE STRUCTURE

The structure reflects closely the analysis and is composed of three sections.
The panel that manipulates the types, a panel that is used to manipulate the associations between the "types" and a panel that is used to actually edit the values of the objects.
The right part is used to define the types and the relationship between them, the left part is used to associate different types and to define the relationship between associations; the lower part is used to edit the values being inserted.

4.2 MANIPULATING TYPES

From the previous section about building basic knowledge it appears that when we build knowledge we start from some abstract point and attach to that. We therefore need to define the "root node" as the starting point of all our knowledge. The root node does exist, it is logically essential but its definition is recursive since its existence is defined as the node that groups together all of the children of the knowledge.
To the root node we attach the basic items that we discover while we classify. There must be no need to be precise at the beginning and to have a precise classification tree in mind since the relation between nodes should be modifiable like it is our knowledge.
In the same way as our mind try to define an object into a well specified type the system should also provide this kind of coherency. So, one item belongs to only one well specified type. We will see how the association is the way to make an item appear to have multiple aspects.
It is important to note that it is the user that defines the types; nothing is constrained by the system. It can be argued that it would be good to have a uniformed set of types, and this is a good point but it is beyond this paper.
In the above example our expert has chosen to classify "Koshi" as a region and that itself is part of geography and general culture. Adding a new node should be as easy as selecting a parent node and adding to it a new definition. The software allows just that.
It should also be possible to delete a "type object" and more important to move an object bound to a type under another type. In this case the interface is not so intuitive since it uses the association window.

4.3 MANIPULATING ASSOCIATIONS

Associations allow our sword expert to attach correlated knowledge to an atom.
In the above image we can see how "Ekizen" is associated with "Fukui" and they are of different type. (The types are written in between square brackets. Each node provides a set of operations.

- Show Type: Displays on the right panel the given node as a type. This is extremely useful if we need to know how the node was classified.
- New Association: Allow the user to create a new association between the type currently being selected on the right panel (the type panel) and the node currently selected.
- Del Association: Deletes the association currently selected.
- Set Type: Sets the type of the object currently selected taking the value from the currently selected type in the type panel.
- Print: Prints a sub-tree of the association starting from the given node.

4.4 DEFINING VERBS AND ADVERBS

As described in the study the verb and adverb definition is peculiar of the association panel. For each node it should be possible to define what is the verb and adverb that specify the association. In the next image we see how it is possible to set the verb and adverb for an association shown in the left panel.
The association between "Mempo" and pg. 259 is defined in terms of "described" and "in" that defines that the Mempo Armour Mask is described in page 259 of a book. If we want to know what book we are referring to it is possible to select the object and ask the system to show the type tree of the page, in that tree we will find what is the book that contains page 259.

4.5 IMAGES AND OBJECT VALUE

An image can be associated to an object to better describe it. Image association is done in the same panel that is used to specify the values for the object. Defining an image for the object is a simple of dragging it into the image display and then press the save button.
Once the image is associated with the object a thumbnail is shown in the editing panel and it is possible to see a full size version by using the show option of the menu associated with the image. The main purpose of images is to better describe what the object is.

4.6 MULTILINGUAL DATA ENTRY

We aim to describe how the system provides support for input in different languages. First of all there is the provision to select a main language that is the language used as a reference and a working language that is the language being used to input the data.

The input panel will the show the main language value for the selected object and allows the user to write the content in the working language. An example of this is shown below.
In the above image we can see how the "Descrizione" is translated differently in Italian and English. Having completed or partially completed an English translation it is then possible to have as a main language English and do the work in German or any other language, including Asian languages.

5. CASE STUDY

In this part we are looking at the result of the work of our sword expert. It is important to remember that the work is not finished and that what is at the moment classified in a certain way may not be in the future. This reclassification work is done often for part of the tree and it is an essential feature that even if objects are moved around the database is still meaningful, meaning that there are no lost objects in it.

5.1 THE CURRENT TYPE DEFINITION

What follows is the current tree type. One interesting note is that the current classification is somewhat different from what I would expect. While I was studying the subject I kind of vision the first level of types to be places, then time, then abstract concepts, but it is not so. It happens that it is especially important not to force the user to a particular classification definition. Human minds are different and see things in different ways.

5.2 A SAMPLE OF THE ASSOCIATIONS

The association tree is somewhat similar to the type's tree. The main difference is that an object may be associated many times with other objects. Besides seeing the content of the knowledge as a tree like structure it may be useful to see it in some other format. Upon thinking what may be a suitable visual representation we did find the following one quite useful.
The element at the top is the association root node. It is possible to note that Blades – Oshigata is associated with both the "Blade Catalogue" and with the "Technical terms Dictionary (H-O)". This form of representation is quite useful to find out where a certain object has been classified with.

6. INNER SYSTEM STRUCTURE

This part aims to convey the inner system structure so it would be possible to replicate the finding in this paper.
The system is extremely simple given the power it has. It is made by a few logically correlated tables of data.

6.1 THE TYPES TABLE

The types table is in charge of storing the relationship between types, it represents a pure tree by using only two columns.

- obj_id: This column contains a unique id in the whole system of a given object/concept. It is essential that the id is unique since this is the basis for a "consistent" system. Note that no meaning is bound to the id, it is a pure number.
- obj_type_id: This column contains an id taken from the set of obj_id or contains null if this row describes one of the top types.

The two above columns are able to fully represent but no meaningful content is attached to them. To correlate some content there is the need for another table.

6.2 THE NAME TABLE

This table holds the information that describes an obj_id. This table contains data that can be read by a person. The structure contains the seed to make the system multilingual.

- name_obj_id: Contains an id from the sets of id defined in the types table in the column obj_id.
- name_lang_id: Contains and id describing the language this content is written.
- name_val: The actual value for the name of a given obj_id in the given language.

The table holding the description for an object has the same structure of this table with similar names.

6.3 THE ASSOCIATION TABLE

This table is again a purely numeric table since the content that should be display is defined in the names and desc table. Its purpose is to capture the web of knowledge that is being built using the objects being described in the types. The columns are as follows.

- assoc_obj_id: This column contains an id taken from the list of available id in the obj_id in the types table. The aim is to convey the information that this id is associated with another one that is being described in the next column
- assoc_with_id: This column contains an id taken from the list of available id in the obj_id in the types table. It states that the previous id (assoc_obj_id) is associated with another id.
- assoc_verb_id: This column contains an id taken from the list of available id in the obj_id in the types table. It states what kind of relationship there is between the two id described above.
- assoc_adverb_id: This column contains an id taken from the list of available id in the obj_id in the types table. It further describes what kind of relationship there is between the two id described above.

It is important to note that no textual information is actually stored in this table or in the types table. Both tables are purely abstract and the only aim is to capture relationships, not content.

7. RESULTS

Here we summarise the results in terms of the end product and its capability.
The resulting program is able to manipulate the type tree and allow types to be moved from one branch to another without compromising the integrity of the knowledge, in other words all types always have a parent type.
Associations can be as wide as needed and can describe a single type by associating many other different types to the one being under scrutiny. In addition it is possible to define the relationship between associations by defining a verb and adverb the describe them.
Multilingual is a key point of the system allowing coherent understanding of what is being described. A translation of a name or description is facilitated by the fact that it is possible to know more information surrounding the atom being translated.
Images are supported allowing for different images to be associated to an atom depending of the selected language.

CONCLUSION

The analysis or the methodology of work of our Japanese sword expert has shown an interesting definition of how classification can be performed. From this analysis a system design has been derived that can satisfy it.
The implementation and testing has been a fascinating part of the work since its result is a way to make our knowledge immortal and with the possibility to retain the relationships between knowledge, something that is normally lost in a standard document.
The work is by no means finished a further research should be done into searching capabilities and consistent definitions of verbs and adverbs but it is an interesting starting point.

FIGURES

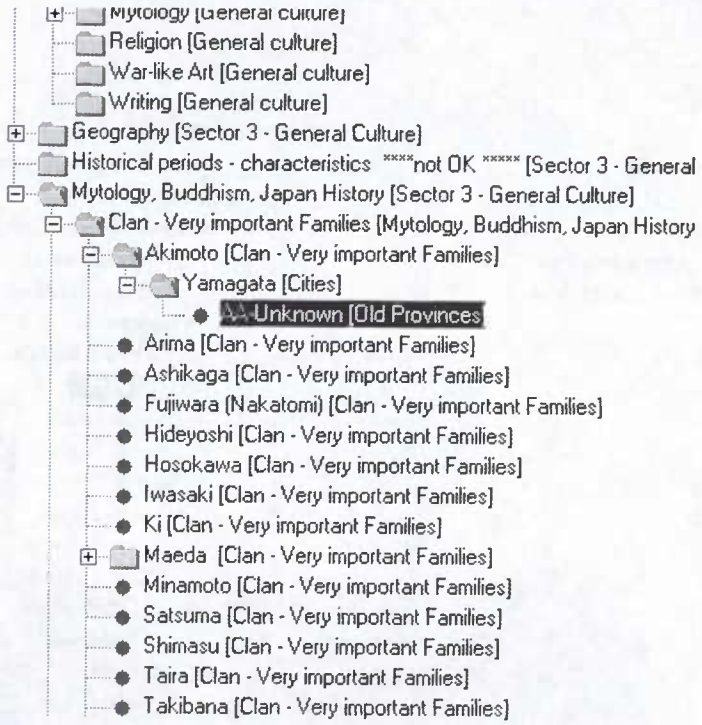

Fig. 1 – Example of association.

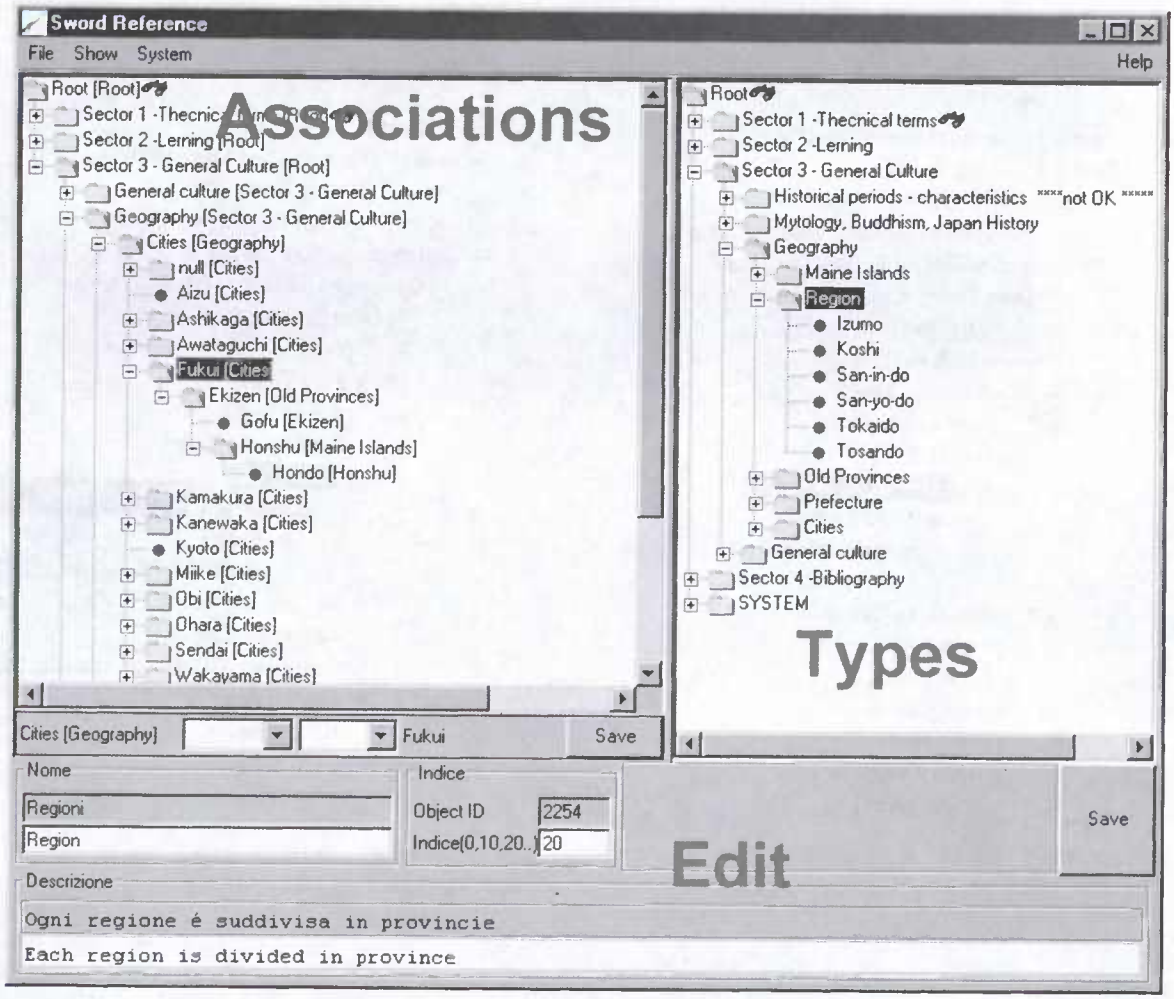

Fig. 2 – The main system structure.

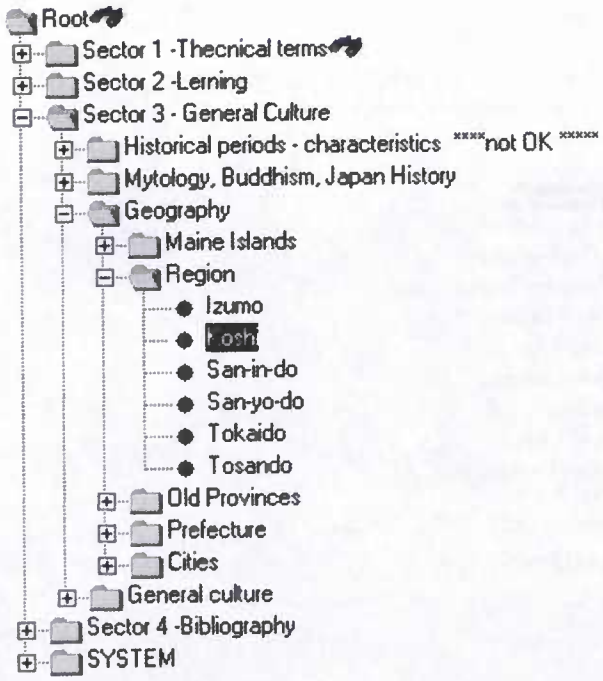

Fig. 3 – Classification by type.

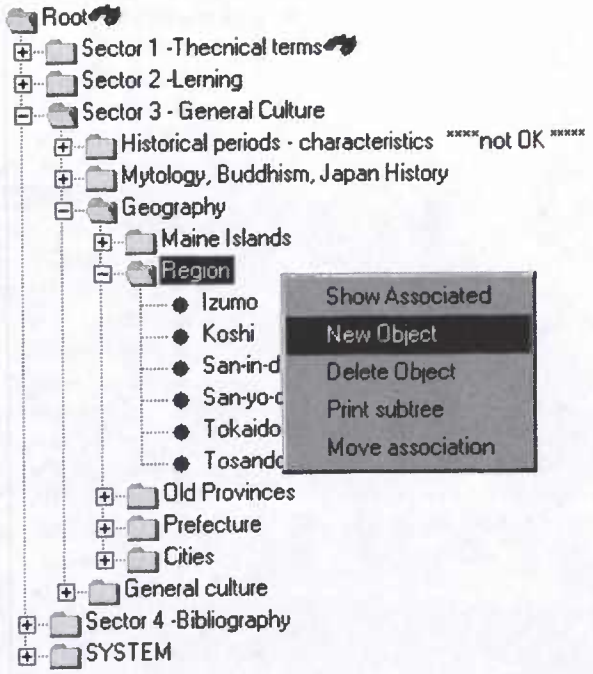

Fig. 4 – Adding an object to a type.

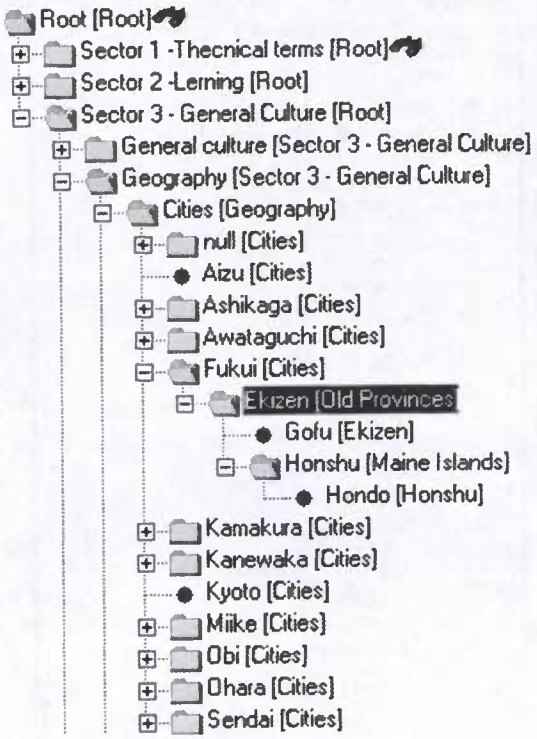

Fig. 5 – Association manipulation.

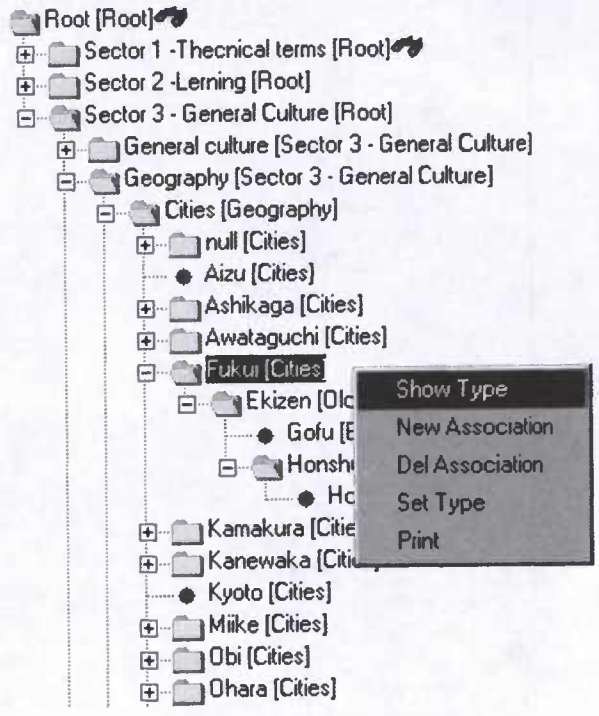

Fig. 6 – Possible operation on associations.

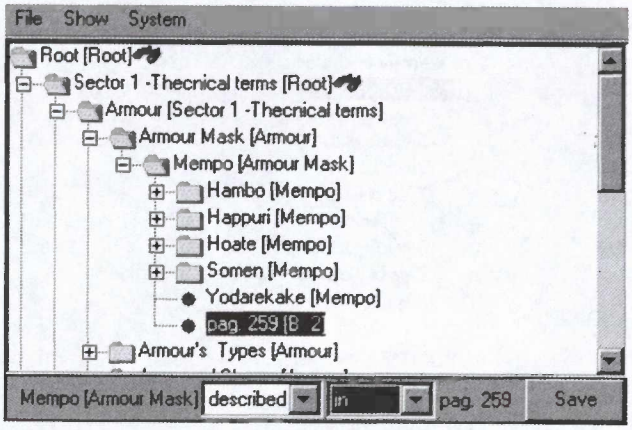

Fig. 7 – Verbs and adverbs definition.

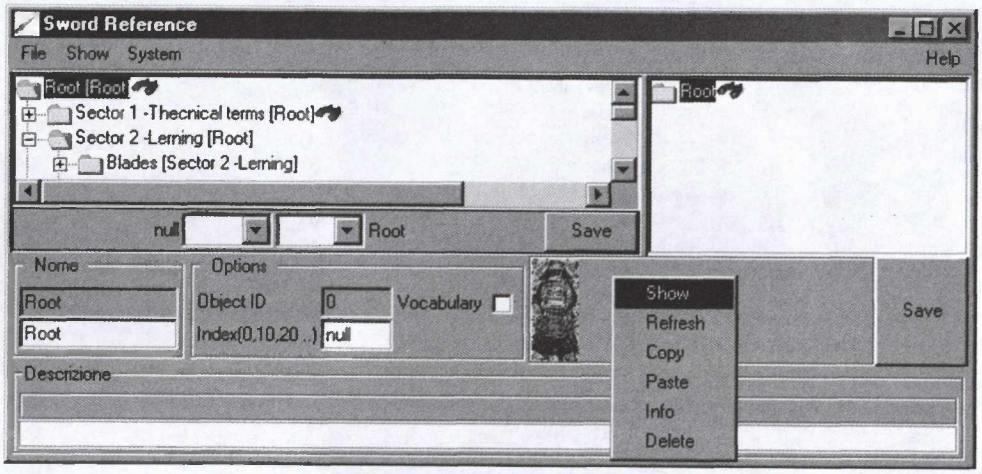

Fig. 8 – Images operation.

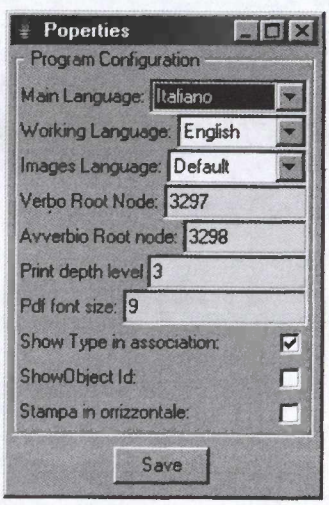

Fig. 9 – Control panel.

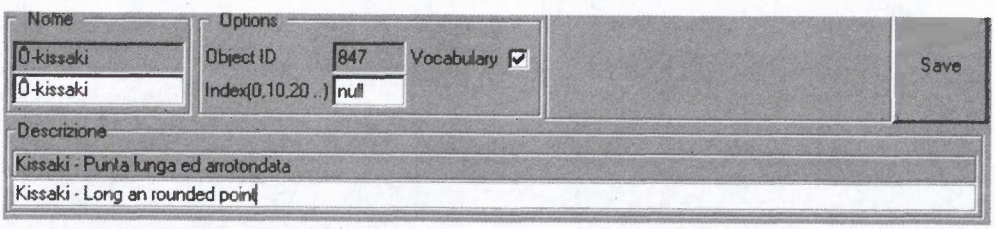

Fig. 10 – Multilingual input.

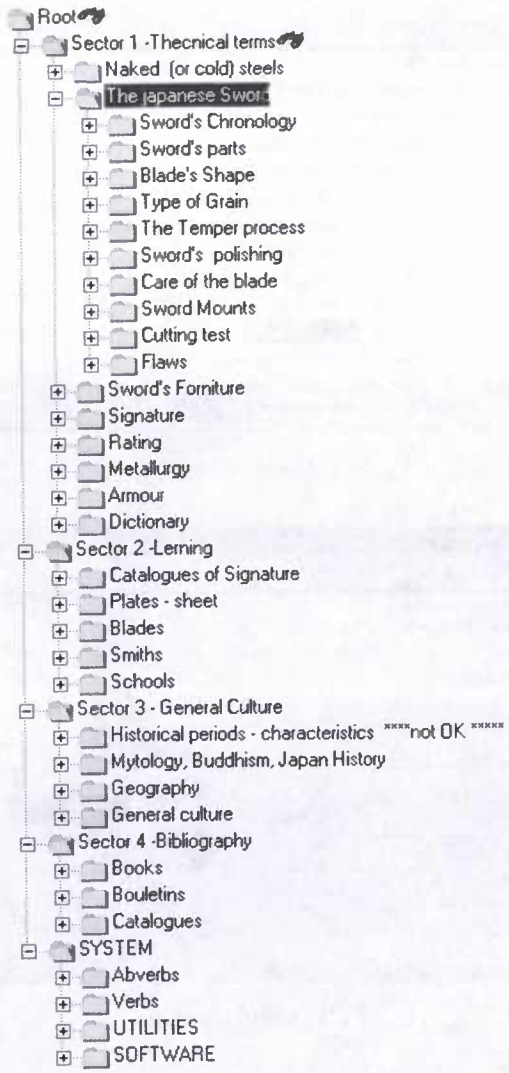

Fig. 11 – Part of Japanese swords types.

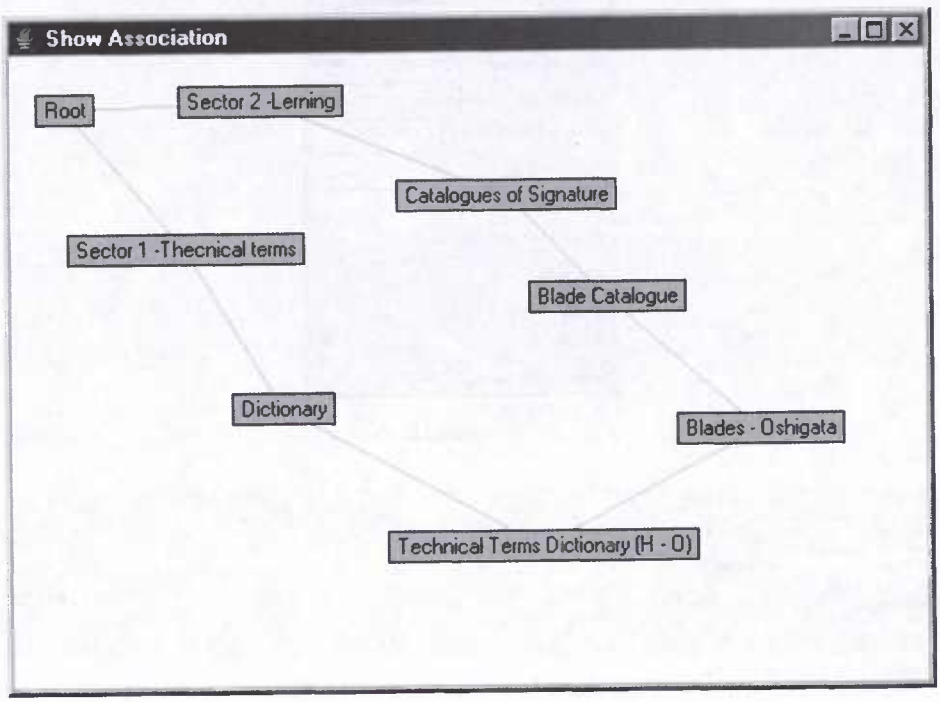

Fig. 12 – Another possible view for associations.