Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

# An XML-based information model for archaeological pottery

LIU De-zhi (刘德智)[†1], RAZDAN Anshuman[2], SIMON Arleyn[2], BAE Myungsoo[2]

([1]*School of Computer Science, Zhejiang University, Hangzhou 310027, China*)
([2]*Partnership for Research in Spatial Modelling, Arizona State University, Tempe, AZ 85281, USA*)
[†]E-mail: liudezhi@zjuem.zju.edu.cn

Received Sep. 29, 2004; revision accepted Nov. 29, 2004

**Abstract:** An information model is defined to support sharing scientific information on Web for archaeological pottery. Apart from non-shape information, such as age, material, etc., the model also consists of shape information and shape feature information. Shape information is collected by Lasers Scanner and geometric modelling techniques. Feature information is generated from shape information via feature extracting techniques. The model is used in an integrated storage, archival, and sketch-based query and retrieval system for 3D objects, native American ceramic vessels. A novel aspect of the information model is that it is totally implemented with XML, and is designed for Web-based visual query and storage application.

**Key words:** Geometric modelling, Feature extraction, XML, Content-based 3D search
**doi:**10.1631/jzus.2005.A0447 **Document code:** A **CLC number:** TP391

## INTRODUCTION

There is growing consensus among computational scientists that observational data, result of computation and other forms of information produced by an individual or a research group need to be shared and used by other authorized groups across the world through the entire life cycle of the information (Williams, 1998; Rowe *et al*., 2001). The Web has revolutionized the electronic publication of data. It has relied primarily on HTML that emphasizes a hypertext document approach. More recently, Extensible Markup Language (XML), although originally a document mark-up language, is promoting an approach more focused on data exchange. XML is a set of rules for defining semantic tags that break a document into parts and identify the different parts of the document. It is a meta-markup language that defines a syntax used to define other domain-specific, semantic, structured markup languages (XML, 2004).

As for archaeology, three-dimensional knowledge, including shape information and feature information are very important in the archaeological research. For example, archaeologists study the 3D form of native American pottery to characterize the development of cultures. Quantitative methods of reasoning about the shape of a vessel are becoming far more powerful than was possible when vessel shape was first given a mathematical treatment by Birkhoff (1933). Significant research effort was devoted to identify and build upon the descriptive and cataloging standards that had been used to describe ceramic artifacts (Rice, 1987; Shepard, 1976; Staudek, 1999). However, traditional measurement tools, such as calipers, rulers and hand drawn sketches are subjective and prone to inaccuracies. Recent theoretical and technological breakthroughs in mathematical modelling of 3D data and data-capturing techniques bring a new opportunity to advance quantitative analysis of archaeological vessels. The research problems are twofold: (1) How to model the artifacts to permit more accurate representation and measurement; and (2) How to create a system to catalog, query, search, retrieve, and display the information to satisfy the demand of sharing information. In this paper, we

[*]Project (No. IIS-9980166) supported by the National Natural Science Foundation of America

present an information model consisting of shape information and feature information for archaeological vessels. Shape information is collected from 3D Laser Scanners and geometric modelling techniques. Feature information is generated from shape information via feature extracting techniques. A novel aspect of the information model is that it is totally implemented with XML, and is propitious to a Web-based storage and query application.

## ANALYSIS OF ARCHAEOLOGICAL POTTERY

According to archaeological definition (Birkhoff, 1933) there are four kinds of feature points to calculate dimensions and proportions of vessels. They are End Points (EPs), Points of Vertical Tangency (VTs), Inflection Points (IPs) and Corner Points (CPs) found on the vertical profile curve of a vessel (Fig.1).

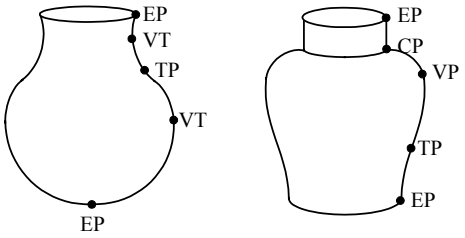


**Fig.1 Feature points of vessel profile curves**

Four features are common to all vessels: (1) Orifice: the opening of the vessel, the minimum diameter of the opening, may be the same as the rim, or below the rim; (2) Rim: the finished edge of the top or opening of the vessel. It may or may not be the same as the orifice. It may have a larger diameter; (3) Body: the form of the vessel below the orifice and above the base; (4) Base: the bottom of the vessel, portion upon which it rests, or sits on a surface. The base may be convex, flat, or concave, or a combination of these (Fig.2).

From the above definition for characteristic points and common features for all vessels, we can formalize feature representation of vessels as follows:

<Point Feature>:=<End Point Feature>
|<Point of Vertical Tangency Feature>
|<Inflection Point Feature>
|<Corner Point Feature>;
<Curve Feature>:=<Rim Curve Feature>|<Orifice Curve Feature>|<Base Curve Feature>;
<Rim Curve Feature>:=<End Point Feature><End Point Feature>;
<Orifice Curve Feature>:=<Corner Point Feature><Corner Point Feature>;
<Base Curve Feature>:=<End Point Feature>
|<End Point Feature><End Point Feature>
<Region Feature>:=<Neck Region Feature>
|<Body Region Feature>
|<Base Region Feature>;
<Neck Region Feature>:=<Rim Curve Feature><Orifice Curve Feature>;
<Body Region Feature>:=<Orifice Curve Feature><Base Curve Feature>;
<Base Region Feature>:=<Base Curve Feature>;
<Volume Feature>:=<Unrestricted Volume Feature>
|<Restricted Volume Feature>.

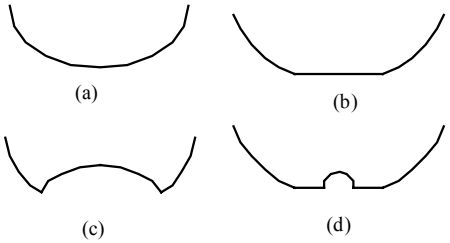


**Fig.2 Four kinds of bases. (a) Convex base; (b) Flat base with zero curvature; (c) Concave based; (d) Composite base**

## AN XML-BASED INFORMATION MODEL

We design an information model for archaeological vessels for storage and query application. Apart from the non-shape information, such as age, material, etc., the powerful information model for vessels includes shape feature information. It is easy to embed the non-shape information into XML-based information models. Therefore, our research focuses on the acquisition and representation of shape information. Shape feature information consists of shape raw data, curve/surface modelling information, and shape higher-level information, feature information. Shape raw data of archaeological vessels are 3D triangulated meshes composed of points, edges and triangles. They are collected from scanning vessels by 3D Laser Scanners. Curve and surface modelling information is generated from shape raw data via geometric modelling techniques. Feature information is extracted from geometric models, and is organized

according to formal description. The relationship of three levels of the information model is described in Fig.3.

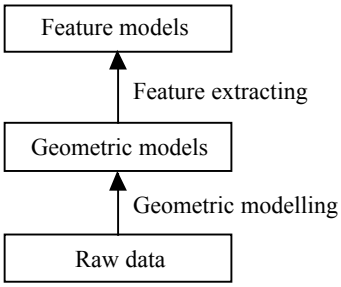


**Fig.3 Three levels of the information model**

**Shape raw data**

Shape raw data are collected from scanning archaeological vessels by 3D Laser Scanners, and are represented by polygonal meshes consisting of faces, edges and vertices and their connections. Raw data are used for further shape analysis of pottery.

Raw data in our information model are treated as binary large objects (BLOBs) since outcomes of most commercial 3D Laser Scanners are in the format of PLY files, which are binary (PLY, 2004). The PLY file format is a simple object description designed as a convenient format for researchers who work with polygonal models. However, XML does not support binary data, and it is "illegal" to embed BLOBs into XML documents directly. There are three methods to solve the problem: (1) CDATA method; (2) Multipurpose Internet Mail Extensions (MIME) method; and (3) BASE 64 en/decoding method. BASE 64 en/decoding method is adopted by most applications. We use BASE 64 en/decoding method to encode binary data before we embed the data into XML files.

An arbitrary bit stream encoded in the BASE 64 can be specified in an XML document as the content of an element, as long as any special characters such as "<" are represented as entities ("<"). An application reading the document would need to look for the element that contains the binary data, and decode the BASE 64 string to recover the original binary stream. That is fine for two cooperating applications, but not all applications will be able to recognize which elements should have BASE 64 encoded binary data, which have hex-encoded binary data, or which simply have character strings. It would be better if the data were self-describing.

For this reason, the XML-Data paper proposed a "dt" attribute that would allow embedding binary data into XML documents like:

```
<stuff dt:dt ="binary.base64" >
84592gv8Z53815Zb82bA68g
</stuff>
```

This example signals to other applications that the contents are binary stream encoded and use the BASE 64 notation.

**Geometric information**

Geometric information for archaeological vessels includes curve models for profile curves and surface models for reconstruction results. Geometric information is generated from raw data via geometric modelling techniques.

One way of representing or modelling surfaces is via parametric surfaces such as B-Spline surfaces or Non-Uniform Rational B-Spline (NURB) surfaces. We use such representation to enable us to rebuild models, analyze properties such as curvatures, make quantitative measurements as well as "repair" incomplete models. A NURB surface can be represented as

$$\boldsymbol{P}(u,v)=\frac{\sum_{i=0}^{m}\sum_{j=0}^{n}w_{i,j}\boldsymbol{d}_{i,j}N_{i,k}(u)N_{j,l}(v)}{\sum_{i=0}^{m}\sum_{j=0}^{n}w_{i,j}N_{i,k}(u)N_{j,l}(v)} \quad (1)$$

where $\boldsymbol{d}_{i,j}$, $i$=0, 1, …, $m$; $j$=0, 1, …, $n$ are control points; $w_{i,j}$ are weights; $N_{i,k}(u)$ and $N_{j,l}(v)$ are B-Spline basis functions. When weights equal 1.0, it reduces to a non-uniform B-Spline surface.

NURB surface is generated by fitting points of raw data with least squares approximation (Bae, 1999; Farin, 1996). Degrees, control point coordinates, weights, knot vectors are stored in information models because they uniquely determine a NURB surface.

In order to generate the reasonable surface modes from unorganized triangle meshes, the proper parameterization must be taken into account. Most vessels in our research are round and symmetrical. So the parameterization can be done by projecting mesh points onto a cylinder. For the $u$-direction, we obtain the angle, $\theta$, where $0\leq\theta<2\pi$ for each point, then map

to the range from 0 to 1. For the *v*-direction, we obtain the distance for each point from the bottom of the vessel, then normalize the range from 0 to 1. However, the points near the vessel bottom always have the same parameter values under the cylindrical projection. One solution for this problem is the combination of cylindrical projection and spherical projection for the points near the vessel bottom. Fig.4 shows the surface modelling result.

Geometric information also includes curve information for contour shape study. Contour shape information plays an important role in analysis of archaeological pottery. We use two kinds of models, chain codes and NURB curves to represent profile curves of archaeological vessels. Use of 2D geometric models can simplify the problem, and reduce 3D problem to 2D problem.

1. Chain codes

In order to get a 2D profile curve from a vessel, archaeologists use a cutting plane to intersect the vessel (polygonal mesh) and can get intersection points, then connect all the points according to some order, and get the chain code. Point coordinates of chain codes are stored in information models.

2. NURB curves

NURB curves are generated by fitting points of chain codes with least squares approximation. Since curvature has useful information such as convexity, smoothness, and inflection points of the curve needed by vessel analysis, we adopt cubic NURB curves to approximate profile curves of vessels.

$$\boldsymbol{P}(u)=\frac{\sum_{i=0}^{n} w_i \boldsymbol{d}_i N_{i,k}(u)}{\sum_{i=0}^{n} w_i N_{i,k}(u)} \tag{2}$$

where $\boldsymbol{d}_i$, $i$=0, 1, …, $n$ are control points; $w_i$ are weights; $N_{i,k}(u)$ are B-Spline basis functions. Degrees, control point coordinates, weights, knot vectors are stored in information models.

One of the important characteristics of a curve is the curvature, which is the magnitude of the rate of change of the tangent vector with respect to arc length. It is very useful for analysis and classification of vessel shape. In 3D space the curvature of curves is unsigned. However, for planar curves in 3D space, positive curvatures can be converted into signed curvatures.

Let $X(t)=(x(t), y(t), z(t))$ be a parametric curve, then the curvature at $t$, $\kappa(t)$, is defined as

$$\kappa=\kappa(t)=\frac{\left\|\dot{X}\wedge\ddot{X}\right\|}{\left\|\dot{X}\right\|^3} \tag{3}$$

where $\dot{X}=\dfrac{\mathrm{d}X}{\mathrm{d}t}$ and $\ddot{X}=\dfrac{\mathrm{d}^2X}{\mathrm{d}t^2}$, and "$\wedge$" denotes the cross product. In practice the curvature of a NURB curve can be computed by converting the NURB curve into a series of Bezier curves, and calculating the curvature of these Bezier curves (Farin, 1996). It is more robust and efficient, and is convenient to change always positive curvature value in 3D space into signed value in 2D space. It is very useful for our research to determine the Inflection Points (IPs) on profile curves.

**Feature information**

As defined before, feature information is of hierarchical structure. Feature points are basic features and other features, such as region features and volume features are defined on them. Analyzing geometric

**Fig.4 (L to R) A scanned vessel rendering by wire frame and shading, and its surface model generated by least squares approximation**

properties of profile curves can extract point features. All four kinds of feature points defined in Section 2.1 can be easily determined by analyzing curvature information, point position information and tangent line information of profile curves. Several algorithms for extracting point features are listed below.

**Algorithm 1** Algorithm for extracting end point features

Input: a profile curve represented by a B-Spline curve and a chain code respectively.

Output: end point features

1. end point 1 := start point of the chain code;
   end point 2 :=end point of the chain code;
2. center point :=center point of the chain code;
3. find the base section around center;
4. if base section is flat or concave then
   total end point number :=4;
   end point 3 := left terminate of base section;
   end point 4 :=right terminate of base section;
   else { base is convex}
   total end point number :=3;
   end point 3 :=center;
5. calculate feature information for each of the end points, include space coordinates, parameter value, position on the chain code, and so on.

**Algorithm 2** Algorithm for extracting corner point features

Input: a profile curve represented by a B-Spline curve and a chain code respectively.

Output: corner point features

1. calculate curvature value for each of the points on the chain code;
2. find points with local maximum (minimum) curvature value as candidates for corner points;
3. for each candidate do
   if angle at the candidate point<a predefined value
   then
   the candidate point is a corner point;
4. calculate feature information for each of the corner points, include space coordinates, parameter value, position on the chain code, and so on;

When computing the angle between points ($x_l$, $y_l$), ($x_0$, $y_0$) and ($x_r$, $y_r$) in Algorithm 2, the value of angles is sensitive to sample errors. In order to reduce errors due to sampling, instead of taking ($x_l$, $y_l$) and ($x_r$, $y_r$) as points of the curve, the coordinates of these points are calculated by averaging the coordinates of a group of neighbors to perform a less noise prone resampling.

Let us consider the mid point ($x_0$, $y_0$) of $n$ contiguous points in a chain code of a curve, where $n$ is an odd number, and let $p=n/2+1$ be the point ($x_0$, $y_0$). Thus, the initial point of the angle ($x_l$, $y_l$) is calculated from the $n/2+1$ previous point as

$$x_l = \frac{\sum_{i=1}^{p} x_i}{n/2+1}, \quad y_l = \frac{\sum_{i=1}^{p} y_i}{n/2+1} \tag{4}$$

and similarly for the end point of the angle ($x_r$, $y_r$)

$$x_r = \frac{\sum_{i=p}^{n} x_i}{n/2+1}, \quad y_r = \frac{\sum_{i=p}^{n} y_i}{n/2+1} \tag{5}$$

As for inflection features and point of vertical tangency features, they are easy to find by analyzing the curvature value and tangent lines.

## WEB APPLICATION

After getting point features we continue finding curve features and region features based on feature hierarchical definition. Then we use XML to represent the result. The purpose of using XML to represent information is that we can develop a distributed and Web-based visual query interface for archiving and searching 3D Archeological vessels. Embedding data in XML adds structure and Web accessibility to the inherent information of Archeological vessels. Fig.5 shows a sample of the XML-based information model.

We offer users content-based retrieval tools on the client site. Originally content-based retrieval system was mainly designed for 2D still image libraries. There are lots of unsolved problems remaining in the content-based retrieval of 3D models. Our research tries to implement the search engine for 3D models, especially 3D ceramics.

We combine MS IIS, Tomcat and MS Access to design the Web server for the visual query interface


**Fig.5 A sample of the XML-based information model**

(VQI). Tomcat supports JSP and XML/XSL. MS Access serves as database to store information models of vessels. A Netscape/IE plug-in was developed using C++ and OpenGL, and allows users to draw profile curves on screen. Drawn curves with other retrieval parameters, such as height, diameter, area, and volume are submitted to Web server, and process a content-based retrieval.

The query process in VQI combines a sketch-based interface and searches by traditional text and metric data. Representative vessel shapes can be selected from the supplied palette and modified, or a freeform profile sketch can be created in the interface window. Text and numeric fields support parallel query of descriptive and derived data within the databases. Query results from database are stored in XML format, and are visualized via a pre-designed Extensive Stylesheet Language (XSL) file. Fig.6 describes the flow chart of the Web-based VQI.

A hierarchical indexing structure for the database design is used to speed up the query procedure. The structure includes the compactness value of 3D solid objects, feature points and profile curves of 3D ceramics.

**Compactness computing**

The initial database search field is the indexing of compactness. The basic descriptive properties of rigid solids are the enclosing surface area and volume. A measure of compactness for solids relates the enclosing surface area with the volume. Thus, a classical measure of compactness can be defined by the ratio $C=(area^3)/(volume^2)$, which is dimensionless and minimized by a sphere.

For a sphere, its area is equal to $4\pi r^2$ and volume $(4/3)\pi r^3$. Therefore, $C=36\pi$ is the minimum compactness of a solid, since the sphere encloses maximum volume for a constant surface area. We can define regular compactness $C_{reg}=C_{min}/C$, where $C_{min}=36\pi$, and $C$ is classical compactness. Several compactness values of classic vessels can be found in Fig.7.

The compactness value is a real number. Totally different vessels may have the same compactness values. But these values can narrow the search range, and speed up the search procedure. A solid 3D model of a submitted 2D sketch curve is generated by CGI via surface of rotation modelling techniques. Then the compactness value of this solid model is used to search the database. Tens of similar vessels are returned as the initial search result by comparing the compactness values.

**2D sketch matching**

From the initial search result, a curve match process is called to perform the best matching of the submitted sketch curve and profile curves of 3D vessels. A curve-based Iterative Closest Point (ICP) algorithm (Besl and Mckay, 1992) is implemented for this. The ICP algorithm can be stated as following. A "data" shape P is moved (registered, positioned) to be

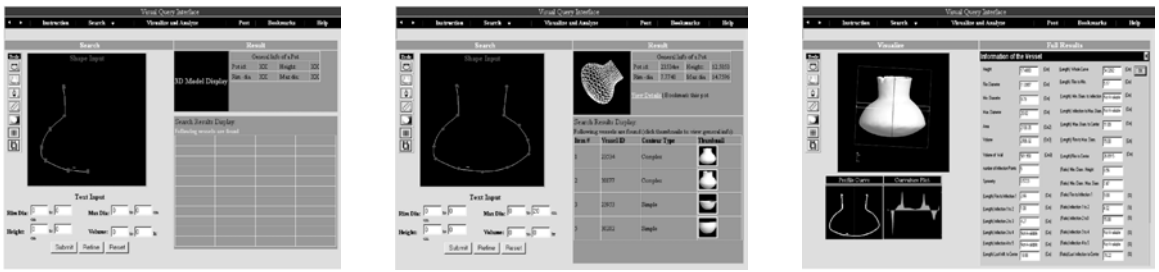

**Fig.6 (L to R) Query interface screen with sketch-based, numeric, and text-based input fields. Initial response screen with resulting thumbnail images and summary data, and wire frame of first matching vessel. Detailed individual display screen with 2D, 3D, and descriptive vessel data**

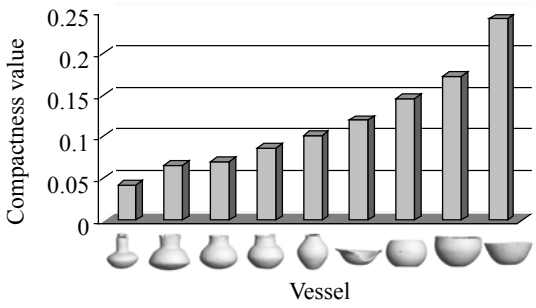


**Fig.7 Regular compactness values of some vessels**

in best alignment with a "model" shape X. The data and model shape may be represented in any of the allowable forms (point sets, line segment set, implicit/parametric curves, triangle sets, implicit/parametric surfaces, in our project, we will use triangle sets). The number of points in the data shape will be denoted by $N_p$. Let $N_x$ be the number of points, line segments, or triangles involved in the model shape. In our case, chain codes of profile curves in the Database are treated as line segments. The 2D sketches drawn by users are discretized into line segments too.

There are three basic computational components for ICP algorithm:

Step 1: Computing the closest points [costs: $o(N_pN_x)$ worst case, $o(N_p\log N_x)$ average];

Step 2: Computing the registration [costs: $o(N_p)$];

Step 3: Applying the registration [costs: $o(N_p)$]

The three basic computational components make up the iterative process of ICP algorithm, which always converges monotonically to the nearest local minimum of a mean square distance metric. Experience shows that the rate of convergence is rapid during the first few iterations. A trick to accelerate the ICP algorithm is that we use feature points on profile curves to calculate the initial position estimation.

## CONCLUSION

The final goal of the project is to learn about vessel uniformity and proportionality for different functions as indicators of developing craft specialization and complex social organization among prehistoric cultures. Use of metric rulers and visual inspection are inadequate to accurately capture the complex curvatures and proportionality of vessel forms and sizes. The project uses two Cyberware scanners, the M15 and 3030 to scan ceramic vessels. While individual scans take only 17 seconds, the total average scanning time for each vessel is about two hours, depending on complexity, color, texture, etc. Currently, scanning has focused on complete (or nearly complete) ceramic vessels from the Classic Period (A. D. 1250–1450) of the prehistoric Hohokam culture area of the Southwest (Salt/Gila River Valleys) near present-day Phoenix, Arizona. The ceramic vessels are part of the Roosevelt Platform Mound Study collection and the Department of Anthropology Whole Vessel collection conserved at the Archaeological Research Institute (ARI) at Arizona State University.

As the first stage of the project, this paper presented an XML-based information model for archaeological vessels, and introduced a Web visual query interface that allows users to access scientific information on Internet. In the further study we will extend our information model so that it can represent other scientific information, such as archeological Lithic tools, Anthropologic bones, etc.

## References

Bae, M., 1999. Curvature and Analysis of Archaeological Shapes. MS Thesis, Arizona State University.

Besl, P.J., Mckay, N.D., 1992. A method for registration of 3-D shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **2**:239-256

Birkhoff, G., 1933. Aesthetic Measure. Harvard University Press, Cambridge, MA, USA.

Farin, G.E., 1996. Curve and Surface for Computer Aided Geometric Design. 4th Edition, Academic Press, Boston.

PLY, 2004. Ply File Format. http://www.cc.gatech.edu/projects/large_models/ply.html.

Rice, P., 1987. Pottery Analysis: A Sourcebook. University of Chicago Press, Chicago.

Rowe, J., Razdan, A., Collins, D., Panchanathan, S., 2001. A 3D Digital Library System: Capture, Analysis, Query, and Display. 4th International Conference on Digital Libraries (ICADL), Bangalore, India.

Shepard, A., 1976. Ceramics for the Archaeologist. Carnegie Institution of Washington, Washington, D.C.

Staudek, T., 1999. On Birkhoff's Aesthetic Measure of Vases. Technical Report, Faculty of Informatics, Masaryk University, Czech.

Williams, R., 1998. Interfaces to Scientific Data Archives. Workshop Report, California Institute of Technology, Pasadena.

XML, 2004. Extensible Markup Language 1.0. http://www.w3.org/TR/REC-xml.