XML Encoding of Archaeological Unstructured Data

Marco Crescioli
Unirel srl, Firenze

Andrea D'Andrea
Istituto Universitario Orientale, Napoli

Franco Niccolucci
Università di Firenze

Abstract: The use of databases to archive archaeological records is well-established since the very beginning of the diffusion of personal computers. Using DBMS with more or less complex features is thus nowadays common practice, even in school excavations and in the most conservative and computer-unaware contexts. In some cases, however, the internal complexity of records is badly managed by the rigid structure of database management software, which requires a repetitive structure and a high level of homogeneity among data. This is the case, for example, of historical documents, which need to be linked to archaeological sources in comprehensive historical archaeology investigations. Databases are of little utility also for research that spans over many years and needs to integrate recent database archives with previous excavation diaries. In general, whenever descriptive text documents have to be analyzed together with structured archives, comparison and overall search are laborious, if ever possible, since, as it is well known, fitting text documents into a relational database may determine a substantial loss of information.

Some recent pioneering work suggests that XML may be profitably used to encode historical texts, but as search engines are not yet easily available, this may appear to be only a wishful statement.

In this paper we suggest a simple way to create and manage an archive of archaeological texts and images and a search system. This follows a research line in which we try to improve the computer tools available to archaeologists both developing extensions to manage peculiar conditions (uncertain or unstructured information) and using inexpensive software and easy-to-use, web-savvy user interfaces. The archives created in this way may then be accessed and searched using a Web interface, as stand-alone, on a Local Network or over the Internet.

Introduction

As far as excavation databases are concerned, attention has recently been drawn to an increasing proliferation of proprietary solutions, mainly oriented to a "local" data structuring and unrelated to the goal of integrating information with other solutions. It is not casual, therefore, that many applications, created to give an answer to specific investigations, have been abandoned after the publication of the research results. Notwithstanding the positive trend produced by a large process of computer literacy that invested at different levels also this sector, there is no unified perspective that may hint a final destination where all these applications may converge. Such a condition considerably limits information sharing and re-use. Thus traditional paper publication is still the main channel of knowledge diffusion, often as an interpretation synthesis and only sporadically at data level.

A short and certainly partial survey on the most recent procedures (see, for instance Arroyo-Bishop 1999 – for a comment on this paper see below – Constantinidis 2001, D'Andrea and Niccolucci 2001, Lang 2000, Madsen 2001) shows that excavation systems have been as yet characterized by solutions implemented in such a way that porting to different case studies is not easy. Even when this limit was recognized as an obstacle to data diffusion, the discussion seems to mainly concern the creation of advanced data standards from the semantic viewpoint or at the meta-structure definition level. This is, for instance, Madsen's proposal (1999: GARD project; 2001: GUARD project), who considers a difficult task the integration of different systems as far as field information description and formalisation are concerned and thus thinks to be simpler to favour an integration of different solutions by means of homogenous formal structures.

Anyway, the setup of semantic or structural standards does not appear as urgent, so invites to define common description protocols risk remaining unheard. Archaeologists differ too much in sensibility, culture, geographic areas to achieve quickly a formal homogeneity as far as description criteria and field information organization are concerned. Paradoxically, the diffusion of front-end software end-user applications and authoring systems to create personalized solutions increased this particularism by giving directly the archaeologists the responsibility of planning their data models. So archaeologists build the information representation model following their data formalization schema and aiming the system implementation to a specific archaeological investigation. This happens even with good computer solutions and with procedures that guarantee uniform digital documentation and archival criteria, without sacrificing the information management and overall processing. But the conclusion deriving from these experiences is

incommunicability between different data management systems.

Between the exigency of unifying different solutions by means of standards or the creation of meta-structures, possibly it may be more useful to start reflecting and working on simplified data diffusion tools, i.e. solutions that that may use formats independent from particular software or operating systems, above all open-source. XML is such: as it is well-known, it is a meta-language based on public standards independent from the alternate fortunes or the commercial strategies of software producers. This technology uses text files and has an essential structure, so it will be able to guarantee the preservation and future use regardless new development and changes in information and technologies.

The availability of languages of the same family, aimed at the description of different data typologies already allows, moreover, levels of integration among data that would not be possible with the same flexibility using different tools. So, using different dialects of the XML family it is going to be possible not only to access on-line different archives, but also to integrate vector graphic information, for which the XML-based public standard SVG is already available, and spatial data for which GML (Geographic Markup Language) has been proposed as the XML description language of geo-referred objects.

This paper examines a first significant experiment we carried out in such a perspective. It is more than a simple hypertext-hypermedia application using the Internet interface to remotely and easily access data. Our project starts, on the contrary, from the need to convert a large database created with a proprietary application to integrate it, in a second stage, with unstructured information taken from reports published at the beginning of the twentieth century; data have then to be related to databases archiving specific object types as ceramics.

In order to avoid exporting all archives (both structured and unstructured ones) to the most recent release of a commercial package whatever, in an endless process to comply with the latest version of the same software (consider, for instance, the outcome of for distinct version of Office 95, 97, 2000 and XP in a short period, most of which show backward incompatibility for the Access DBMS format), we planned an created a meta-solution that could overcome the constraints imposed by the software market with no sacrifice on data, archives and existing information access. Thus we chose to implement the project with XML and to create a simple dynamic HTML interface to visualize and search alphanumeric and graphic data. With such a solution it is possible to use a browser, on the client side, to navigate among all these different data types and query the documentation using an appropriate search engine. It must be noticed that since XML tagging is a cumbersome job, the work is still in progress and as long as different documents are tagged there is a feedback to improve the structure of archived documents. Anyway, the principal database, created after modern excavations, is already available and may be consulted on-line and on a CD-ROM with no special software requirements.

Excavation data recording systems

Systems for archaeological excavation data recording vary from global solutions integrating alphanumeric and graphic data in the same environment to hypertext solutions addressing a simplified management of excavation data, with a remote access and easier and friendly GUI. A common tract is effectiveness as far as on-field data access is concerned. Moreover these system show a wide set of query functions and a hierarchical architecture. Most of them integrates data collected on field (stratigraphic units, drawings, photos) with laboratory-generated information (phase maps, joint maps, finds description).

In the wide diffusion of excavation systems, a specific application segment concerns the integration of both graphic and alphanumeric data within GIS systems: these are the so-called global solutions that recreate the information unity represented by the physical description of the layer-object with its spatial position but may also include data management and display or address particular inferential statistics procedures (see, for instance Djindjian 2000 with bibliography on intra-site artefact analysis). Other solutions do not take into account the spatial component and produce exclusively alphanumeric databases. Perhaps the wide diffusion of database software – consider, for instance, the wide availability of Access, distributed as part of the Microsoft Office suite, which made it the probably most used package for data management notwithstanding his limits – and its apparent easiness of use led to a vast use of file management systems and/or (quasi) relational databases.

A negative note that accompanies most of these solutions concerns the narrow-minded attitude shown by archaeologists toward similar experiences more than software development or computer methodology. Since these solutions are extremely personalized, as noticed before, neither from the computer nor from the archaeological point of view it is relevant to compare one's experience with other solutions, in order to verify the possibility of developing common procedures or to analyze problems met while processing data. It is sufficient to give a glance to bibliographical references of different papers to notice that most teams follow their research line with little concern for similar experiences. The most noticeable case is the paper by Arroyo-Bishop (1999) on the ArcheoData project, where all the twenty-two (!!) bibliographical quotations refer to papers by the same author, with no bibliographical reference to any other solution (or to anything else). Apart from such a self-referential attitude that automatically disqualifies itself and its author and thus deserves little or no attention, the general impression is that there is a widespread superficiality and little interest, if any, for different problems and solutions (a malicious reader might perhaps say little knowledge). It seems that there is little conscience (again, knowledge?) that computer procedures have their own methods and theory and a scientific approach requires comparison and generality: for some "ground-scrapers" the world ends at the border of their site, and whatever is done elsewhere is neither applicable nor important, and needs to be quoted only as a courtesy.

However, in general, the diffuse opinion that one's excavation is different from any other in the universe, which by the way leads to the uncontrolled proliferation of recording forms (helped, at least in Italy, by the obsolete official ones), is the psychological basis to create slightly different records and recording systems, on paper and in the archaeologist's mind before than in the computer. Even if there is something true in this attitude, it does not favour the growth and the diffusion of excavation data and above all it commits to traditional paper publication the diffusion of scientific results. On the contrary, our solution aims at easing data access (without necessarily having special software) and is based on a development strategy/ philosophy that avoids proprietary data formats and uses instead public domain ones to integrate different data types. So, in our perspective the goal did not focus on creating a multimedia product even if the system is oriented to hypertext navigation among various archives and documents. However, the solution nucleus consists of the use a technology aimed at structuring documents lacking a rigid structure and at establishing a dialogue among different document types, thus enriching the information content of the overall database.

To make this principle assumptions concrete, and test in practice the theory enounced above, we applied this approach to a case study. It is an excavation of Cumae, a site of the greatest archaeological interest.

The excavation of Cumae

Very little is known of the organisation of urban space in Cumae, the most ancient of West Greek colonies, founded in the second half of the 8th century BC (Fig. 1).

Since 1994 the Istituto Universitario Orientale of Naples in co-operation with the Soprintendenza Archeologica di Napoli e Caserta started again an archaeological excavation focused on the topography of the ancient town. The little data as yet known about the urban settlement and the walls came from excavations carried on in early 1900 or from rescue excavations. The city extension is known according to an hypothetical reconstruction of the wall enclosure, while archaeological witnesses of the urban settlement limit to a few monumental buildings of the Sannite (5th – 4th century BC) and the Roman period (3rd century BC . 2nd century AD). The road system is witnessed by a N-S axis that crossed at least four E-W roads, discovered during a limited rescue excavation performed to build modern sewers. The choice of investigation areas was therefore finalized to the attempt of reconstructing the road network, characterizing the function of spaces with the urban perimeter, reconstructing the wall enclosure and establishing its phase chronology.

Investigations as yet mainly focused on the area between the northern walls and the forum (Fig. 2).

They allowed the discovery of the development of fortifications in a zone which probably was a hinge between the public area and the northern strip, where perhaps a lagoon landing was placed (D'Agostino and Fratta 1995).

In the same area, some sondages had been excavated between the end of the 19th century and the beginning of the 20th: there are, unfortunately, few records of these investigations, and they are kept in the Soprintendenza archives or in few concise excavation reports published in early 1900. In most cases there is no indication about the precise sondage position, while information concerning ancient buildings still visible in the 19th century may be derived from several maps compiled during the 19th century in at least three different topographic surveys on the area (Fig. 3).

Data collected during the field investigation (over 3000 stratigraphic unit records and related documents as photos, drawings, materials, samples) were recorded using the Syslat program (Lattara 10).

The SYSLAT system: archaeological data

Originally conceived as a tool to record excavation data of the proto-historic site of Lattes (Montpellier – France), since 1984, when it was first experimented, SYSLAT (Système Lattes) has undergone several releases that progressively assimilated the suggestions deriving from use. Outside France, the system has been used for instance in some excavations in Italy (d'Agostino and Fratta 1995, Gastaldi 1998, Santoriello and Scelza 2001). The experimentation so created a portable system now provided with a wide set of personalization functions, which may configured according to the needs of any excavation.

SYSLAT is not only a mere digital adaptation in digital format to the stratigraphic excavation model formalized by Harris. Even if it is based on the fundamental principles of stratigraphy, it is a tool for a "guided" organization of field data (from action recording to graphical and photographic documentation and sample collection) and for an integrated management of all data, including statistic-quantitative analysis of materials.

Created using Hypercard, one of the first authoring tools, SYSLAT represents a great container in which appropriate scripts allow to access recorded information: from stratigraphic units to "fact" and "set" records, from the photos to the graphic archive, from quantification records of materials to the typology of individual ceramic finds and so on. The global archive is structured in five different modules (terrain, objects, samples, documentation, utility) from which additional hierarchical sub-levels may be accessed.

The system includes a wide iconographic dictionary to classify ceramics (Lattara 6), and several vocabularies guide the operator in compiling forms. SYSLAT has also a module to personalize the database and to add new definitions to glossaries and dictionaries. With the latest version, 3.1, dating to 1997, the system has been exploited until reaching the limits of Hypercard: anyway, it has no multi-user version and cannot visualize spatial data. The lack of these features makes it difficult to plan further developments for the application: moreover, the fact of being limited to a specific platform does not allow porting archives to other operating system. Restructuring the system is now still more necessary since Hypercard, the software on which SYSLAT is based, will no more be supported on the current and on future versions of the Apple operating system.

Addressing the conversion of our SYSLAT archives, we had to face several problems:

a) since the original system was based on a hypertext model, descriptive fields were encouraged and these had been intensively used by the compiler; so a quantity of useful information was hidden inside the fields;
b) updating the software was unavailable, as mentioned above;
c) migrating to a different database management system would have solved the obsolescence problem only temporarily;
d) converting the archives to a RDBMS and extracting the information buried into descriptive fields would have been difficult, if not impossible; the lossy alternative would have been to leave it unchanged, renouncing to retrieve this hidden information with usual RDBMS searches.

Thus we decided to adopt a solution based on XML.

The conversion of digital archives

For the conversion, the fields included in the SYSLAT form were grouped distinguishing the ones containing information from the ones used by the system to manage the archive (a characteristic of Hypercard software). Only the latter fields were converted, associating to each one of them an XML semantic component, with identical content. Moreover, it was noticed that some elements (for instance the code of stratigraphic units) were often present inside the fields, in particular within the "Description" and "Interpretation" fields: so they were evidenced (i.e. tagged as such) in the final document, using as much as possible an automatic procedure based on the easy formal recognition of these parts inside a text (e.g. a numeric value preceded by the code US).

With such objectives, that is automatic transfer of "data" fields and automatic recognition of some references inside the fields, conversion routines written in Hypertalk, the programming language of Hypercard, created the XML files corresponding to the original records.

Having thus "saved" the database, now available as tagged text files, it may be published on Internet or consulted by the research team with no particular software or computer, but simply using an Internet browser as Internet Explorer or Netscape. Since only the latest versions of the browser are XML-compliant, to make access easier XML documents were batch converted to static HTML pages using a stylesheet and a conversion utility. The XSLT (a transformation language of the XML family) stylesheet that was created for this purpose automatically generates the link to the records that are referenced inside the fields (stratigraphic units, facts, materials, photos, drawings and maps) to allow easy navigation; different pages may also be accessed from an index page.

This solution is a temporary choice, since it has the evident drawback of requiring batch processing for data update; it has been adopted since it guarantees an immediate access to information, even with a minimal set of software resources (data can always be accessed from http://localhost) and on standalone workstations. More advanced searches require, on the contrary, the availability, at least on the network, of search engines and an XML-compliant HTTPD server as Cocoon, a freely available software which is a part of the Apache Open Source project. Moreover, as one can expect, the Microsoft browser has a tiny non-standard requirement that is easy to take into account but may make IE-compatible documents incompatible with anything else. So a more effective solution will leave the burden of XML-to-HTML conversion to the server and then send the user's browser a standard HTML document: thus pages will be dynamically created on user's demand using the XSLT stylesheet and will take into account all the variations on data. Integration with graphic vector information is already available via SVG (an XML-family description language for vector graphics) using a plug-in freely distributed by Adobe, one of the sponsors of this language; unfortunately, possibly due to Hypercard limitations original drawings were created on paper and are still such.

Other archives

In the past many archaeological investigations had been carried out in the same area as 1994 excavation. Notices about the former are, however, very scarce: they consist of a few memories and documents conserved in the archives of the Soprintendenza Archeologica and some reports concerning sondages of the beginning of the last century. Particularly interesting for the knowledge of the stratigraphy appears the information given by Maraglino (1906): He described some 1903 sondages documenting settlements in the area since the Greek period until the Roman one. The interest of this source consists not only of the rich stratigraphic information, comparable with data acquired from the most recent excavations, but also of some topographic reference, useful for an approximate placement of the sondages. They may be positioned by means of topographic references inserted into the published text that may easily be recognized in a contemporary cadaster map. Reading Maraglino's report, the stratigraphy of the area emerges clearly: it is characterized by superimposition layers under which remains of ancient buildings, overlapping preceding structures, were discovered. Of each layer the source gives a summary description, the deepness and the description of ceramics, which are relevant to date the stratigraphic unit.

Here is an example of Maraglino's text, translated into English:

> On the end of 1903 prof. Innocenzo Dall'Osso, looking after the excavations on the Correale farm-land owned by Signor Ernesto Osta, lawyer, did not omit to do some surveys around the mount of Cumae, to recognize and study the ancient features of such places.
>
> In one on these surveys, walking on the Gigante farm-land, placed some 300 m. South-West of the Acropolis, he happened to notice, in the recently moved ground, some fragments of black paste with an Italic graffiti decoration. This was a very good hint for him to suppose that there

could be settled the village, or at least the necropolis, of primitive inhabitants, occupying those place before the arrival of Greek colonists, thus preceding the colonization of Cumae.

Fortunately the owner of the Gigante farm-land is Signor Enrico Origlia, a person well educated and passionate for antiquity.

Thus prof. Dall'Osso easily persuaded Signor Origlia to pay for a sondage, which was directed by prof. Dall'Osso himself, with three workers; for example, since the Gigante farm-land was planted with a vine-yard, as almost everywhere in that place, he had to choose a point where the vines, being less dense, left a free space of about two square meters. After removing the humus layer, of the usual thickness of 30 cm., a yellowish soil followed, evidently of alluvial nature, more than one meter thick; then there appeared remains of ancient dwellings, with walls built in big blocks of yellow tuff, a local rock. Basing on numerous fragments of black painted ceramic, of Italic Greek provenance, found among the ruins, there were sufficient hints to refer the remains to the Greek-Roman period or to the Sannite domination.

Below them, however, there were remains of foundations of other more ancient buildings, among which some fragments of proto-Corinthian vase were found, undoubtedly belonging to an archaic period of pure ellenism. This first excavation shows that in that place a new population had overlapped on a preceding one and then still another until the Roman period.

As it can be seen, the text includes very useful information about early discoveries, and since the place-names appear on historical maps the settlement described in the report can be georeferenced, though imprecisely, and the information can be spatially compared with other data (Fig. 4).

To be able to perform structured searches in the text and compare results with the modern database, it is however necessary to adequately structure the source, using an element set as compatible as possible with the one used for the modern database.

The need of comparing Maraglino's information induced us to design a solution integrating all data (structured and unstructured) available for the studied area.

A further enrichment of our global solution may come from the integration within the system of another database specific to the study of ceramics coming from the latest investigations, created in order to have a precise chronology grid of recent and previous findings and to have a precise quantification of ceramic fragments for each layer, grouped by production classes and typology. This database had been implemented using Microsoft Access. It consists of two tables, linked by a one-to-many relation based on the stratigraphic unit code. The fragment table contains information on each fragment as part of the vase, number of fragments, category, production class, typology.

The multiplicity of available data (collected on field, in the laboratory, from archives) required thus a new computer solution and as already mentioned XML was the chosen one.

In particular, a great attention has been paid to the markup process of Maraglino's text since a correct definition of a translation model for similar documents was determinant for the subsequent integration of data extracted from excavation diaries, archive memories and generic texts.

It should be clear at this point that adopting a traditional technology, as relational databases, would probably imply a loss of information due to the necessity of using a common framework for rather different information structures, in particular historical documents as the above mentioned, which are usually in a discursive format and do not easy comply with database structure requirements.

Below there is a sample of marked text. It does not correspond to actual markup, which is done using the original Italian source and meaningful Italian words as tags. So this translation is given here only for the sake of clarity and to make it understandable to an international audience.

<*report*>On the end of <*year*>1903</*year*> <*responsible*>prof. Innocenzo Dall'Osso </*responsible*>, <*introduction*>looking after the excavations on the <*place_name*> Correale</*place_name*> farm-land owned by Signor <*owner*>Ernesto Osta<*owner*>, lawyer, did not omit to do some surveys around the mount of Cumae, to recognize and study the ancient features of such places.</*introduction*>

In one on these surveys, <*place_description*>walking on the <*place_name*> Gigante </*place_name*> farm-land, placed some 300 m. South-West of the Acropolis</*place_description*>, he happened to notice, <*layer level="0"*>in the recently moved ground, some <*finds type="ceramics"*>fragments of black paste with an Italic graffiti decoration</*finds*></*layer*>. This was a very good hint for him to suppose that <*interpretation*>there could be settled the village, or at least the necropolis, of primitive inhabitants, occupying those place before the arrival of Greek colonists, thus preceding the colonization of Cumae</*interpretation*>.

Fortunately the owner of the <*place_name*>Gigante</*place_name*> farm-land was Signor <*owner*>Enrico Origlia</*owner*>, a person well educated and passionate for antiquity.

Thus prof. Dall'Osso easily persuaded Signor Origlia to pay for a sondage, which was directed

by <*director*>prof. Dall'Osso</*director*> himself, with <*resources*>three workers</*resources*>; for example, since <*sector*>the Gigante farm-land was planted with a vine-yard, as almost everywhere in that place, he had to choose a point where the vines, being less dense, left a free space of about <*dimensions size="2" unit="m2"*>two square meters</*dimensions*></*sector*>. After removing the <*layer order="1"*>humus layer, of the usual thickness of <*layer_depth size="0.3"*>30 cm. </*layer_depth*><*layer*/>, <*layer order="2"*>a <*color*>yellowish</*color*> soil followed, evidently of <*geology type="alluvial"*>alluvial nature</*geology*> more than <*layer_depth size="1"*>one meter thick</*layer_depth*></*layer*>; then <*layer order="3"*>there appeared <*layer_intepretation*>remains of ancient dwellings, with <*layer_description*> walls built in big blocks of yellow tuff, a local rock</*layer_description*> </*layer_interpretation*>. Basing on <*finds type="ceramics"*>numerous fragments of black painted ceramic</*finds*>, <*finds_interpretation*>of Italic Greek provenance</*finds_interpretation*>, found among the ruins, there were sufficient hints to refer the remains to the <*period*>Greek-Roman period</*period*> or to the <*period*>Sannite domination </*period*></*layer*>.

Below them, however, <*layer order="4"*>there were <*layer_intepretation*> <*layer_description*>remains of foundations of other more ancient buildings </*layer_description*>, among which <*finds type="ceramics"*>some fragments of proto-Corinthian vases</*finds*> were found, undoubtedly <*finds_interpretation*>belonging to an archaic period of pure ellenism</*finds_interpretation*></*layer*>. This first excavation shows that in that place <*site_interpretation*>a new population had overlapped over a preceding one and then still another until the Roman period</*site_interpretation*>. </*report*>

As it can be noticed, only very wide concepts have been used to tag the text. However, they allow a much more effective search when combined with the other databases.

The search engine

Since the work is currently in progress, several search engines are being tested. Moreover, while the markup of historical sources goes on, the element list (what is technically called the DTD), is often revised. So we created a very simple query interface, which can be easily modified according to variations in the DTD or different search engines.

In general, the user is presented with a HTML dynamic form where query conditions may be set. At present, the lists of searchable elements are kept separate for different document types: this is due to the above mentioned variability of the elements used to tag historical sources. Thus the list is kept in a file, which is read by the form and displayed as a combo-box. It is planned, however, to offer hints for the equivalence of elements within different element types and to propose a more limited number of search choices to a generic user.

The form, written in PHP, prepares the query strings as needed and passes them to the search engine. The result is then displayed by another page.

As far as the query involves only a database, this process is straightforward; using a native XML database makes everything a little more complicate, because such engines are still largely experimental and their interfaces are awkward.

So after the user has filled the form fields with the search parameters and the form has been submitted, the values are extracted from the fields and a command is compiled, adding the necessary codes, that is "switches" or element names. This step is transparent to the user, who needs, however, to know the data structure to choose in a correct way where the text string has to searched by the program. There is a number of options that have to be chosen: for instance one may search a specific content within an element, regardless the number of nested sub-elements occurring between the specified element and the required content, or the latter may be required to be contained directly in the indicated element. The content of attributes may be specified as a search parameter as well, thus giving a great flexibility. The equivalent of relational join operations is, on the contrary, sometimes difficult to express: a search as "construction materials used in layers containing a determined type of ceramics" is much more complicate to write than using, for instance, SQL, as the search logic is to move on the tree representing the element structure of documents, obtaining as query result a set of nodes of this tree.

The best operational results have so far been obtained using sgrep, a structured text search program developed at the University of Helsinki by Jani Jaakkola and Pekka Kilpelainen and freely available at http://www.cs.helsinki.fi/~jjaakkol/sgrep.html. While this program is very quick and effective, it uses a structure based on text "regions", which is completely different from XML nested elements philosophy, however this requires only translating from one metaphor to another and back. A program more compliant with XML specifications, in particular with XQuery, would make things simpler. Recently the beta version of an XML native database, named dbXML, has been released; it looks very promising and it is being tested, with provisional results that might be improved if a better documentation on inexplicable error messages were provided with the software.

To illustrate the procedure from the user point of view, we include a sample query form (Fig. 5).

It must be underlined that the one shown is not the real form, which is written in Italian and uses Italian names for the elements; it is just an English translation prepared for this paper,

to be more intelligible to an international audience. Also the results are not the real ones, which would be in Italian as well: the real results have been translated and a specimen has been prepared. In summary, the illustration shown here is fake, but it is aimed at giving a non-Italian-speaking reader an idea of the appearance of the system to actual users. Moreover, as stated before, the form frequently changes according to variations in the DTD of historical sources, or to experiment associations among different element structures.

Conclusions and further research

As previously stated, what we present here is work in progress: it needs refinement and, above all, to prove its effectiveness it requires to turn from the present experimental status into an operational one. However, even in the present form it suggests some considerations.

Introducing XML data encoding in the humanities, as it is proposed in the present paper, might have substantial, and positive, consequences beyond the ones herein described. The Caere project (Moscati, Mariotti and Limata 1999) first introduced markup languages in archaeological data recording, with a remarkable intuition and very interesting results. It still used SGML as a reference meta-language, instead of XML, a choice which might have possible advantages but some certain drawbacks (Internet compliance first of all), and might therefore be argued. However, if incompatible SGML features are not used, as it seems from the published reports, there is practically no difference and this paper stands for its authoritative support to markup solutions.

The potential impact of XML use in human sciences and particularly in archaeology has already been evidenced in workshops (Niccolucci forthcoming) and is proving its importance as far as another related field, 3D modelling and Virtual Reality applications, is concerned (Cantone 2002; Niccolucci and Cantone forthcoming). A comprehensive and authoritative review of the importance of XML in metadata description and in general, in archaeological computing is also available (Ryan forthcoming).

Taking only archaeological databases into account, as noticed above the comparison and interchange between different investigations is still low and requires a long and fatiguing conversion among different data formats, which would be greatly facilitated by a common paradigm as XML.

Preserving digital archives is another important issue, and using proprietary data formats forces repositories (see e.g. ADS) to maintain different, and sometimes obsolete, DBMS formats to assure that data are still accessible and/or to convert them to ASCII delimited text (Richards and Robinson 2000). On the contrary, a simple and common text format, together with metadata description, would allow data access by means of current data management software and would not suffer from obsolescence.

Moreover, present trends in public administration using ICT (the so-called e-government) will lead in a near future to a vast amount of documents in electronic format, most probably XML-encoded: this seems the choice both of the Italian government [DPCM 2000] and of other European states. Thus administrative documents, for instance in Italy the administrative forms adopted by the Ministry of Cultural Heritage for monuments and archaeological excavations, could become interoperable with those produced by scientific research.
So XML might not only improve the efficiency of scientific research in the archaeological field, but also positively affect the management of cultural resources and perhaps fill the still existing gap between scientific research and administrative management, endowing both of a very powerful computer tool.

Bibliography

ADS Archaeology Data Service http://ads.ahds.ac.uk

Arroyo-Bishop D. 1999. From earth to cyberspace: the unforeseen evolution, *Archeologia e Calcolatori* 10, 7-16.

Cantone F. 2002. 3D Standards for Scientific Communication, this volume

Cantone F. and Niccolucci F., forthcoming, Legend and virtual reconstruction: Porsenna's mausoleum in X3D. Paper to be presented at the Web3D 2002 Conference.

Constantinidis D. 2001. Introspective Sitescaping with GIS, in Z. Stancic AND T. Veljanovski (eds.), *Computing Archaeology for Understanding the Past. CAA 2000. Computer Applications and Quantitative Methods in Archeology,* Proceedings of the 28th Conference, Ljubliana April 2000, BAR International Series 931, Oxford: Archaeopress, 165-172.

D'Agostino B.and Fratta F. 1995. Gli scavi dell'IUO a Cuma negli anni 10994-1995, *AION ArchStAnt* n.s. 2, 201-209.

D'Andrea A.and Niccolucci F. (eds.). 2001. *Atti del Primo Workshop nazionale di archeologia computazionale*, Napoli-Firenze 1999, Firenze: All'Insegna del Giglio.

Djindjian F. 2000. Artefact Analysis, in in Z. Stancic and T. Veljanovski (eds.), *Computing Archaeology for Understanding the Past. CAA 2000. Computer Applications and Quantitative Methods in Archeology,* Proceedings of the 28th Conference, Ljubliana April 2000, BAR International Series 931, Oxford: Archaeopress, 41-52.

DPCM. 2000. *Decreto del Presidente del Consiglio dei Ministri 31 ottobre 2000 - Regole tecniche per il protocollo informatico di cui al decreto del Presidente della Repubblica 20 ottobre 1998, n. 428* (Decree of the President of the Council of Ministers of Italy, 31 October 2000 - Technical rules concerning the computer protocol according to the Decree of the President of the Republic, 20 October 1998, nr. 428), art. 18, available on the Official Journal of the Republic of Italy and at http://www.aipa.it/servizi[3/normativa[4/leggi[1/dpcm311000.asp.

Gastaldi P. 1998. (ed.), *Studi su Chiusi arcaica*, Annali di

Archeologia e Storia Antica, Nuova Serie n. 5, Istituto Universitario Orientale, Napoli.

Gray J. and Walford K. 1999. One good site deserves another: electronic publishing in field archaeology, in Internet Archaeology, 7.
http://intarch.ac.uk/journal/issue7/gray_toc.html.

Madsen T. 1999. Coping with complexity.Towards a formalised methodology of contextual archaeology, *Archeologia e Calcolatori* 10, 125-144.

Madsen T. 2001. Transforming Diversity into Uniformity – Experiments with Meta-structures for Database Recording, in Z. Stancic and T. Veljanovski (eds.), *Computing Archaeology for Understanding the Past. CAA 2000. Computer Applications and Quantitative Methods in Archeology,* Proceedings of the 28th Conference, Ljubliana April 2000, BAR International Series 931, Oxford: Archaeopress, 101-105.

Maraglino V. 1906. *Cuma e gli ultimi scavi*, Memoria Letta alla Reale Accademia di Archeologia, Lettere e Belle Arti di Napoli, Stabilimento Tipografico della Regia Università di Napoli, A.Tessitore e C.: Napoli.

Moscati P., Mariotti S. and Limata B. 1999. Il "progetto Caere": un esempio di informatizzazione dei diari di scavo, in *Archeologia e Calcolatori* 10, 165-188.

Lang N. 2000. *Beyond the Map: harmonising research and Cultural Resource Management*, in G. Lock (ed.), *Beyond the Map: Archaeology and Spatial Technologies*, NATO Science Series, vol. 321, IOS Press, 2000, Amsterdam,, 214-228.

Lattara 6 = Py M. (ed.). 1996. Dicocer, Lattes.

Lattara 10 = Py M. (ed.). 1997. *Syslat 3.1. Système d'Information Archéologiques. Manuel de référence*, Lattes.

Niccolucci F., forthcoming, *Dalla fonte alla rete: applicazioni di XML alla ricerca storica e archeologica* (From the source to the net: XML applications to historical and archaeological research) Proceedings of the "XML Workshop", University of Florence, March 2001.

Richards J. and Robinson D. (eds.). 2000. *Digital Archives from Excavation and Fieldwork: A guide to Good Practice.* Published for Archaeology Data Service. Oxford: Oxbow Books.

Ryan N. S., forthcoming, Documenting and Validating Virtual Archaeology, in *Archeologia e Calcolatori* 12.

Santoriello A. and Scelza A. 2001. Un sistema informativo archeologico: l'applicazione del Syslat a Fratte di Salerno, in D'Andrea and Niccolucci 2001.

All Internet references have been checked on October 10th, 2001.

Figures

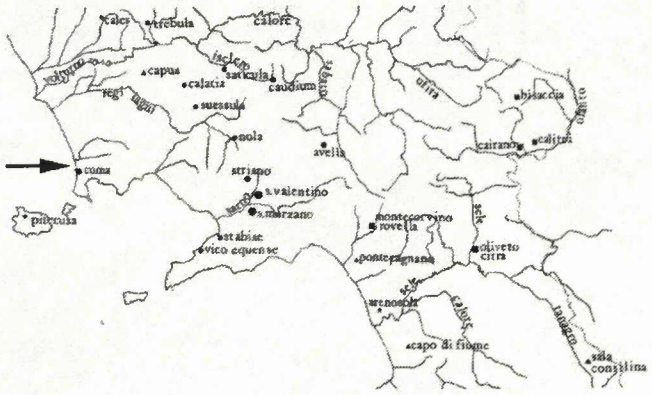

Figure 1: Cumae in Campania (Italy)

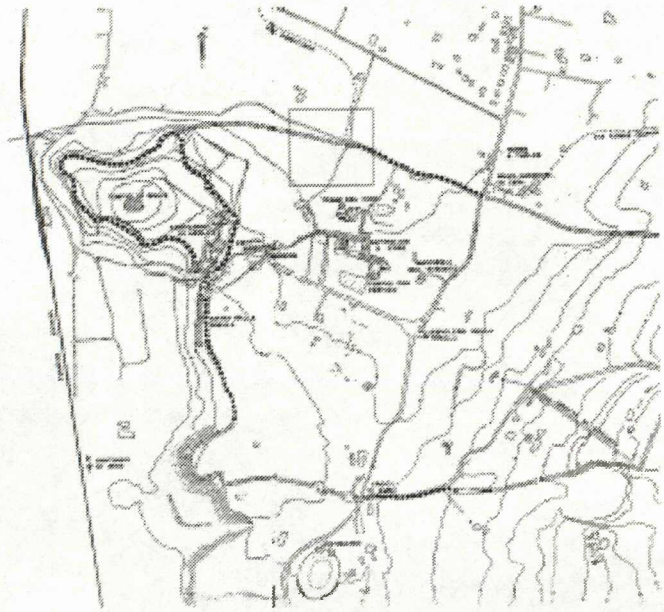

Figure 2: Archaeological remains of the ancient town of Cuma. The black square shows the localisation of the recent investigations carried out by Istituto Universitari Orientale on the area between the northern walls and the forum.

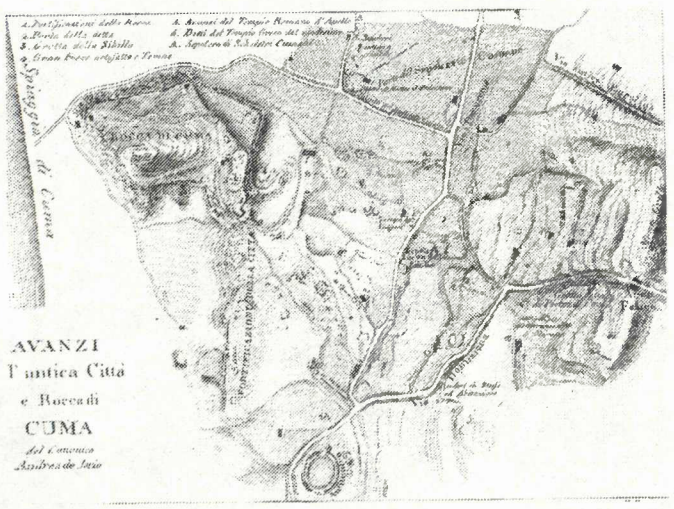

Figure 3: An example of historical map of Cumae (De Jorio 1822).

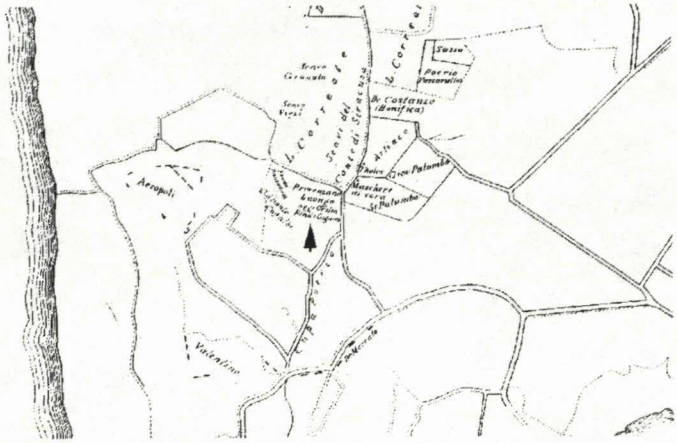

Figure 4: An historical cadaster map showing the farm-land described in the Maraglino's report.

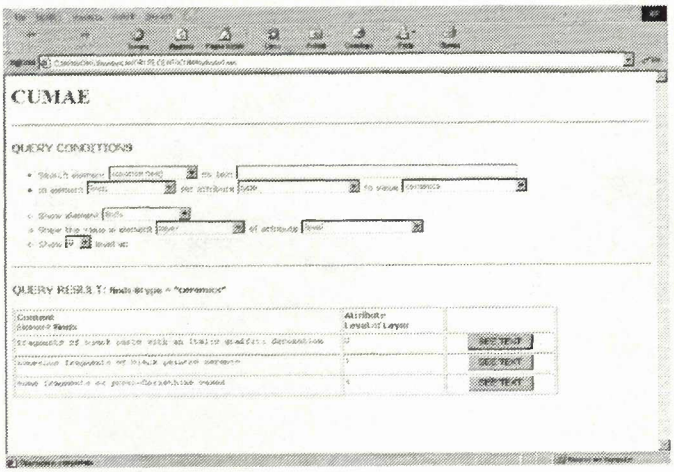

Figure 5: A sample query carried out using the search engine.