A Methodology for Evaluating Archaeological Web Sites

Nicoletta Di Blas
Politecnico di Milano, Milan, Italy

Franca Garzotto
Politecnico di Milano, Milan, Italy

Maria Pia Guermandi
IBC, Bologna, Italy

Franco Niccolucci
Università di Firenze, Florence, Italy

Abstract

Even if nowadays almost every cultural institution, including archaeological museums and sites, owns an institutional web site, there are few investigations on evaluation criteria. MiLE, a methodology for web sites evaluation, was recently specialized for cultural web sites by a team of experts from Politecnico di Milano, original developers of the general methodology, and from cultural institutions, coordinated by IBC. It precisely defines scenarios, i.e. abstract tasks performed by abstract users: evaluators assign marks to specific attributes, as clarity or accessibility, acting as the supposed user. Thus evaluation may take into account the target user population and the intended web site goals. A group of students in Architecture from the Università di Firenze was selected to perform an extensive application of this methodology to a large sample of web sites of archaeological institutions, including practically all such Italian sites, a number of Spanish and Polish ones and a sample of sites from other European and extra-European countries. The results give an insight into the effectiveness of such Internet presentations and test the validity of the methodology, suggesting an extensive application of it and the adoption of widely accepted and objective guidelines for multimedia cultural communication.

Introduction

Evaluating the effectiveness of multimedia for cultural communication is a task which as yet has been performed basing only on the intuitive appreciation of their global appearance. Objectives, intended audience and services to be offered are often unspoken and evaluation is carried on, when it is, with an empiric, subjective and qualitative approach. The quantification of multimedia quality remains therefore somewhat fuzzy, with no guidelines to make it as objective as possible, and comparing different applications or evaluating the attainment of a quality threshold, for instance, rely on a totally individual and perhaps unreliable judgment. Multimedia play, on the contrary, an increasingly important role in communicating Cultural Heritage and for some authors they are in fact the best, or only, solution to store and communicate the intrinsic reflexivity, contextuality, interactivity and multivocality of the archaeological record (Hodder 1999, Wolle, Tringham 2000, Biehl 2002, Tringham 2003); others have evidenced that Cultural Virtual Reality applications need to improve substantial aspects as validation, annotation and philological approach (Frisher et al. 2002), which also possibly impact on the archaeological method and theory, so far as valid 3-D models can provide new tools to archaeological investigation (Niccolucci 2002); finally, the widespread presence on the Web of cultural institutions, including museums and archaeological sites, does not clearly correspond to a well-tested business model, i.e. a set of criteria that guide the planning and enable to evaluate the effectiveness of such cultural web sites. So a specialized and effective methodology for the evaluation of multimedia is an urgent need not only for the valid communication of culture but also for the definitive acknowledgement of the theoretic importance of such technology in the realm of archaeological theory.

The present paper is a first step in that direction. It starts from the application of the MiLE evaluation method, developed at Politecnico di Milano, to cultural web sites as proposed in (Di Blas et al. 2002, Bolchini et al. 2003, Triacca et al. 2003) and it tests the proposed methodology on a large set of archaeologically related web sites. The evaluation was carried on by students in Architecture as part of their yearly assignments for the course of "Urban Models" held by Prof. Niccolucci at the University of Florence in Autumn 2002. The students were concluding their studies (5th year, in practice corresponding to what is called elsewhere "graduate students") and their skills were appropriate for the job, which consisted in filling on-line forms; moreover, they exercised in performing the task during a seminar and were supervised and received remote assistance by the teacher while completing it. For each evaluation record they also had to prepare a short comment to justify the scores, an useful feedback to tune the methodology. All the scores and report were reviewed by the supervisor and in a few cases the evaluator was asked to correct mistakes.

The sites under evaluation were chosen by the teacher to represent a wide panorama of Italian web sites with a good number of Spanish and Polish ones; a selection of other European sites, mainly from UK with a few German and French ones, were also taken into account; other sites dealt with non-European institutions, mostly from USA. The reason of the selection were language skills: students were asked to evaluate sites using their mother tongue, with a few more in a foreign language they declared to understand. A group of Erasmus Spanish and Polish students, attending the course, enriched the international flavour of the experiment.

The experiment had multiple goals: to test the method on a large group of evaluators; to evaluate the archaeological web sites; and to analyze sites usually less considered, for langua-

ge reasons, than those in English. For space reasons, it is impossible to describe here all the sites under evaluation or to appropriately cluster them (e.g. those related to large institutions, those related to museum collections versus those related to archaeological sites, etc.) and analyze the results in terms of such clusters. However, the original data (i.e. individual marks and scoring) will be available for some time on the web and a more detailed analysis will be performed in future work.

SUMMARY OF MILE

MILE is based on two main concepts, the Abstract Task (AT) and the User Profile (UP).

An Abstract Task (AT) is a general type of action that can be performed by visiting a web site only, e.g. "Get practical information for a visit as opening time, ticket cost, etc." or "Get the data necessary for writing a student's report". An AT needs to be defined in a unique and clear way. ATs may be classified according to scope as specific, complex or general, and according to concern as practical, operational and cognitive.

A User Profile (UP) is a general, but detailed, description of the user visiting the web, e.g. "A university student in Humanities, female, with good knowledge of English and a good (wide bandwidth) Internet connection" or "An adult (male) person, with average cultural background but no knowledge of foreign languages, with curiosity for archaeological treasures". The combination of an AT + a UP makes a Scenario.

A Web site has attributes, that is relevant properties, which are scored by the evaluator. Attributes may depend on the scenario, but it is preferable to give an overall list and then choose only the relevant ones. The current overall list of attributes includes the following:

1. Efficiency: the action can be performed successfully and quickly.
2. Authority: the author is competent in relation to the subject.
3. Currency: the time scope of the content's validity is clearly stated. Information is updated.
4. Consistency: similar pieces of information are dealt with in similar fashions.
5. Structure effectiveness: the organization of the content pieces is not disorienting.
6. Accessibility: the information is easily and intuitively - accessible.
7. Completeness: the user can find all the information required by the AT.
8. Richness: the information required is rich (many examples, data, etc.).
9. Clarity: the information is easy to understand.
10. Conciseness: the basic pieces of information are given; texts are not too long and redundant.
11. Multilevel: different levels are available according to user's profile.
12. Multimediality: different media are used to convey the information.
13. Multilinguism: the information is given in more than one language.

As stated before, attribute relevance may depend on scenario, which therefore will also include a relevance coefficient for each of them.

A web site is then evaluated:

- Considering all possible scenarios (or those considered relevant or defined as such by the site mission)
- Giving marks to attributes by direct inspection, using an agreed scale as 1 = very poor to 10 = excellent (0 = N.A.).

For the evaluation it is therefore necessary beforehand:

- To list all ATs
- To list attributes
- To list all UPs
- To create all scenarios by matching every AT with every UP, discarding incoherent couples
- To decide which attributes are relevant, and how much, for each scenario, that is to assign the attribute relevance coefficients for each scenario
- To decide which scenarios are relevant, and how much, for the site, that is to assign relevance coefficients for each scenario as far as the objectives of the evaluation are concerned.

After this preliminary process there will exist scenarios si, with attributes aj and relevance coefficients wik (0 £ wik £ 1) defined for each attribute ak and scenario si

- The evaluation is performed by assigning marks xik to each ak assuming to be in scenario si
- The overall mark mi for the site regarding scenario si is then given by the weighted average

$$m_i = \sum_k w_{ik} x_{ik}$$

- The same for the overall score S, the weighted average of mi using weights pi (0 £ pi £ 1) expressing the relative importance of each scenario in the site's overall goals, or simply averaging them (pi =1 for all i)if the site mission is unclear or unspoken.

$$S = \sum_k p_i m_i$$

This procedure standardizes the evaluation task as far as possible. ATs, UPs and attributes are defined by the evaluation team as well as weights, but subjectivity may be reduced using an agreed set of such features and in any case evaluation transparency is highly improved by the availability of evaluation criteria. Weights may take into account objective factors (e.g. the actual incidence of specific visitors' types) and the mix of scenarios may be precisely tuned to the target audience and the desired objectives of the Web site. The same method may be generalized for other multimedia applications, e.g. 3D or VR models, what will be the object of future work.

The Evaluation Experiment

In order to test the method and to have an extensive evaluation of archaeological museum web sites, the experiment was carried on as described in the Introduct ion. It was decided to privilege coverage (i.e. the number of Web sites under evaluation) versus depth (the number of scenarios taken into account for each Web site), to achieve a better understanding of the method effectiveness and have a first insight into the quality level of archaeological web sites. In other words, we were less interested in thoroughly evaluating individual web sites than in extensively examining of a number of such sites, to test the impression that in most cases these added little, if anything, to the visitors' understanding and satisfaction. Unfortunately, as it will be shown below, this was in fact the case.

After searching the Internet, 164 sites were selected for evaluation by the 18 students, but only 134 were in fact evaluated for different reasons (unavailability of the web sites, fault of the evaluator, excessive complexity of the site for the scope of the task, etc.). Of these, 98 related to museums, 23 to archaeological sites, 4 to complexes including a museum and a site, 1 to a temporary exhibition, and 8 concerned networks of cultural institutions.

The geographic (and linguistic) distribution of the sites was the following:

State	IT	PL	ES	UK	DE	FR	Other EU	Total EU	US	Other non-EU	Total
Number	72	18	9	8	4	2	8	121	9	4	134
Percentage	53.7%	13.4%	6.7%	6.0%	3,0%	1,5%	6.0%	90.3%	6,7%	3,0%	100,0%

IT, PL, ES, DE and FR sites were examined in their home language; the others in the English version, which of course corresponds to the home language for UK and US sites. Other European include sites from Belgium (2), and one each for Denmark, The Netherlands, Sweden, Croatia, Switzerland and Greece. Other non-European include sites from United Arab Emirates, Malaysia and Israel (2).

The test was performed on scenarios deriving from the following User Profiles and Abstract Tasks:

UP	Name	Description
1	Student	M/F High school student; good general culture; language knowledge at a school level
2	Educated tourist	Adult tourist (M/F) age 30-40; good general culture; cultural interests over the average; fair language knowledge

AT	Name	Description	Concern	Scope
1	Historic knowledge	To get general information about the historic context	Cognitive	General
2	Tourist evaluation	To evaluate the potential interest of a visit	Cognitive	General
3	Data collection	To get information about the holdings of the institution	Operational	Complex
4	Information collection	To get operational information for a visit	Operational	Complex

The scenarios under evaluation did not correspond to all possible combinations (UP, AT) but only to the significant ones: s1 = (1, 1); s2 = (1, 3); s3 = (2, 2); and s4 = (2, 4).

Results and Conclusions

As one of the goals of the experiment was testing the method on a rather large sample of persons, the scoring has been examined to verify the behaviour of the evaluators. Scores concentrate in the range 4 to 7, partly for statistical reasons and partly, perhaps, as it is typical of questionnaires, in which response tends to concentrate on central values. They show that variability is a personal factor and scores should perhaps be normalized before comparing the evaluation by different evaluators. However, the difference among scenarios was rather well perceived by students since scores for different scenarios about the same site show appreciable differences.

Of the 72 Italian sites, 40 (56%) have been evaluated below sufficiency, i.e. score below 6 over 10, and 24 (33%) are barely sufficient, between 6 and 7. Only 8 (11%) are "good", with no excellent ones. This bad evaluation is mainly due to very low scores for attributes related to multilevel, multimediality and multilinguism, all receiving an average score between 3 and 4 ("very poor"), while other attributes generally receive an average score between 6 and 7 ("sufficient"). The situation of the 9 Spanish sites (perhaps too small a sample to be significant of the Spanish presence on the Web in this field) is similar. Polish sites, a larger sample with 18 cases, do not behave much better, with 4 evaluated as "very poor" and a majority of "insufficient" ones. Also in this case the fault is caused by poor multimediality and multilinguism, with also some lack of consistence in the presentation of pages. All evaluated sites are in fact very weak as far as multilevel presentation is concerned, that is the possibility of graduated approaches for different users.

In conclusion, the experiment has shown that the method is feasible and requires little training. The results confirm that the use of the Internet as a communication tool for archaeological heritage is far from optimal and in most cases still unsatisfactory: the pages do not avail of the multimedia potential of the web and in general add little to printed text, which perhaps is still the reference communication model for most curators.

It is the intention of the authors to report in greater detail the aggregate results of the

evaluation, here summarized for space reasons. Moreover, they intend to complete the investigation on archaeological web sites in order to obtain a larger sample and report comparable evaluations for the archaeological presence on the web. A first attempt aiming at evaluating virtual reality archaeological recreations with a similar methodology is also on the way, and some preliminary outlines were discussed during a seminar (2003) at the Cultural Virtual Reality Lab at UCLA.

References

BIEHL, P.F., 2002. Hypermedia and Archaeology: Towards a Methodological and Theoretical Framework. In Niccolucci, F. and Hermon, S. (eds.), Multimedia Communication for Cultural Heritage, Archaeolingua, Budapest:147-153.

BOLCHINI, D., SPERONI, M. and TRIACCA, L., 2003. MiLE: a reuse-oriented usability evaluation method for the web. In Jacko, J. and Stephanidis, C. (eds.), Human - Computer Interaction: Theory and Practice (Part I), Vol. I of the Proceedings of HCI International 2003, LEA Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey.

DI BLAS, N., GUERMANDI, M.P., ORSINI, C. and PAOLINI, P., 2002. Evaluating the Features of Museum Web Sites. In Bearman, D. and Trant, J. (eds.), Museums and the Web 2002, Selected Papers from an International Conference, Archives & Museum Informatics, Pittsburgh, U.S.A.

FRISHER, B., NICCOLUCCI, F., RYAN, N.S. and BARCELÒ, J.A., 2002. From CVR to CVRO: The Past, Present, and Future of Cultural Virtual Reality. In Niccolucci, F., Virtual Archaeology, Archaeopress, Oxford:7-18.

HODDER, I., 1999. The Archeological Process - An Introduction. Blackwell, Oxford.

NICCOLUCCI, F., 2002. Virtual Archaeology: an introduction.. In Niccolucci, F., Virtual Archaeology. Archaeopress, Oxford:3-6.

TRIACCA, L., BOLCHINI, D., DI BLAS, N. and PAOLINI, P., 2003. Wish you were usable! How to improve the quality of a Museum Web site. In Cappellini, V., Hemsley, J. and Stanke, G. (eds.), EVA 2003 Florence, Proceedings, Pitagora Editrice, Bologna.

TRINGHAM, R., 2003. Interweaving Digital Narratives with Dynamic Archaeological Databases for the Public Presentation of Cultural Heritage. This volume.

WOLLE, A-C. and TRINGHAM, R., 2000. Multiple Çatalhöyüks on the World Wide Web. In Hodder, I. (ed.), Towards reflexive methods in archaeology: the example at Çatalhöyük, McDonald Institute Monographs, Cambridge:207-217.