III

CODIFICA, FORMALIZZAZIONE E ANALISI DEI DATI
ESPERIENZE A CONFRONTO

DATA ENCODING, FORMALISATION AND ANALYSIS
COMPARISON OF EXPERIENCES

Archeologia e Calcolatori
15, 2004, 201-218

ON THE CODING OF ARCHAEOLOGICAL FINDS

1. Introduction

The coding of the information is a crucial step previous to any kind of statistical or mathematical treatment. This is evident, since it is impossible for any data analysis tool to extract more information than the one that was coded; in other words, the aim of the coding is to allow to analyse the information one wants to study. Thus this task is very delicate, since it is a true interface between the observed reality, as seen by the researcher, and the data analysis tools, developed by other scientists, such as statisticians, data analysts, mathematicians, computer scientists, etc. This is complicated by the differences existing in the two domains: whereas the researcher aims at coding his/her material according to the criteria typical of his/her scientific frame of reference, the data analysis methods require a very strict way of coding, in particular such that every unit composing the analysed population or sample is described in the same way by the chosen coding. Only in this way an effective data table can be built and analyses can be performed.

It is obvious that only a specialist in a specific field can do a coding able to take into account all the information he/she considers of importance. On the other side, he/she must be aware that the analysis techniques appropriate for the aims of his/her investigation require that the coding observes some rules, in order to be applied. For this reason, it is advisable that such a work is done with the cooperation of a data analyst, who can check the specialist's needs and ensure that the coding matches the requirements of the methods that must be applied.

2. Coding in archaeology

The problem of coding archaeological finds is far from a solution. Such a problem involves some concepts, such as classification and type, which are very difficult to deal with. So broad is the variety of the material that it is very difficult even to describe all the possible coding.

For some categories, such as ceramics and lithic material, there are traditional ways of coding, tied to typologies based on very different principles, both empirical and systematic (see Adams, Adams 1991, for a synthesis). During the last decades, it became more and more clear, that *the best* typology does not exist, but different typologies may be appropriate for different study aims, based on an *a priori* classification of items. In addition, an archaeological typology is usually based on a hierarchical classification, namely a set of encapsulated partitions[1], sometimes difficult to both code and analyse suitably, since the final types may be subcategories of larger ones.

To quote some, the French group working around J.-C. Gardin tried to develop a universal coding for cylinder seals (Digard 1975), pottery morphology (Gardin *et al.* 1976), and decoration (Gardin 1978), but the attempt was not completely successful, since it seemed too difficult to reduce such a broad variety to a univocal coding. In fact, at least according to the majority of scholars, the border among types is too fuzzy to be defined in a simple general way and, in each situation, it must be adapted according to some *local* criteria, local meaning a criterion limited to a set of artifacts in a particular study. In general, the approach aiming at a universal coding, highly complicates the coding with redundant information that hides the truly significant characters of each corpus of material with too general ones. As a consequence, the general codes are not useful – the attribute *cylindric* does not help in distinguishing a Chinese vessel from a prehistoric one – whereas in the detail the system is too complex and difficult to follow.

So, the variety of the items to code is so broad and different, that the scholars had rather leave the utopian aim of a perfect universal coding and choose both the set of characters to take into account in a specific study and the appropriate way of coding such information. As an alternative, one has to think about a technique able to translate a universal code into some coding more handy for particular purposes: Camiz and Rova (2001) compared a qualitative coding with a textual one, both used for the study on images of seals and forecast some possible method for unifying the two coding in this particular case. Nevertheless, it is possible to describe the different kind of data used in archaeology, that can be classified according to their nature. Since the coding highly influences the analysis methods that can be used for their treatment, some attention will be paid to this matter.

It must be emphasized that some particular problems arise when coding archaeological finds. The first problem is represented by the *missing information*, depending either on the condition of the object or on the loss of the associated information. Furthermore, most of the characters require a qualitative coding, a coding that may raise problems in the analysis treatments. The qualitative coding requires that specific levels are defined prior to the coding, in order to attribute the finds to them correctly. Since the types are usually defined in a hierarchy, one may wish to keep track of the hierarchy in the coding, a task very difficult to fulfill, unless by dramatically increasing the number of characters taken into account, that,

[1] A *partition* of a set is a family of disjoint subsets of one set, whose union is the set itself. Two partitions are *encapsulated* if each subset of either is partitioned in subsets belonging to the other.

in addition, could highly complicate the following analyses: in fact these methods should be able to handle the hierarchical structure, a task that would require specific techniques.

In the coding other specific problems raise from the nature of the finds: in fact the border among the attributes of a character may be uncertain, so that it may be difficult to suitably code an item. This difficulty can depend as well on the incompleteness of the item or of the find, so that an incertitude can result in the coding. Both problems may be solved in the frame of *fuzzy coding*, where a *degree of belonging* of an item to a set is considered, instead of the common dichotomy (Zadeh 1965; for its application to archaeology, see Hermon, Niccolucci 2002). As a result, the typical qualitative coding, where each possible value of a character is specified as a level, may be no longer used and must be replaced by a complete set of degrees of belonging, that is a set of real values ranging 0-1, one for each possible level. Andrenucci (1998) used a similar coding for the uncertain date of some Egyptian scarabs. All these problems, specific of archaeological finds, force to adjust the currently used analysis methods or to develop new ones, that take into account the specific coding conditions.

3. Coding the archaeological information

The coding of archaeological finds concerns different information that usually deserve to be considered in their description. So one may distinguish among:

– *outer information*, i.e. the information that concerns the context of the find, but not pertains to the object itself, such as the site where the object was found, its function, its association with other objects, etc. In addition, the state of the object (complete, broken) may be considered here, since this does not pertain to its original characters, but is an accidental result of its later “history”;
– *inner information*, the information concerning the object itself, that may be subdivided according to the different kind of features that characterise the object:
  a) *physical properties*, such as material, dimensions, colour, etc.
  The coding concerning both the outer and inner information, the latter limited to the physical properties, is usually a classical one: the characters may be of presence/absence or multistate kind, so that a qualitative coding is suitable; measures, thus quantitative; in some cases, like the dates, they may be of scale kind. So, such characters are not difficult to deal with, since their treatment is well known and practised.
  b) *morphology*, the information concerning the shape of the object, when this is a major issue in distinguishing the objects belonging to a corpus under study.

This is the traditional basis of every archaeological typology, that is usually performed by adapting a traditional one, coding the presence/absence of some specific type. Nevertheless, this is one of the cases that may take advantage of some new coding technique, in order to solve the problems outline above. In particular, a spatial coding, based on 2- or 3-dimensional coordinates of specific key-points identified on the object, allows the use of the *shape analysis* (Dryden, Mardia 1998) a technique for the study of homogeneous objects, that enables a very fine tuning of the smallest details.

c) *textual/iconographical content*, is the information concerning the content of the text or of the image represented on the finds.
Such information may be simply coded referring to the most important meaning of the text or the image, such as *dedication*, *proverb*, *list of names*, *account* or *war scene*, *sacred procession*, *animal contest*, etc. Instead, should one need a greater detail in the description, the whole text or a thorough description of the image must be taken into account. This claims for a textual coding, obvious in the case of the engraved texts, but relatively new in the case of iconography.

- *Texts coding*. The simple storing of the text is certainly the solution for most of the needs. Nevertheless, some further treatment, such as indexing, tagging, and the addition of qualitative characters, such as those concerning palaeography, or line subdivision and paragraphing, may be helpful for a deeper study.
- *Images coding*. The image coding depends largely on the level of detail one wishes to consider for his/her investigation. First of all, the structure of the drawing/painting/engraving that appears on the object, the colour, etc. should be considered, then the iconographical content. This may vary from the simple identification of a decoration (*flowers*, *leaves*, *geometric elements*, etc.) or a type of traditional representation (*war scenes*, *country scenes*, *mythical scenes with Achilles*, etc.), to the most detailed description of all elements that appear in the image itself, together with their particular attributes and their mutual relations. Of course, whereas for the simplest level the coding is relatively easy, for the latter the coding should be carefully organised, in order to consider all the image aspects one wishes to record.

The two cases of shape and image coding deserve a high interest, since they are susceptible of new developments in coding. The coding of the shape of the finding can be in some cases of qualitative character, as for instance when one wants to distinguish a *jar* from a *cup* or a *dish*, etc. As an example, dealing with seals, one can distinguish *stamp* from *cylinder seals*, and different types of both. Actually, we refer here to a different case, in which a set of finds have a similar aspect (like points of arrow, scarabs, pots of the same basic shape) but some variation in shape that may not be caught with a different term.

In the following, I shall try to describe the new ways of coding the information in the cases in which I worked recently, that seem promising for further developments: the information concerning the shape of objects of a very similar kind and the information concerning the iconographical content of images. Both cases seem applicable to very large corpora with outstanding results, but their potential is not yet fully investigated.

4. Beyond the qualitative coding of morphology: shape coding

The coding of the shape of an object, when one wishes to distinguish it among a set of analogous ones, like arrow points, scarabs, vessels of similar profile or shape, etc. may not be done through a qualitative coding, since either two shapes are identical, thus coded in the same way, or are different, thus coded differently, without any information on how different they are. A coding of objects distinguishing among *type_1*, *type_2*, etc., is possible, but does not seem very helpful. For this reason, it is advisable to choose a quantitative coding, able to evaluate the amount of shape difference among the objects.

As an example, considering Egyptian scarabs (that may as well be studied on the basis of the text engraved on the bottom), their different shape depends on many different details in their carving. Andrenucci (1998) built a data table of scarabs by selecting five different values for each of 22 selected characters of the carvings, something similar to the classification of scarabs heads shown in Fig. 1 (Tufnell 1984). Albeit this allowed a classification and an attempt to select the characters that could be used for the dating of the scarabs (Andrenucci 1998; Camiz, Venditti in press), it was not sufficient to fine tune the differences among samples, since, in this case, only a drawing could render exactly the true shape of the scarabs and allow a good comparison. Thus, in order to deal with shapes, the *landmarks* technique may be used, considered a sufficient approximation of the true shape.

The assumption of the landmarks technique is that on each object belonging to the corpus under study, a set of key-points may be located, the landmarks, whose reciprocal position is sufficient to identify the whole shape of the object, in particular in comparison with other similar ones. As an example, dealing with a 4-vertices polygon (Fig. 2), the vertices are such landmarks, since from their position all the polygon can be identified. A distance among the two configurations may be defined, based on the shape analysis. The underlying theory is that there exist non-linear plane operators that transform the coordinates of one sample landmarks into those of another sample. In particular, *pairs of thin plate splines* (*PTPS*) may be used, a family of curves that can be used as generalised coordinates systems on the plane. Given two images, two such pairs may be defined, so that the position of the

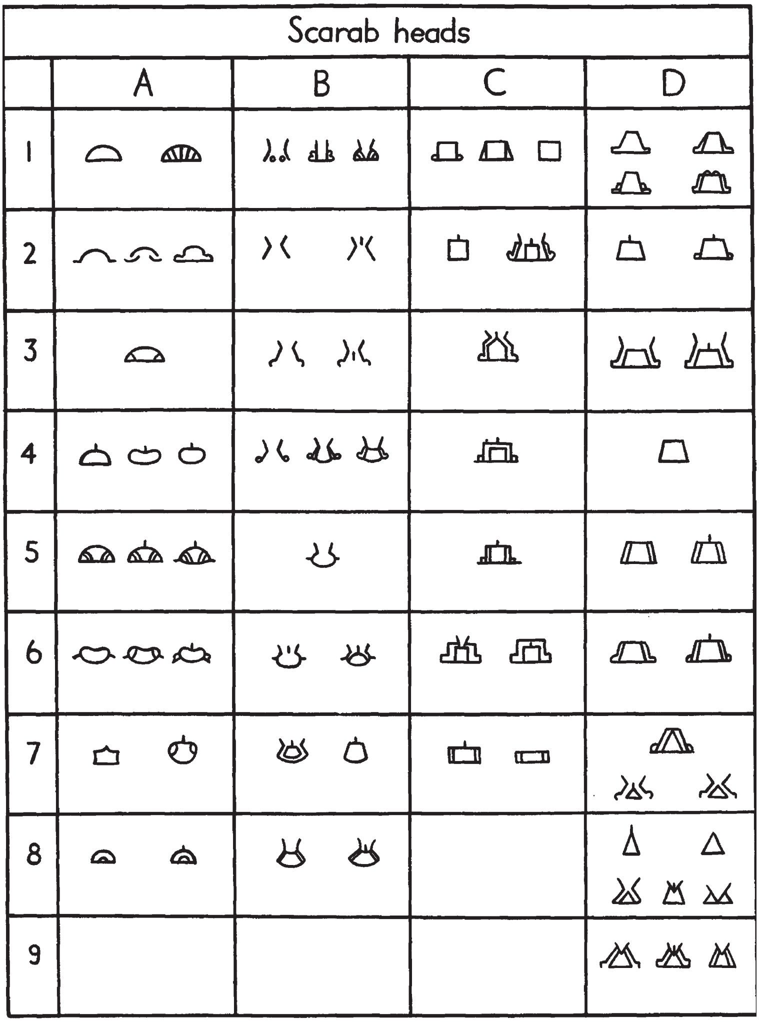


Fig. 1 – The classification of scarab heads, according to TUFNELL (1984).

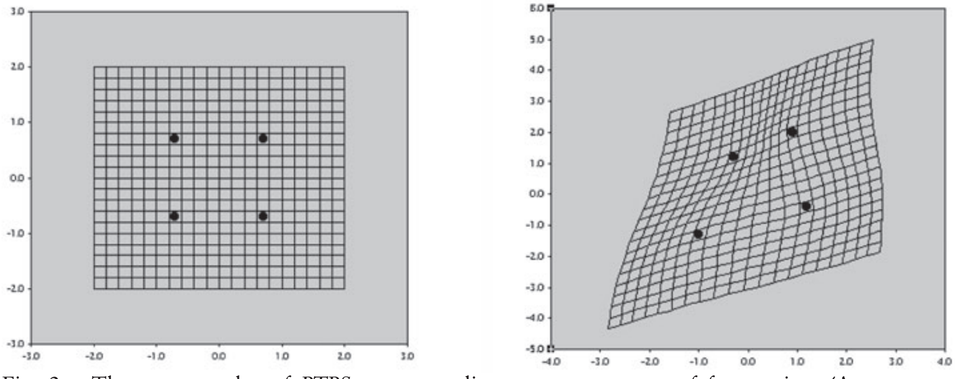
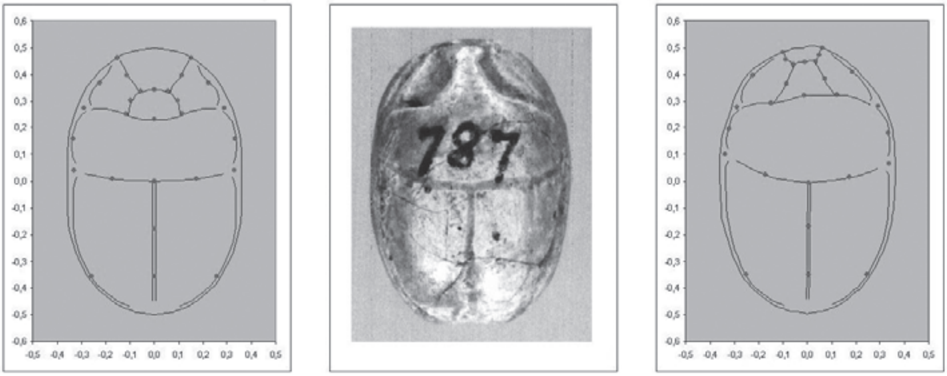


Fig. 2 – The two couples of *PTPS* corresponding to two patterns of four points (ANDRENUCCI, ANDRENUCCI 2002).



a) b) c)

Fig. 3 – The use of landmarks for the description of Egyptian scarabs: a) the landmarks of an ideal scarab; b) the scarab 787 of Florence Museo Egizio (Andrenucci, Andrenucci 2002); c) the landmarks of the scarab 787.

landmarks of either scarab on the corresponding *PTPS* grid is the same in that particular grid (Fig. 2). Analytically, this corresponds to a deformation operator of the plane, represented by the parameters of the transformation that lead from a *PTPS* grid to another. From these parameters, an association measure among the samples can be defined, to be used for classification.

Such a coding needs some extra treatment, since the position of the object in respect to the origin of the coordinates and to the axes and its size heavily influences the landmark coordinates. So, it is advisable:

– to *rotate* the objects, in order to represent them always with the same orientation and from the same point of view;
– to *center* the coordinates, setting the origin to the centroid of the landmarks[2];
– to *normalise* them in some way.

As a result, *standardised* coordinates become comparable and the *size* of the object is represented by a different quantitative variable[3]. If there are curve edges, extra points are necessary, in order to reproduce sufficiently well the curve with a set of segments (a polyline). Clearly, the number of extra points should be the same for all objects in the corpus. So, having fixed a two – or three – dimensional orthogonal reference system, the two or three coordinates of each landmark on each object are taken.

[2] The *centroid* of a set of points is the point whose coordinates are the average of the coordinates of the points of the set.

[3] In statistics, when one wants to compare two variables, regardless of their variance or their range, it is customary to *normalise* them, that is to reduce them to equal variance or range. In this way, the comparison does not depend on the intrinsic distribution of a variable. If a variable is both centred and normalised to have variance equal to 1, it is called *standardised*.

The system of coordinates of the landmarks is thus a coding for the shape of an object. As an example, in Fig. 3 *a)* an ideal scarab is shown, with a number of landmarks on it. In Fig. 3 *b)* a scarab is portrayed, whose landmarks are shown in Fig. 3 *c)*. Andrenucci and Andrenucci (2002) applied the landmarks technique to scarabs, in an attempt to overcome the problems raised during the exploratory analyses based on the qualitative coding.

5. Textual coding

Also the textual coding deserves a special consideration in our discussion. Textual coding means that a character of an object is described through a text, by no means the most flexible and appropriate code for transmitting an information, apart from the scanning of an image. Indeed, when the iconographical content of an image must be investigated, the textual coding is even better, since it puts in evidence the interpretation, say the meaning of the image, as understood by the scholar, rather than the image itself. Actually, defining an image as a *woman craftsman* is by far more precise than simply *woman* or, worse, *human*, as it could be automatically recognised by an automatic algorithm. Anyway, should a very precise algorithm exist, it would be a tool for providing automatically a classical coding: most useful, but nothing more.

The description of an object through a text can take into account any particular aspect one wishes to record. This is a major advantage for a very thorough coding of every detail one wishes to put in evidence, since the description can be as detailed as a very complete textual description can be. This shifts the discussion to the textual studies. Their development started once the textual analysis was introduced, first as a method for studying literature texts and further in surveys, in order to avoid to submit to the interviewed a set of possible answers already organised. Indeed, it was proved that, in this case, the interviewed is conditioned by the grid of proposed answers, since:

- it is likely that one does not find among the given items the one that he/she really means, so that the chosen one is only a rough approximation of his/her meaning;
- the presence of given answers drives the interviewed to pay attention to items than do not pertain necessarily to the sphere of his/her concerns;
- even the order in which the answers are shown can influence the interviewed.

In the recent years, *textual analysis* (Lebart, Salem 1994) became an important topic in the frame of data analysis studies, considering the large number of different fields of investigation that can take advantage of such techniques: suffice here to remind, apart from literature and linguistics, that it found applications in sociological, psychological, and marketing surveys, in political sciences, in the automatic classification of information, etc.

It is evident that, once the archaeological studies concern finds containing written texts, the textual analysis techniques may be applied, exactly as they can be applied to other kind of texts. In addition, it can be also applied once that the text is used as a special coding, an interface for the description of particular structures, as in the case of images. A long lasting investigation on this topic was carried out with Elena Rova, based on a corpus of Uruk/Jamdat Nasr cylinder seals, whose images were aimed at being classified according to the iconographical contents (see, among others, Rova 1994; Camiz, Rova 1996). This textual coding will be discussed in the following section. Paola Moscati used the textual coding for the study of stone cinerary urns produced in Volterra during the Hellenistic period (Moscati 1997a; 1997b). The aim was to go beyond the more traditional coding based on the presence/absence of characters. Thus a textual formal description was achieved for both the iconographic information (the series of architectural mouldings, which characterise specific types of framing and have been considered as countermarks, i.e. distinctive trade-marks of workshops) and the stylistic data (drapery and hairstyle of the human figures represented on the chests or lying on the lids). This coding was proved effective to study the problem of architectural mouldings – avoiding an *a priori* typological definition – to verify the significance of the relation between crown and base mouldings, and finally to analyse the modes of production (Moscati 2004).

6. Beyond the qualitative coding of images: textual coding

When studying an image by considering its iconographical content, it may be sufficient to simply record the contents of the image in a very general way. This is what Camiz and Ferrazza (in press) did for their investigation on the different facets of the Ajax myth represented on ceramic artifacts in ancient Italy. In their work, the images were coded according to different representations, such as *Ajax playing dice*, *Ajax duelling with Hector*, *Ajax's embassy to Achilles*, *Ajax carries Achilles*, and others.

Such a coding was not sufficient for Camiz and Rova (1996, 2001) study on cylinder seals. In fact, in this case the aim was not limited at classifying the seals according to a general image subject, moreover very difficult to partition in a small number of types. Rather, we aimed at identifying different types of seals, based not only on their general subject, but also on the complexity of the engraved images. So, aiming at studying seals in the deepest possible detail, we decided to take into account at least three different levels:

– the elements that appear in the image, like different kinds of *human beings*, *animals*, *objects*, or *symbols*, and their attitudes, like *sitting*, *passing by*, *with open arms*;
– the small, sometimes repeated subpatterns that compose the image and may occur identical on different seals, such as *a woman with open arms sitting*

on a bench carrying a vessel, or *two rampant animals in front of each other with an object in-between*. These sets have major importance in the study of object with engraved or painted images traditionally composed of a sometimes repeated set of subpatterns, whose combination allows to compose different images; indeed, this suggests to broaden the description to:

– the relations among subpatterns in the image composition, such as *image composed by two sub-patterns, the first subdivided into three*.

The problem of coding images, when the iconographical contents is of importance, cannot take advantage of the landmarks or more generally of the image processing techniques, since they are useful for the identification of spatial pattern or, when image recognition techniques are used, for the identification of objects, that may be found by aggregating the image pixels. Not even the shape coding is appropriate, since it is impossible to fix the landmarks once that the iconographical context is different.

It is clear that, according to the choice of a level of detail, a specific coding should be chosen. In principle, each different level may require a different coding, that must have specific characteristics, in order to suit the descriptive needs.

When dealing with the most detailed level, one should consider the following points:

1) For the first level, it is important to identify a set of elements/attitudes whose presence or frequency in each image should be taken into account. So, a classical data table crossing images with either presence/absence or multilevel characters, each identifying a specific element/attitude, may be used. It is clear that the identification of iconographic elements and positions, etc., involves a certain degree of arbitrariness, which should be part of the scholar's (in this case, the archaeologist's) responsibility. Of course, the problem is not limited to the archaeological framework: every scholar in every field selects a particular subset of a population of reference (the *sample*) and, for each sampled unit, he/she selects a particular subset of the information available.

The definition of some of the iconographic elements involves a certain degree of interpretation, since they are defined both by formal features and by their function inferred from the image context. A purely descriptive codification, as advocated by SUTER (1999, 49-51), in which figures and objects are identified in terms of their postures, gestures and formal features alone, could be adopted, but it would result in a too large number of elements, many of which would not be useful for the specific analysis aims (see, e.g., the type of coding proposed for seal images by DIGARD 1975).

For the definition of positions, more strict formal criteria can be used, so that a cross-check for the identification of the iconographic elements may derive. If one wants to keep track, in his/her study, of the elements taxonomy, a hierarchy of characters can be used: a specific element, say a *female craftsman*, may be taken into account at the same time as *human being*, *female*, and *craftsman* by simply coding the presence of all three characters in the image. As well, specific *animals* may be coded in addition, as *lion*, *caprid*, *snake*, etc., and the same may be done for objects. As concerns the attitudes, one can consider general characters as *passing*, *rampant*, *sitting*, etc., and variants of them, like: *with parallel arms*, *with open arms*, *with parallel paws*, *with turned head*, *upside down*, etc. This is the classical coding, whose limitations are evident, since no information concerning the subpatterns or the general image structure can be derived from the coded data, unless specifically previewed.

2) For the second level, the said classical coding is not sufficient. At this level, the subpatterns are the structures which deserve the highest interest, but it is not possible to code each of them as a specific character, since too many different characters should be taken into account, and the small differences among them could not be identified correctly. A textual coding could be suitable, once, for each image, one builds a specific text, fully describing its content. The advantage of such a coding is its ability to describe the relations among elements and/or attitudes and to represent at the same time the subpatterns, simply through a subtext sufficiently detailed.

Such description of the image contents is a tool more flexible than the first coding. In fact, once are defined very strict rules for the construction of a formalised descriptive text, the information transferred in this way is sufficiently complete to understand the content of the image: apart from the *style* of the image and some minor differences, one could actually rebuild completely the image based on its descriptive text. In particular, special care must be devoted to use constantly the same form for the description of the same element or the same attitude. Furthermore, all terms should not be inflected according to grammatical rules, otherwise the 1-1 correspondence between descriptors and described objects would be lost. Finally, when using a textual coding, special attention must be paid, in order not to include in the coding some uncontrolled bias, due to ambiguities in the meaning of the words or to differences in the literary style used for the coding.

Several advantages may be attributed to such coding: with a good practice, it may be easier to proceed to this coding than to the previous one; instead, with particular care, the previous one may be included in, or at least automatically extracted from the latter, via a computer program. Furthermore, the text may be inspected not only for the identification of the occurrences of a single form, corresponding to an element or an attitude, but as well for either *repeated segments*, i.e. sequences of forms that appear exactly in the same way in different texts, or *nearly-segments*, i.e. sequences of forms differing from each other only for one or two forms (BÉCUE, HAEUSLER 1995). Segments and nearly-segments are very important for our

Animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left above animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left above animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left on animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left and animal undefined passing left.

Row plus man craftsman passing with parallel arms left; altar; man king priest passing with symmetric arms right; standard type_3 left and standard type_3 left, on altar, on bovide passing left; man craftsman sitting left plus row. In boat.

(S.S.S.S.S)·(S.S.S.S.S)·(S.S.S.S.S)·(S.S.S.S.S)

((x+S).(X).(D).(((s.s)/(x))/(S)).(S+x))·(X)

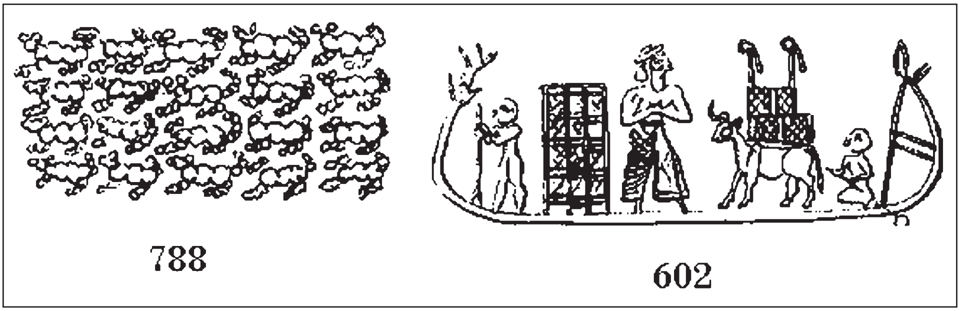


Fig. 4 – Two seals images from Rova (Rova 1994): on the left (788) a single element repeated on multiple lines and on the right (602) a non-periodical complex image. Above the textual coding and below the symbolic one.

descriptive task: in fact, there are objects and/or attitudes that may be described only through *polyforms*, i.e. sequences of forms having a unique meaning, as the "*king priest*" that appears in the centre of the seal 602 in Fig. 4; in addition, the association among elements and attitudes is as well described through polyforms, i.e. segments. Furthermore, image subpatterns involving two or more elements may be described thoroughly through segments, or at least through nearly-segments, when only minor differences exist among them.

In the study on seals, the construction of the formal text for the description of the images was done according to a set of fixed rules. Starting from the top left of the image, continuing rightwards and from the top to the bottom, each icon was described by means of a sequence of lexical forms, defining, in this order, the iconographical element, its position, the position of arms and/or paws (according to the same criteria used in the classical coding) and its orientation (*right*, or *left*). Furthermore, additional lexical forms were added to record specific attitudes (e.g. the presence of animals *with turned* *head*). Different elements were connected through relation markers (*and*, *plus*, *on*, *intertwined with*, *inside*, *above-below/alongside*), while different subpatterns were divided through punctuation marks: *comma*, *semicolon*, *period*. A detailed description of the original coding procedure of seals images is given in ROVA (1994; see also CAMIZ, ROVA 2001). In Fig. 4 two seals images, one with only one element repeated many times on several rows and the other with a complex structure of subpatterns are shown with the corresponding textual coding.

Unlike the classical use of textual analysis, where differences in style are considered important, in this case special care was devoted to constantly use the same form for the description of the same element or attitude. In this way, no information is available concerning the style of the image itself, since it is very difficult to include this information in the text. This could be coded in a separate way as a qualitative character, distinguishing several different pre-defined styles, as a different text describing some particular style features, or by using synonyms in the coding.

3) For the third level, the syntactical structure of the image must be taken into account. This is a complicate task and, up to now, rather difficult to implement. This problem raised a further study, carried out in CAMIZ *et al.* 1998, aiming at finding a suitable coding. This will be discussed in the following section.

7. BEYOND THE TEXTUAL CODING OF IMAGES: SYMBOLIC CODING

For the upper level study, that is the syntactical structure connecting the image subpatterns, an appropriate coding must be defined, able to describe the organisation of the elements composing the subpatterns. In CAMIZ *et al.* 1998 we used sequences of symbols to describe the relations among elements composing the image. The pattern was coded considering the presence of generic elements, together with their orientation and their spatial relation with other elements. A couple of parentheses encloses those sets of symbols that compose a subpattern. This allows to distinguish among *terminal* elements, i.e. the elements corresponding to the said symbols, and *non-terminal* ones, namely those corresponding to a set of symbols enclosed by a couple of parentheses, i.e. the subpatterns that form a subimage. We fixed several rules in the construction of sequences, in order to avoid ambiguities and get uniform all the corpus structure, in what concerned the precedence of the connections, the use of the parentheses, etc.

The symbols are chosen according to the kind of elements and/or attitudes the archaeologist had in mind and the relations are those among the elements and/or the subpatterns contained in the image: they are represented

	Elements		*Relations*
D	Main element right oriented	.	adjacent to
S	Main element left oriented	+	joined with, touching, attribute
X	Main element not oriented	*	interlaced with
F	Main element doubly oriented, main right	/	on
J	Main element doubly oriented, main left	v	on / under and by
		|	above
d	Secondary element right oriented	∩	into
s	Secondary element left oriented		
x	Secondary element not oriented	*Subpattern*	
f	Secondary element doubly oriented, main right	(	beginning
j	Secondary element doubly oriented, main left	)	end

Tab. 1 - The symbols used in the pattern description strings in Camiz *et al.* 1998.

by special symbols. In Tab. 1 the sequence of symbols used in Camiz *et al.* 1998, with their meaning, is shown[4]. It is to be noted that the subpatterns syntax is enclosed in parentheses. This produces the so-called *hierarchical sequences*, that are suitable to represent the complex structure of an image composed of a pattern of subpatterns, etc., so that the analysis methods developed can recognize them and take them into account when evaluating distances among sequences. In Fig. 4 the corresponding descriptive strings are shown of the two seals already described by the textual coding.

8. Exploratory analysis of archaeological data

Nowadays it is no longer necessarily a special coding to store information into a database, thanks to the most large dimensions of computers storage and the very sophisticated tools for the information retrieval. Instead, it is in the analysis of the stored data that the coding finds its rationale. We refer here mostly to the *exploratory data analysis* phase (Camiz 2000; Camiz, Rova 2001), since the further phases of a study may seldom be applied properly to archaeological finds. In particular, it is questionable to apply statistical inference to non-randomly selected archaeological material, and not even mathematical models seem appropriate. Indeed, no special requirements are necessary during the exploratory phase, so that exploratory data analysis has recently become the most used framework for a general investigation of

[4] "*On/under and by*" refers to the case in which an object lies aside another and partially covered by a part of it. "*On*" refers to a single element lying on top of another, while "*above*" refers to compositions on two or multiple rows.

collected data in any scientific environment, in particular where the construction of a mathematical model is very far from its conceivability. Exploratory data analysis revealed a most useful investigation tool, prior to any further study, and the recent studies in archaeology could take advantage of it too (BAXTER 1994).

The exploratory data analysis was first developed in the French school headed by J.P. Benzécri, who introduced the main analysis tools (BENZÉCRI *et al.* 1973-82). Nowadays, it is largely adopted all around the world. Recently it merged in the larger world of *Data Mining*, that is its "translation" in the Anglo-Saxon scientific framework. Indeed, the data mining originates from the computer scientists involved in the extraction of information and knowledge from the very large databases and data warehouses of the big enterprises, but the two terms seem really synonymous. There is an enormous literature on data mining: a quick search in the on-line bookstore www.amazon.com returned over 1400 titles, to be compared with the only 350 of *exploratory data analysis* (plus 130 of *analyse des données* in the french bookstore www.amazon.fr); just as a quote, consider KANTARDZIC (2002).

In the exploratory data analysis phase one defines a frame of reference for his/her work and the aims of his/her investigation. In this phase the treatment of the data aims at searching structures and relationships, in order to formulate some hypotheses. Collected data are thus submitted to specific (exploratory) tools, to recover as much synthesized information as possible, in order to reveal any existing data structure and, in particular, to see whether or not the research aims are reachable on the basis of the collected data.

The exploratory data analysis tools are able to reorganize the data, in order to reveal the structures that may exist. Such structures represent a way to synthesize the information contained in the data, since the exploratory data analysis aims at describing this information through a *strong synthesis*. In particular, this leads to describe which relations exist among the characters considered during the observation and which resemblance can be detected among the observed units. For such syntheses the *important* information is produced in graphical form. The hypothesis underlying the exploratory techniques, say their paradigm, is that the important information contained in the data may be extracted via some mathematical techniques, based on *association measures*. Two main graphical approaches are generally used, based on two different mathematical models:

- the *objects* are represented as *vectors* and *clouds of points* on an affine plane, forming scatter diagrams, useful to show the influence of characters on the observations and to find *factors* that best describe these influences;
- the *objects* are represented as *nodes* of a graph, whose *edges* represent *relations*; particular graphs used are dendrogram trees, useful to build data

taxonomies, thus showing structures and suggesting partitions, obtained by cutting the branches of the dendrograms.

Both approaches are useful to reveal the structures contained in the data, since they lead to the identification of:

- *factors*: the characters that best fit the objects diversity; these may be merely descriptive but they are also useful to represent the objects position in their respect; in many cases they can also be interpreted as the causes of the diversity.
- *classes*: building homogeneous classes of objects allows to consider the obtained partition as a structure of the considered set.

The search of factors and classes is usually performed using exploratory factor and cluster analyses respectively. The most known factor analysis techniques are Principal Components Analysis and Correspondence Analysis (Bry 1994; Lebart *et al.* 1995; Bolasco 1999), all aiming at representing both objects and characters as points of a geometrical space. Cluster analysis techniques may be either hierarchical or agglomerative (Anderberg 1973; Lebart *et al.* 1995; Bolasco 1999; Gordon 1999). The first technique builds a dendrogram that, by appropriate cutting, provides a complete hierarchy of encapsulated partitions, whereas the second builds only partitions in given number of classes.

Indeed, both techniques are useful in the archaeological studies, for the identification of main trends of variation of the samples under study and for the construction of classes of homogeneous objects, but it is in their synergy that a very good synthesis of information is obtained. In particular, this would be necessary if one suspects that the classification on the original data could be too heavily influenced by some characteristics of the coding. This may be the case of too many highly correlated characters, hierarchical coding, co-occurrence of populations with different structure, etc. According to the kind of data (qualitative, frequencies, quantitative) a different technique should be applied and different association indexes should be used for describing the suitable dissimilarities. For both the landmarks and the symbolic coding, special methods may be developed in both frames: in order to remain in the frame of the classical methods, it is sufficient to find a way to represent the dissimilarity among objects through a numerical value. Factor analysis would require that this value has the characteristics of a *distance*[5], but similar techniques are available for simple dissimilarities. So, once a coding is defined, it is necessary to define measures of association that can be computed among objects, based on the coding, having suitable mathematical properties, in order to use exploratory analysis tools.

[5] A *distance* between two objects is a numerical positive value independent on the order of the objects, valued zero if the two objects are identical, and such that it is less or equal to the sum of the distances of the two objects to a third.

Most of the studies discussed in this paper took advantage of the exploratory analysis techniques, either existing or developed on purpose: the way they were applied and the consequent results are reported in the quoted papers.

Sergio Camiz
Dipartimento di Matematica "Guido Castelnuovo"
Università degli Studi di Roma "La Sapienza"

Acknowledgements

I am most indebted to both Paola Moscati and Elena Rova, that followed the development of this paper, reviewed it, and suggested most of the improvements. Their friendly contribution is gratefully acknowledged.

REFERENCES

Adams W.J., Adams E.W. 1991, *Archaeological Typology and Practical Reality. A Dialectical Approach*, Cambridge, Cambridge University Press.

Anderberg M.R. 1973, *Cluster Analysis for Applications*, New York, Academic Press.

Andrenucci S. 1998, *A computer-aided approach to the classification and dating of ancient Egyptian artefacts*, in *XII Table Ronde Informatique & Egyptologie*, Utrecht.

Andrenucci M., Andrenucci S. 2002, *Statistical shape analysis for Egyptian scarab classification*, in *XIV Table Ronde Informatique and Egyptologie*, Pisa, Università di Pisa, Consorzio Pisa Ricerche, on Cd-rom.

Baxter M. 1994, *Exploratory Multivariate Analysis in Archaeology*, Edinburgh, Edinburgh University Press.

Bécue M., Haeusler L. 1995, *Vers une post-codification automatique*, in S. Bolasco, L. Lebart, A. Salem (eds.), *JADT 1995. III Giornate Internazionali di Analisi Statistica dei Dati Testuali*, Roma, CISU, 1, 35-42.

Benzécri J.P. *et al.* 1973-82, *L'analyse des données*, 2, Paris, Dunod.

Bolasco S. 1999, *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione*, Urbino, Carocci.

Bry X. 1994, *Analyses factoriales simples*, Paris, Economica.

Camiz S. 2000 (in press), *Exploratory 2- and 3-way Data Analysis and Applications*, Lecture Notes of TICMI, Tbilisi International Centre of Mathematics and Informatics, 2. Università di Roma La Sapienza, Dipartimento di Matematica "Guido Castelnuovo", preprint n. 14.

Camiz S., Ferrazza E. (in press), *L'immagine di Aiace nell'Italia antica*, «Archeologia e Calcolatori», 16.

Camiz S., Rova E. 1996, *Metodi di analisi per lo studio di un gruppo di sigilli cilindrici vicino-orientali e di altre immagini strutturate*, in P. Moscati (ed.), *III International Symposium on Computing and Archeology (Rome 1995)*, «Archeologia e Calcolatori», 7, 647-659.

Camiz S., Rova E. 2001, *Exploratory analyses of structured images: a test on different coding procedures and analysis methods*, «Archeologia e Calcolatori», 12, 7-46.

Camiz S., Rova E., Tulli V. 1998, *Exploratory analysis of images engraved on ancient Near-Eastern seals based on a distance among strings*, «Statistica», 58 (4), 669-689.

Camiz S., Venditti S. (in press), *Unsupervised and supervised classifications of Egyptian scarabs based on typology qualitative characters*, Paper presented at CAA 2004 (Prato).

Digard F. 1975, *Répertoire analytique des cylindres orientaux publiés dans les sources bibliographiques éparses (sur ordinateur)*, Paris, Éditions du CNRS.

Dryden I.L., Mardia K.V. 1998, *Statistical Shape Analysis*, New York, John Wiley & Sons.

Gardin J.-C. 1978, *Code pour l'analyse des ornements*, Paris, Centre National de la Recherche Scientifique.

Gardin J.C., Chevalier J., Christophe J., Salomé M.R. 1976, *Code pour l'analyse des formes de poteries*, Paris, Centre National de la Recherche Scientifique.

Gordon A.D. 1999, *Classification*, London, Chapman and Hall.

Hermon S., Niccolucci F. 2002, *Estimating subjectivity of typologists and typological classification with fuzzy logic*, in F. Djindjian, P. Moscati (eds.), *XIV UISPP Congress* (*Liège 2001*). *Proceedings of Commission IV Symposia. Data Management and Mathematical Methods in Archaeology*, «Archeologia e Calcolatori», 13, 217-231.

Kantardzic M. 2002, *Data Mining: Concepts, Models, Methods, and Algorithms*, New York, John Wiley & Sons.

Lebart L., Morineau A., Piron M. 1995, *Statistique exploratoire multidimensionnelle*, Paris, Dunod.

Lebart L., Salem A. 1994, *Statistique textuelle*, Paris, Dunod.

Moscati P. 1997a, *Ricerche informatiche sulle urne volterrane*, in AA.VV., *Atti del XIX Convegno di Studi Etruschi e Italici* (*Volterra 1995*), Firenze, Leo S. Olschki, 339-345.

Moscati P. 1997b, *Un gruppo di urne volterrane con rappresentazione del "viaggio agli inferi in carpentum"*, in AA.VV., *Etrusca et Italica. Scritti in ricordo di M. Pallottino*, Pisa-Roma, Istituti Editoriali e Poligrafici Internazionali, 403-423.

Moscati P. 2004, *Per la descrizione computerizzata delle urne volterrane. Problemi di formalizzazione*, in M. Fano Santi (ed.), *Studi di archeologia in onore di G. Traversari*, Roma, Giorgio Bretschneider, 647-655.

Rova E. 1994, *Ricerche sui sigilli a cilindro vicino-orientali del periodo di Uruk/Jemdet Nasr*, Orientis Antiqui Collectio, 20, Roma, Istituto per l'Oriente "C. Nallino".

Suter C.E. 1999, *Review of Rova 1994*, «Journal of Near Eastern Studies», 58, 47-53.

Tufnell O. 1984, *Studies on Scarabs Seals*, II. *Scarab Seals and their Contribution to History in the Early Second Millennium B.C.*, Warminster, Aris & Phillips.

Zadeh L.A. 1965, *Fuzzy sets and systems*, in J. Fox (ed.), *Systems Theory*, Brooklyn NY, Polytechnic Press, 29-37.

ABSTRACT

The problem of coding archaeological finds is discussed. The different items susceptible to coding are described according to the kind of information that must be collected. Some new coding techniques are described in particular: the landmarks technique, to be used for the shape analysis of *corpora* of finds all having a similar shape; the textual coding, useful for the study of images, once both the elements and attitudes and the sub-images composing the image are taken into account; a symbolic coding, to be used in the study of the syntactical structure of the images, describing the relations among items, regardless of the iconographical content. An overview of the exploratory analysis issues is given as conclusion.