Stuart Dunn – Nicolas Gold – Lorna Hughes

CHIMERA: A Service Oriented Computing Approach for Archaeological Research

Abstract: Archaeological data is available in greater complexity and volume than ever before, which presents significant challenges and opportunities. In this position paper, we describe how a Service Oriented Architecture (SOA) could be deployed not only to bring federated resources together with a unified searching capability, but also how those capabilities could themselves be used to create a 'bottom up' domain ontology for the resources' subject area or areas. By linking with Virtual Research Environment (VRE) technology, this system could deliver access to a range of digital resources for excavators in the field, and at the same time provide a valuable body of semantic information about the archaeological research workflow, as expressed in the natural language archaeologists use to conduct their collection searches.

Introduction

All archaeological research problems are multifaceted, most are interdisciplinary, and many will require the expertise of more than one scholarly community. Technological advance, often at a rapid pace, in scientific disciplines contiguous to archaeology increases the number of facets, complicates the interface between disciplines, and increases the need for collaboration. Many examples of successful technological innovation can be cited in support of this: it is impossible to imagine archaeology today without radiocarbon methods (e.g. Renfrew 1999); visualization techniques (Winterbottom / Long 2006) and GIS (Wheatley / Gillings 2002). The benefits that advances in these disciplines have brought are there for all to see. However, the benefits of computing – and more specifically software engineering (SE) – have been less visible, although certainly not less important (the three-decade plus longevity of the CAA conference testifies to such importance). A vast array of digital material has become available to archaeologists over the last ten years, in the form of data content resources, tools and websites. Although computational methods have developed a healthy bibliography since the 1960s, we believe that there is an urgent need now to consider how advanced computational approaches can be brought to bear on this new digital landscape, and that recent developments in SE give us an opportunity to adopt more formal and documented approaches to the past.

One such approach is the relatively-recent concept of Service Oriented Architectures (SOA). SOA was developed for businesses with complex and evolving IT needs, for which existing IT infrastructures could not be modified rapidly enough to meet business needs. The promise of SOA, is the possible use of remotely distributed functional services, composed at the time of need to form systems capable of meeting the desired functionality. The key significance for a research environment in archaeology (or more specifically a Virtual Research Environment [VRE], as set out in detail below), is that an SOA requires the business processes it enables to be computationally defined. We argue that the same principle of computational definition can be applied to the archaeological research workflow. By defining in this way what archaeologists do, we are able to consider more systematically the implications of mass digitization, ever more complex tools, and federated computational research systems, for archaeology; while at the same time building on recent work on the formal ontological structuring and manipulation of archaeological information (e.g. Doerr et al. 2004).

We argue that this renders us more able to consider how resources and tools can be coordinated and deployed. "Data Mining" is a well-known term in computer science and so-called "e-science", which describes the retrieval of information from massive digital corpora. We suggest that the term "Data Excavation" could be equally well applied to the plethora of digital archaeological material using an SOA approach. The distinction is an apt one: whereas mining is the isolation of material that one is interested in from a large corpus, excavation is at a smaller scale, yet far more nuanced, with more sys-

tematic separation and interpretation of evidence (or digital representations of the evidence).

We propose here a system termed CHIMERA: "Collaborative Harvesting of Information from Museums, E-Records and Archives". Our work in this area is at an early stage. This paper does not present the outcomes of any research, but is more a position statement on which the archaeological and scientific communities are invited to comment. It has become clear that a key area where the SOA can be applied is in novel ways of constructing of domain ontologies for archaeology, by capitalizing on the so-called "folksonomy" concept (Guy / Tonkin 2006). A folksonomy is a collection of user-generated metadata tags. Well known folksonomy systems include del.icio.us and flickr. Users upload their own multimedia data objects and tag them with their own descriptive terms. These tags are then used for searching, organization and management of the data. The folksonomy approach contains a number of challenges for deployment in any academic system. For example, Users A and B may look at Object C and disagree totally as to what tags to associate with it. Issues such as misspellings, typographic errors, use of plurals (two users may tag the same picture of a vase as respectively "vases" and "vase") and – crucially for archaeology – differing interpretation of the same object, may all lead to inconsistency and imprecision. We outline here a SOA-based system, employing recent developments in SE, which treats natural-language search terms entered into a common interface to federated data collections as folksonomy tags. These tags are then preserved in a metadata repository, which refreshes itself every time a new search is made, updating all the connections made between data objects and the tags used to describe them. Unlike conventional folksonomy systems, employing web services in this way will allow us to capture and measure generalized uncertainty about the objects users are querying, as well as linking with existing Virtual Research Environment concepts to deliver far more sophisticated data services to archaeologists in the field than have hitherto been available.

VREs and the Archaeological Workflow

The archaeological workflow onto which such a system must map may be summarized as follows:

- Discovery. The artefact (data object) is recovered from the field, either through survey or excavation.
- Attribution. One or more attributes, including (but not necessarily exhaustively) object class, type, colour, dimensions and a provisional date are attached to the artefact, based on its physical characteristics.
- Cross referencing. These attributes are compared with a range of attributes from comparable artefacts. This point in the process feeds in to the construction of four-dimensional artefact typologies, which are at the core of most reconstructions of material culture.
- Interpretation. The artefact's place in the wider regional and spatial context is determined.
- Publication. Representations and/or textual descriptions of the artefact are published either electronically or on paper. The artefact itself may enter a museum collection or other archive, where it will be given a non-random place within a context of other artefacts sharing its attributes.

It is worth considering some means with which archaeologists, long familiar with this research workflow, have responded to the challenges of the information age. The Silchester Roman Town project is a well-known example of this. Silchester is a large and complex archaeological excavation in Hampshire, England, concerned primarily with the investigation of urban settlement patterns in Roman Britain. It therefore deals with the full, massive, range of types of archaeological evidence: numismatics, architecture, ceramics, paleoenvironment, spatial data, etc. The data are large and complex, and the concomitant range of humans needed for its analysis and interpretation geographically dispersed and unable to meet regularly at the site, or discuss the finds either in situ or in their physical presence, even though the project already has established an "Integrated Archaeological Database" (IADB) across several servers. The project therefore established a VRE (see Introduction above) to a) enable digitization of material onsite and its direct uploading to the IADB and b) provide experts with real-time access to the entire IADB.

This activity has radically changed the onsite data collection process. In the past, as in most major excavations, this has involved site workers filling in a card when they find an object with details of its description, provenance, context, dimensions, etc. The VRE, however, connects the site directly with the IADB over the internet using a broadband wireless aerial with a standard 1 MB downlink and 256 kB uplink, mounted on a barn 600 m to the south of the excavation area. The initial concept was for fieldworkers to use PDAs and a ruggedized laptop to collect

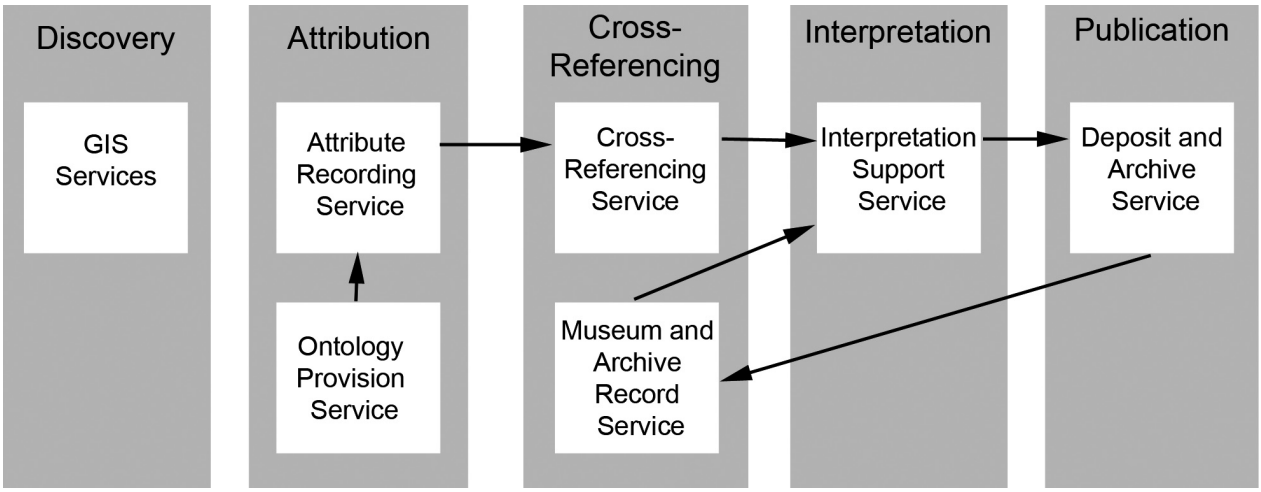

Fig. 1. Computational representation of the archaeological workflow.

data on finds and other features, and upload it directly to the database. When applied, however, it was found that there were some problems ensuring the quality of data being fed directly into the IADB. This highlights the need for a robust workflow (both technical and non-technical) at every stage of the process. Infrastructurally, the main problem with the IADB which the project was intended to address was the server interoperability. Each server operates behind a proxy or firewall, which is surmounted by employing clients using javascript to issue queries to each server and then amalgamate the results. The case of Silchester makes it clear that these issues now extend to the field, and to the analysis and attribution of objects as they emerge from the ground, as well as to the more traditional sphere of scholars working with secondary resources in the library and at their desktops. The next section describes the proposed CHIMERA system in more detail.

The Service Oriented Architecture

Service orientation is being hailed as the dominant future paradigm for the software engineering of large and complex systems (Brereton et al. 1999). A natural development from the foregoing paradigms of object and component orientation, services offer a number of advantages for the development and management of large software- and data-intensive systems:

- Separation of interface from functional implementation: this is the long-recognised principle of information hiding (Parnas 1972) but is worth mentioning in a services context since services extend it beyond "simple" implementation-hiding (as might be found in object-oriented systems) to potentially large-scale examples of distributed collaboration hidden behind a simple interface in a number of ways:
- Separation of integration knowledge from component knowledge: those providing services do not necessarily have to actually implement software but can provide added-value through their integration knowledge.
- Separation of ownership from use: particularly important in business, the idea of paying only for the use of functionality or data is more attractive than paying for the ownership of that function/data and its associated infrastructure. Although business benefits are obvious, these ideas transfer to the academy where the maintenance of infrastructure and costs of managing datasets inhibit sharing between research teams.
- Large-scale distribution: services explicitly allow for distributed data and functionality through open standards.
- Dynamic heterogeneity through standards for communication: services are designed to be composed into an application or aggregated dataset at runtime not development time.
- Potential for semantic description of function and data: there is ongoing work to describe functionality and data in formats amenable to both human analysis and automated reasoning (see work on the semantic web (W3C 2007)).
- Separation of computational resource from computational result delivery: execution of queries and programs can take place remotely from the computer requesting that execution and only the

results provided (either locally to the requester or left remotely for further interrogation).

For many years, software engineers have been dealing with complexity in large-scale software systems and the fact that this causes systems to lag behind the requirements of the businesses they support. Service orientation is perceived as the solution to this "legacy system" problem because it can be seen in both technological and commercial contexts (Bennett et al. 2001). In addition to dynamically composing and recomposing the technological services, one can re-order and recompose the business process (also termed workflow) to meet changing needs. The concept of workflow is very important at all levels of services-engineering and transfers easily to the concept of research workflows (for scientific discovery, grid-based workflows are frequently documented as part of the discovery process and technique, see Syed / Ghanem / Guo 2006). For example, the archaeological research workflow described earlier could be seen (in the context of services support) as a series of invokable services in a workflow as shown in *Fig. 1*. Two types of service may be observed to be in use in this scenario: function-oriented and data-oriented. Function-oriented services (e.g. the interpretation support service) embody executable functionality that can be composed into larger software applications. Data-oriented services (e.g. the GIS services or ontology provision services) are primarily concerned with the provision of data to other services in the workflow.

Data Organization

As noted in the Introduction, the sheer quantity of digital material available to archaeologists has increased at a massive rate in recent years. The archaeological communities have already developed a range of tools and methods for organizing, searching and retrieving such data, but the quantity and complexity of the data makes coordinated and integrated exploitation of these resources extremely difficult, particularly when tackling large research questions at a regional or national level. In fact, from an infrastructural point of view, this proliferation of material could be viewed as a "legacy system" of the kind described in the section above – well known tools, methods and technologies have been employed to create the material, but in spite of the success of the Archaeology Data Service in developing preservation and curation models, there remains the potential to do far more. Using the service oriented approach outlined above, the ADS has already developed the ArchSearch and ArchaeoBrowser facilities; these may be regarded as logical progenitors of a broader integrated architecture, in that they provide the facility for multiple keyword searching across many collections and, in the latter case, for using faceted classification to build many layered queries. Usage statistics for these services provide indications of the breadth of subject matter and extent of employment (see http://ads.ahds.ac.uk/catalogue/index.cfm?CFID=433285&CFTOKEN=24149332). Other initiatives such as the EU DRIVER project are also aimed at making research data available for user services (DRIVER 2007).

As noted above, the CHIMERA system we propose to develop provides a semantic layer based on a formalization of the folksonomy concept. We propose to offer a mechanism for "fingerprinting" descriptions of objects provided by archaeologists via the most intuitive and common method of information retrieval in the archaeological community: text-based searching. New information (i.e. new "fingerprints", or newly identified connections between two or more existing "fingerprints") supplied in the course of any search could affect any part of the corpus of knowledge: our approach will automate the capture of that new knowledge, and make it available to anyone submitting a subsequent search. The web service approach therefore combines the semantic properties of the folksonomy approach with more formal structures such as ArchSearch and ArchaeoBrowser.

Function and Application

Building on the basic service-architecture, we envisage a number of possible applications for service orientation in archaeological research and practice. Given the resource constraints that exist at archaeological digs in terms of the lack of expert availability, the international museum record, and other data sources, it would be ideal if this information could be provided, on-demand, to suit the requirements of the particular query in the field. This would allow the archaeologist to quickly situate their latest finds in the context of the international record allowing a more interpretive approach to the excavation. This is ideally suited to a service oriented approach: Museum and other collections would be exposed as services with functional query mechanisms made available on them to minimise the volume of data and computing power required in the field. The ex-

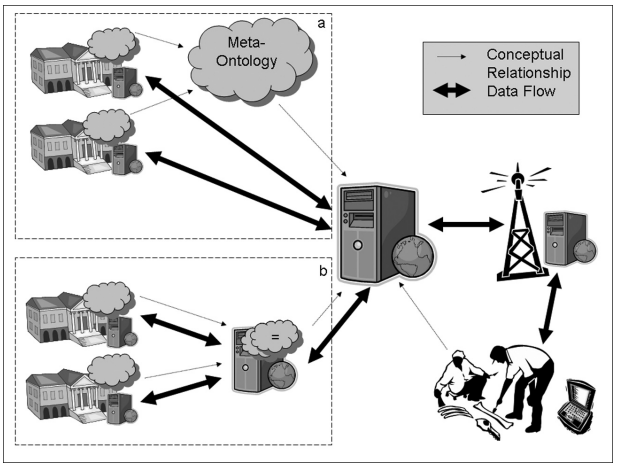

Fig. 2. (a) Meta-ontology approach, (b) Pairwise-matching approach.

cavator would require only a basic laptop computer with a network link (e.g. via GPRS) able to access a portal. The Silchester project has proved that this is possible technologically. The portal would provide tools for dynamically distributing queries across the database services to be used and integrating the results. The approach would require some level of semantic agreement between data providers in order that similar artefacts are described in such a way that they can be compared. Two approaches could be used for this:

1. Meta-ontology: a single meta-ontology could be created to which all local ontologies (for each museum may have its own description schema) are mapped. Queries and results can then be translated and expressed in terms of the meta-ontology and translated back to their local forms as necessary. There are great organisational benefits of such an approach but it could lack flexibility (e.g. it may be hard to extend the meta-ontology to cover all those it aims to model, or else it risks becoming so generic as to lose the important details). Significant up-front effort would also be required to create such a meta-ontology. *Fig. 2a* shows this approach.
2. Pairwise-matching: to avoid some of the problems of the meta-ontology approach, a collection of services could be created to map one data provider's ontology to another. The success of the overall approach is thus dependent on the existence of all the necessary services for doing this but the ramp-up to full functionality can be staged (since the system could still be used even with only a few mapping services). In addition, this market-oriented approach reflects more closely the general ethos of service orientation, allowing emergent properties and behaviours to arise from the creation of a space in which service providers and users can interact. *Fig. 2b* illustrates this approach.

Although the problem of managing and integrating data from multiple autonomous data sources is highly complex, major progress has been made on a similar problem in the domain of health and social care (the IBHIS project, Kotsiopoulos et al. 2003). In the IBHIS system, multiple autonomous data sources voluntarily map their own data schemas and access policies to an information broker's central definitions for roles and data. The files containing this information are stored by the data providers themselves and registered with the broker. When a query is made of the broker, it integrates as many services as are necessary to retrieve the desired information, mapping as necessary from its internal representation to those of the various data providers. In concept, therefore, this is closer to *Fig. 2a* than to *Fig. 2b*. Although there are many advantages to having a standardised central ontology, the approach does require substantial effort to set up and maintain. Adopting a pairwise matching procedure increases the overhead in the services-marketplace overall but reduces the up-front cost of establishing a working system since two services can be paired with far less effort and an initial service launched.

Metadata for all artefacts has to be generated, through user-labelling for instance. Placing a service in either a proxy or observer role would allow user queries to a system like CHIMERA to be used in the generation or refinement of metadata. For example, an excavator at the site of Palaikastro in Eastern Crete discovers a stirrup jar decorated with stylized octopi in the Late Minoan 1B style (e.g. Mountjoy 1984). If they had a wireless enabled device of the type used at Silchester, they would be able to upload to their own network system, which is federated with CHIMERA, a picture of the jar. They could then use the CHIMERA metadata repository to associate with that picture the words (or tags) "Palaikastro", "Stirrup Jar", "Octopus", "Marine Style" and "Ceramic". These tags would be stored in the CHIMERA repository and, because they have all been associated with the same object, related together. Later, a second researcher, working on a completely unrelated research project, wishes to know about ivory kouroi of the Greek Archaic period. They enter into the CHIMERA search interface the words "kour-

ous" and "ivory". The search is then run across all CHIMERA-federated collections, including the Palaikastro researcher's, and a second – museum – collection containing a record referring to the Palaikastro Kourous (MacGillivray / Driessen / Sackett 2000). The search engine will retrieve this record for the user's search because of the word "Kourous"; but it will also retrieve the Palaikastro researcher's stirrup jar find. The researcher, establishing that LM 1B pottery occurs at the same site as a major kouros, clicks on both records. The fact that both records have been clicked is recorded (anonymously). In other words, the Palaikastro researcher's identification and the existing museum record combines with the kouros researcher's search activities to prove a conceptual link between the concept of the kourous and the stirrup jar (and its attributes) and the place Palaikastro. This represents a new, user-driven way of establishing a bottom-up domain ontology.

Conclusion

Although archaeology is not experiencing a terabyte-scale "data deluge" of the kind currently seen in the physical, engineering and life sciences, and ever increasing volume and, critically, complexity of digital data requires new ways of managing and organizing the digital record, and presents opportunities for new kinds of archaeological research. Community generated content and metadata has numerous well-known problems. However, there have been few attempts to combine the acknowledged semantic power of the folksonomy approach with the rigorous ontological structures that academics require. In this paper we have presented a blueprint of how such a combination might work using tried and tested service oriented methods, and have described how a dedicated repository at the heart of such a service oriented system could provide the basis for a user-driven domain ontology. We view the textual information that researchers feed in to search engines as a vast, valuable and largely untapped body of information about the thought processes and workflows that govern how research is done. CHIMERA will realize that body of information's potential in the digital age.

References

Bennett et al. 2001

K. H. Bennett / M. Munro / N. E. Gold / P. J. Layzell / D. Budgen / P. Brereton, An Architectural Model for Service-Based Software with Ultra Rapid Evolution. In: Proceedings of the 17th IEEE International Conference on Software Maintenance (ICSM), Florence, Italy, November 7–9, 2001 (Los Alamitos 2001) 292.

Brereton et al. 1999

P. Brereton / D. Budgen / K. Bennett / M. Munro / P. Layzell / L. Macaulay / D. Griffiths / C. Stannett, The future of software. Communications of the ACM 42/12, 1999, 78–84.

Doerr et al. 2004

M. Doerr / D. Plexousakis / K. Kopaka / C. Bekiari, Supporting chronological reasoning in archaeology. In: F. Niccolucci (ed.), Beyond the Artifact – Digital interpretation of the past. CAA 2004. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 32nd CAA conference, Prato, Italy, April 13–17, 2004 (Prato 2004) 13–17.

DRIVER 2007

DRIVER, http://www.driver-repository.eu/ [4 Sep 2007].

Guy / Tonkin 2006

M. Guy / E. Tonkin, Folksonomies: Tidying up tags? D-Lib 12,1, 2006. http://www.dlib.org/dlib/january06/guy/01guy.html [7 Sep 2007].

Kotsiopoulos et al. 2003

I. Kotsiopoulos / J. Keane / M. Turner / P. Layzell / F. Zhu, IBHIS: Integration Broker for Heterogeneous Information Sources. In: Proceedings of 27th Annual International Computer Software and Applications Conference, Dallas, Texas, November 3–6, 2003, 378.

MacGillivray / Driessen / Sackett 2000

J. A. MacGillivray / J. M. Driessen / L. H. Sackett, The Palaikastro Kouros. A Minoan Chryselephantine Statuette and Its Aegean Bronze Age Context. British School of Athens Studies 6 (London 2000).

Mountjoy 1984

P. A. Mountjoy, The Marine Style Pottery of LMIB/LHIIA: Towards a Corpus. The Annual of the British School at Athens 79, 1984, 161–219.

Parnas 1972

D. L. Parnas, On the Criteria To Be Used in Decomposing Systems Into Modules. Communications of the ACM 5/12, 1972, 1053–1058.

Renfrew 1999

C. Renfrew, Before civilization: the radiocarbon revolution and prehistoric Europe (London 1999).

Syed / Ghanem / Guo 2006
J. Syed / M. Ghanem / Y. Guo, Supporting Scientific Discovery Processes in Discovery Net. Concurrency and Computation: Practice and Experience 19,2, 2006, 167–179.

W3C 2007
W3C, Semantic Web Activity, http://www.w3.org/2001/sw/ [29 May 2007].

Wheatley / Gillings 2002
D. Wheatley / M. Gillings, Spatial technology and archaeology: archaeological applications of GIS (London 2002).

Winterbottom / Long 2006
S. J. Winterbottom / D. Long, From abstract digital models to rich virtual environments: landscape contexts in Kilmartin Glen, Scotland. Journal of Archaeological Science 33,10, 2006, 1356–1367.

Zhu et al. 2004
F. Zhu / M. Turner / I. Kotsiopoulos / K. Bennett / M. Russell / D. Budgen / P. Brereton / J. Keane / P. Layzell / M. Rigby / J. Xu, Dynamic Data Integration Using Web Services. In: Proceedings of 2nd International Conference on Web Services, San Diego, USA, July 6–9, 2004 (Los Alamitos 2004) 262–269.

Stuart Dunn
Lorna Hughes

King's College London
Centre for Computing in the Humanities
Kay House
7 Arundel Street
London WC2R 3DX
United Kingdom
stuart.dunn@kcl.ac.uk
lorna.hughes@kcl.ac.uk

Nicolas Gold

King's College London
Department of Computer Science
The Strand
London WC2R 2LS
United Kingdom
nicolas.gold@kcl.ac.uk