Spatial Data Infrastructures and Archaeological Excavation Data: *SILEX*, the SDI of the Neolithic Flint Mine of *Casa Montero* (Madrid, Spain)

Fraguas, A.[1], Menchero, A.[2], Uriarte, A.[1], Vicent, J.[1], Consuegra, S.[1], Díaz-del-Río, P.[1], Castañeda, N.[1], Criado, C.[1], Capdevila, E.[1], Capote, M.[1]

[1] Research Group "Prehistoria Social y Económica", Instituto de Historia, CCHS - CSIC, Spain

{alfonso.fraguas, antonio.uriarte, juan.vicent, susana.consuegra, pedro.diazdelrio, nuria.castanyeda, cristina.criado, enrique.capdevila, marta.capote}@cchs.csic.es,[2]antonio.menchero@gmail.com

SILEX is a Spatial Data Infrastructure developed for the management and distribution of the primary archaeological information about the Neolithic flint mine of Casa Montero, located in the periphery of the city of Madrid (Spain). It is designed according to an open approach, based on the use of standards and open source software and on the free access to the whole data about the site via Internet. It is a distributed information system with a three layer architecture: the data layer, formed by a GIS level and a complex entity-relationship thematic database; the web service layer, with the use of standard protocols and languages for accessing the database; and the interface layer, a mashup that combines forms and a geographical viewer for querying and retrieving data.

Keywords: Neolithic, Flint Mining, Spatial Data Infrastructure, Open Data, Open Source, XML, OGC, INSPIRE.

1. Introduction: the flint mine

SILEX (*Sistema de Información Locacional en XML*, in English *Locational Information System in XML*) is an information system, under the form of a Spatial Data Infrastructure (SDI), designed for the systematic management and distribution of the complete dataset generated throughout the excavation and laboratory analyses of the Early Neolithic flint mine of Casa Montero (Madrid, Spain). SILEX is accessible through the URL http://www.casamontero.org/ and has been developed by the Research Group *Prehistoria Social y Económica* (*CCHS – CSIC*), one of whose research lines is the development and application of Information and Communication Technologies (ICT) in Archaeology.

The Neolithic flint mining complex of Casa Montero is located on a bluff overlooking the confluence of two of the main rivers in the province of Madrid, the Jarama and the Henares (DÍAZ-DEL-RÍO *et al.*, 2006; CAPOTE *et al.*, 2008; BUSTILLO *et al.*, 2009). It occupies an extension of less than three hectares that have been almost completely mapped and partially excavated. The open area excavation revealed the existence of 3794 cylindrical mining shafts, of one meter mean in width and up to ten meters deep. 338 of them were sampled and excavated, containing 2646 different archaeological deposits. The recovered stone remains amount to 65 tons, more than 1.5 million items, while the total stone processed at the site during the Early Neolithic may have been more than 769 tons (more than 17 million items). The majority of these remains are knapping residues. The main goal of these knapping activities was the production of blades and bladelets, the most common final product being blades with mean dimensions of 5x2 cm. All mining episodes occurred between 5300 and 5200 cal BC, a time span of approximately four generations. These generations were the earliest agricultural and pastoral groups known to have existed in central Iberia.

The abundance and complexity of the archaeological data suggested the need for a precise and systematic methodology, as well as an efficient information management system.

2. An open approach

One of the driving principles of SILEX is the *open* approach, consisting of the unconstrained and free distribution of information and knowledge resources. This approach starts from the ethical stand that supports that sharing knowledge helps to the development of science and society in an equitable and enriching way. The open approach has three aspects (WALSH, 2010): open data, open source and open access.

The concept of *open data* involves that these are “freely available on the public Internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself. To this end data related to published science should be explicitly placed in the public domain” (URL 1, see below). In this context, one of the basic objectives of the *Casa Montero* Project is the distribution of the whole set of the primary archaeological information about the site and the free and open access to it. Moreover, this access is dynamic, through a web interface that allows browsing, querying and obtaining specific datasets according to specific criteria selected by the user. The aim is not to publish a static and flat digital book, but a tool for selecting and downloading information concerning the archaeological site. This allows researchers to get primary archaeological data for performing alternative analyses upon them and obtaining scientific results complementary or even contradictory to those generated by the own excavators.

The second branch of this open approach is the *open source* (URL 2). Most of software applied in the development of SILEX has use licenses that meet this characteristic.

In the third place, *open access* refers to “the international effort to make research articles in all academic fields freely available on the Internet” (URL 3). Our aim is to put the documentation produced about the Casa Montero Project and SILEX on the Web, so as to fully show the theoretical, methodological and technological framework and development of SILEX.

Data exchange demands the use of a common language between producer and user for guaranteeing *interoperability*, that is, the ability between various systems to communicate and share information. Interoperability implies the use of standards. The development of SILEX has involved the use of several standards, both concerning Internet itself, as those created by the *World Wide Web Consortium* (*W3C*) (HTTP protocol, XML language), and geographical and spatial information, as those proposed by the *Open Geospatial Consortium* (*OGC*) (GML language, OGC web services). In fact, many of these *de facto* standards have become *de iure* standards through its conversion into ISO norms, e.g. XML and GML and some OGC web services, as WMS. This trend has its echo in the legal and administrative sphere, through the promulgation of laws promoting the normalisation and open distribution of data produced by public institutions. One leading example is the INSPIRE Directive by the European Union in 2007 (URL 4), requiring the standardisation and publication via Internet of the geospatial data created by its member states.

3. Architecture

As a SDI, SILEX is a distributed information system. The use of web technologies allows the dissemination of information in a friendly way, for it uses Internet architecture itself. So, each resource has its own universal identifier, its URI (*Uniform Resource Identifier*) (BERNERS-LEE, 1996), or a URL (*Uniform Resource Locator*) in navigational terms. Additionally, each spatial resource is accessed by means of calling a remote procedure (RPC, *Remote Procedure Call*), which is located in a URI. The use of URI as pointers for accessing information makes SILEX a real graph-oriented database, a net of information nodes connected by edges or relationships through a distributed system (Figure 1).

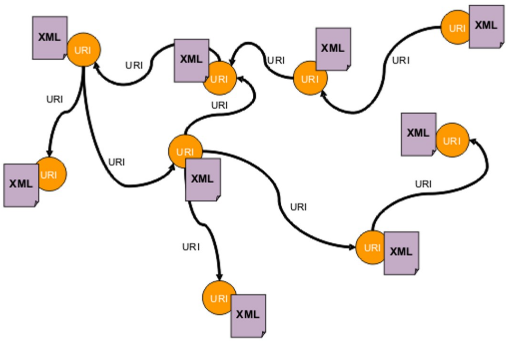


Figure 1: *Graph representing relationships between resources.*

If every resource (e.g. a shaft, a deposit or a lithic item) is given a URI, then one can get it with any software application able to retrieve resources from the Web. If resources are linked by their respective URL, then one can navigate through the full set of information just following the links, as one does when explores web pages. So, if there are several servers publishing information and all sharing a common web publishing framework, every client may retrieve data from it. The web publishing framework may act as a catalog, indexing resources which are located in different servers so that every resource is connected with others.

In this way, for publishing the whole set of data about Casa Montero, we have developed the following web tools:

1.→ A web framework for publishing data, so that any archaeologist may retrieve and analyse them by his own.

2.→ A web admin interface, that is, a friendly browser application, so that users can edit their data according to a shared model and save and upload them to the web publishing framework.

3.→ A web query interface, so that archaeologists can query the whole data set at the publishing framework to retrieve the data subset they are interested in.

SILEX has a typical multilayer architecture, composed by three levels (Figure 2): data storage or persistence layer, services layer or logic layer, and user interface or presentation layer. During several decades, increasing complexity in information systems has led to the autonomous encoding of their constituting parts. Such layers interact through interfaces. The main advantage of this kind of architecture is that it minimizes the impact of changes in one layer upon the others. Multilayer systems are easily scalable, that is, can be extended and adapted to new requirements without putting their functioning at risk. For example, we have selected the choice of using a native XML database for storing thematic information in the data layer; nevertheless we could replace it by a classic relational database without affecting the presentation and services layers.

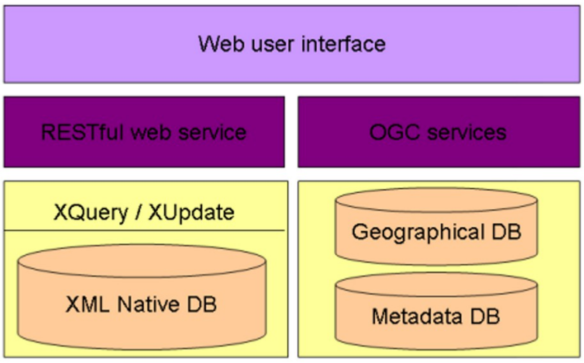


Figure 2: *SILEX web architecture.*

3.1. The database

The archaeological information contained in SILEX has a double dimension: thematic and spatial.

The thematic database is organized according to the Extended Entity-Relationship (EER) Model, a development of the classic Entity-Relationship (ER) Model (CHEN, 1976).

Information is stored in a native XML database. XML has been the selected language code because it is a widely accepted data exchange standard between different systems. For the definition and description of the entities of the SILEX data model, we have created our own XML Schema, CasamonteroML (URL 5-7). As usual in native XML databases, information is structured in collections, subcollections and documents. This native XML database can be queried through XQuery, a powerful language created by W3C.

SILEX has a complex data model, considering around 40 different entity types or classes (URL 8). The main part of the data model is formed by two basic blocks, one about the excavation entities and the other about the archaeological items, mainly lithic industry. This is not a closed model, for it can be extended through the definition of new classes and applied to other archaeological sites.

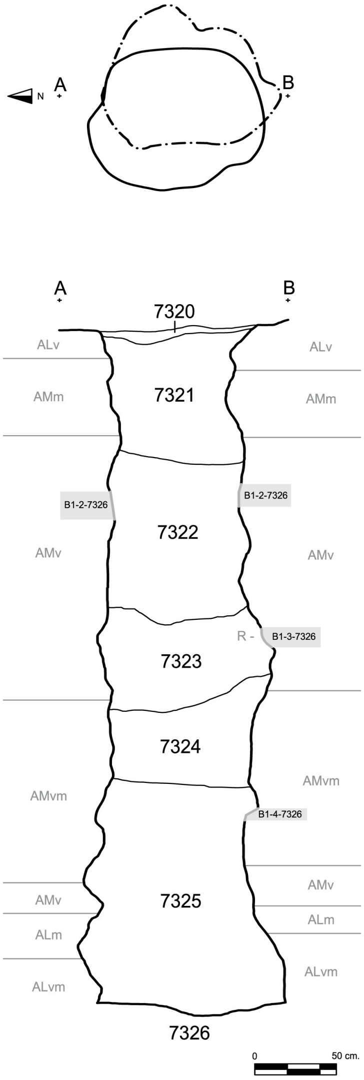


Figure 3: *Archaeological section of a shaft and its deposits.*

The excavation part is modelled according to Harris' system (HARRIS, 1989). The basic entity type is the *stratigraphic unit*, which has several subclasses. On the one hand, we have the *negative stratigraphic units*, that is, the result of removing sediment. The majority of them belong to the subclass *shaft*, which is the result of digging the soil for finding and extracting flint seams. On the other hand, we have the *positive stratigraphic units*, with an only subclass, the *deposit*, which is the result of adding sediment. Most of deposits fill mining shafts, being the result of throwing the soil and waste products of flint knapping and other materials back into the shafts (Figure 3). Moreover, stratigraphic relationships between units are considered in the data model: e.g. a deposit *covers* another deposit, a deposit *fills* a shaft or a lateral excavation *cuts* a shaft.

Archaeological items are objects recovered from deposits and all of them are included in the general class *finds*. Most of finds are *lithic pieces*, for whose description and classification the logical analytical system (MORA *et al*., 1992) has been used, which involves the detailed definition of several lithic industry subclasses.

Linked to thematic data, there is abundant graphic information, such as photographs and drawings (e.g. the sections of the mining shafts). These are stored as files with standard formats (JPG, PDF) in the directory structure of the web server.

As a SDI, SILEX includes spatial information. The spatial data model has been constructed following the OGC standards. Vector information is stored in an object-relational DBMS (PostgreSQL) with a spatial module (PostGIS). Raster data are stored as standard files in the directory structure. Metadata relative to spatial information is stored in the same DBMS that vector layers.

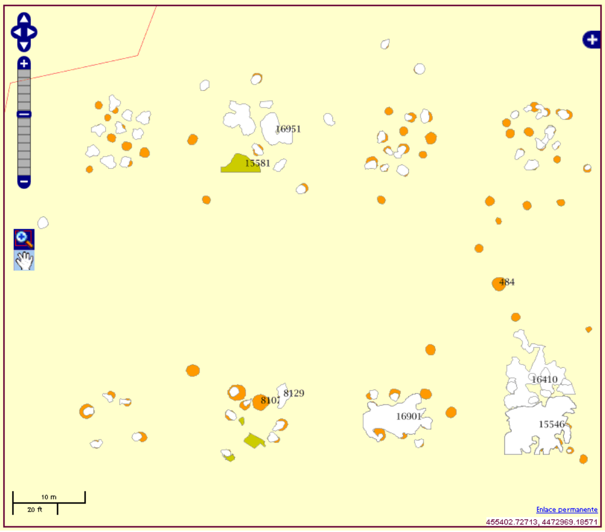


Figure 4: *Spatial entities on the geographical viewer.*

Main spatial data consist of a series of vector layers that correspond to some entity classes in the thematic data structure. All of them are negative stratigraphic units, e.g. shafts. Each vector feature is the planimetric representation of its corresponding negative stratigraphic unit, which has been extracted from site planimetry (Fig. 4), and is linked to its corresponding thematic entitiy by means of an ID, which is the stratigraphic unit number. The rest of thematic entities (e.g. deposits or archaeological items) have no direct spatial representation, and are indirectly georeferenced through the nega-tive stratigraphic units they belong to.

Also, there are complementary vector and raster layers, without linked thematic information, which have been included for visualising purposes, e.g. excavation limits, sample units or referenced oblique aerial photographs. Besides, SILEX use layers served by other SDI via web services, like the WMS of the Spanish aerial orthophoto program (PNOA).

3.2. The web service

SILEX is organized as a mashup that combines several web resources. According to previous SDI prototypes (FRAGUAS, 2008), we have used a mixed software architecture that uses web services with RESTful design criteria for thematic data (see a full description in URL 9) and OGC services for spatial data (WMS, WFS, WCS, CSW).

Thematic data is accessed through services with a Resource Oriented Architecture (ROA), consisting of a set of guidelines (RICHARDSON *et al*., 2007) for the implementation of the REST architecture (FIELDING, 2000). According to its designers, ROA has four advantages: 1) *addressability*: the web service exposes at least one URI for each information resource; 2) *connectedness*: the requested representation includes URI from other resources and forms with which one can build another URI, so that the client can discover new infor-mation just following the links; 3) *statelessness*: each request to the web service happens in isolation to others; when the client makes a request, this includes all the information necessary to generate a reply, that is, the server does not require additional information about preceding requests, but if it was necessary this information would be provided again by the client; 4) *uniform interface*: operations that can be executed are the same for any resource.

The web service used for thematic information turns read-and-write requests into XQuery/XUpdate requests to the native XML database via HTTP. In this way, the thematic web service can use any database as long as it supports XQuery, XML:DB and XUpdate over HTTP.

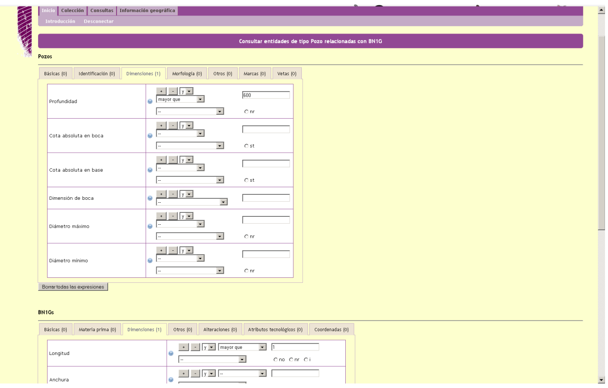
We have implemented the possibility of filtering entities according to their properties and the role of their relationships with other entities. The result of such a query is a list of URI containing the resources that satisfy that filter. For example, the URL 10 (see below) returns a list of shafts more than 600 mm deep related to NaB whose longitude is greater than 1 mm.

Regarding spatial information, we have followed the INSPIRE Directive and have technically implemented it by means of SOAP services created by the OGC. It is thus a Service Oriented Architecture (SOA). Vector features are exchanged in GML format, the extension of XML for spatial data. According to INSPIRE recommendations, SILEX allows the integration of detailed archaeological information layers into local, regional and global SDI. That is, services and resources used by SILEX can be included as a node in a wider SDI.

3.3. The web user interface (WUI)

The Web User Interface (WUI) is a web application based on browser. The browser has to implement XHTML (eXtensible Hypertext Markup Language) and JavaScript. It is a client of the web service previously described and allows users to browse, query and update information resources. In this way, we have developed a Rich Internet Application (RIA) combining AJAX technology (Asynchronous JavaScript And XML), XHTML and CSS (Cascading Style Sheets). AJAX interacts with DOM (Document Object Model) through the XMLHttpRequest object that browsers contain since 2002.

The interface has a menu with links to static descriptive pages about the main characteristics of SILEX and to the list of elements of each type of entity. Also, there is a submenu for making queries. For this, there is a series of forms that generate URL for demanding information to the web service. For example, the URL described in the previous section can be created using these forms (Figure 5). The list of entities returned by a query is a XML file that is visualised by default as a XHTML + JavaScript form dynamically generated by means of XForms definitions and that can be converted into a PDF or CSV document.



Consultar entidades de tipo Pozo relacionadas con BN1G

Pozos

Profundidad	y mayor que --	600 ○ nr
Cota absoluta en boca	y -- --	○ st
Cota absoluta en base	y -- --	○ st
Dimensión de boca	y --	
Diámetro máximo	y -- --	○ nr
Diámetro mínimo	y -- --	○ nr

BN1Gs

Longitud	y mayor que --	1 ○ no ○ nr ○ i
Anchura	y --	

Figure 5: *Form for generating thematic queries.*

Besides, there is a geographical viewer for showing the spatial layers. One can click on a specific spatial entity and get the thematic information linked to it. Inversely, when the result of a thematic query involves entities that have spatial representation, one can see their distribution on the viewer.

Conclusions

SILEX defines a data model for describing and organizing the complexity of different types of archaeological entities (deposits, shafts, lithic artifacts an so on) and the relationships between them. It includes two well established archaeological "ontologies", as the Harris' system for describing and relating stratigraphic units and the logical-analytical system for classifying lithic industry. The data model has been implemented by means of a XML based format.

On the other hand, SILEX is a distributed system that offers this information via Internet. The web framework allows to share data, so that archaeologists may publish information produced by themselves or, inversely, retrieve that created by others and analyse it with their own tools and criteria. A web admin application has been developed to edit, save and publish archaeological data through the web browser; that way archaeologists do not have to know the internals of the publishing framework or XML format and do not need to install proprietary applications. Moreover, a web query application allows archaeologist to query the full data set and retrieve the subset they are interested in. Using open source tools, we have built a mashup that gives access to geographical information through OGC services, following OGC standards based on the INSPIRE Directive.

SILEX will be integrated into more general SDI, such as the *IDEE*, the Spanish SDI for institutional geographical information, and the *IDE-CSIC*, the SDI designed for managing and serving spatial information generated by scientific projects by the *Consejo Superior de Investigaciones Cientificas* (CSIC), the Spanish research council.

Although SILEX is devoted to a specific site and designed according its particular features, it can be useful as a SDI model for new archaeological site information systems. Because of its flexibility, it can be developed to incorporate other levels of archaeological information, such as new sites.

Recently, at the TED 2009 during the "The Great Unveiling" conference in Long Beach (California, USA) (URL 11), Tim Berners-Lee, father of the WWW, asked the public to shout "raw data now!", in order to facilitate linked data. SILEX had already begun serving raw data about the Neolithic flint mine of Casa Montero in 2008. The next step for SILEX (a 2.0 version) should link data through standard protocols (RDF) and define their semantics through common languages (OWL) for defining ontologies in the context of the Semantic Web.

Appendix: Software

The software packages used are the following:

1.→ Apache Tomcat, as a servlet container for mounting web applications.

2.→ GeoServer, as a server for providing geographical information layers via OGC standards such as WMS, WFS and WCS.

3.→ OpenLayers, as a Javascript library for showing dynamic maps in web pages.

4.→ GeoNetwork, for offering metadata according the OGC's CSW specification.

5.→ PostgreSQL and PostGIS, for storing and recovering maps and metadata.

6.→ Apache Cocoon, as a web publishing environment based on XML, for designing web applications, both the web service and the web interface for accessing thematic data.

7.→ Orbeon XForms, for converting XForms into cross-browser forms based on HTML and Javascript.

8.→ Saxon, a XSLT transformation motor, for converting XML documents.

9.→ eXist, a native XML database, for storing thematic data, with a query motor based on XQuery.

Appendix: Cited URL

URL 1: http://www.pantonprinciples.org/

URL 2: http://www.opensource.org/

URL 3: http://www.soros.org/openaccess/

URL 4: http://inspire.jrc.ec.europa.eu/

URL 5: http://www.casamontero.org/webservice/schemas/casamontero.xsd

URL 6: http://www.casamontero.org/webservice/schemas/casamontero.rng

URL 7: http://www.casamontero.org/webservice/schemas/relaciones.rng

URL 8: http://www.casamontero.org/webservice/schemas/modelo.pdf

URL 9: http://www.casamontero.org/webservice/schemas/CMWS1.0.pdf

URL 10: http://www.casamontero.org/wui/pozos/related-to/bn1gs.html?pozo.dimensiones.profundidad=.%20gt%20600&bn1g.dimensiones.longitud=.%20gt%201&

URL 11: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

References

BERNERS-LEE T., 1996. Universal Resource Identifiers - Axioms of Web Architecture. http://www.w3.org/DesignIssues/Axioms.html

BUSTILLO M.A., CASTAÑEDA N., CAPOTE M., CONSUEGRA S., CRIADO C., DÍAZ-DEL-RÍO P., OROZCO T., PÉREZ-JIMÉNEZ J.L., TERRADAS X., 2009. Is the macroscopic classification of Flint useful? A petroarchaeological analysis and characterization of flint raw materials from the Iberian Neolithic Mine of Casa Montero. *Archaeometry* 51 (2), pp. 175-196.

CAPOTE M., CASTAÑEDA N., CONSUEGRA S., CRIADO C., DÍAZ-DEL-RÍO P., 2008. Flint Mining in Early Neolithic Iberia: A Preliminary Report on ‘Casa Montero’ (Madrid, Spain). In Allard P., Bostyn F., Giligny F., Lech J. (eds.) *Flint Mining in Prehistoric Europe. Interpreting the Archaeological Records,* pp. 123-137. BAR International Series 1891, Oxford.

CHEN P., 1976. The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems* 1 (1), pp. 9-36.

DÍAZ-DEL-RÍO P., CONSUEGRA S., CASTAÑEDA N., CAPOTE M., CRIADO C., BUSTILLO M. A., PÉREZ-JIMÉNEZ J. L., 2006. The Earliest Flint Mine in Iberia. *Antiquity* 80 (307). http://antiquity.ac.uk/ProjGall/diazdelrio/index.html

FIELDING R.T., 2000. *Architectural Styles and the Design of Network-based Software Architectures*. University of California Press, Irvine (CA). http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

FRAGUAS A., 2008. The ARANO SDI: A Spatial Data Infrastructure for the Rock Art of Northeast Africa. *Archaeological Computing Newsletter* 68, pp. 1-8.

HARRIS E. C., 1989. *Principles of Archaeological Stratigraphy*. Academic Press, London & New York.

MORA R., MARTÍNEZ J., TERRADAS X., 1992. Un Proyecto de Análisis: El Sistema Lógico Analítico (SLA). In Mora R., Terradas X., Parpal A., Plana C. (eds.) *Tecnología y Cadenas Operativas Líticas*, pp. 173-199. Bellaterra.

RICHARDSON L., RUBY S., 2007. *RESTful Web Services.* O’Reilly, Sebastopol (CA).

WALSH J., 2010. Free Software Model for Open Knowledge. *IV Jornadas SIG Libre (Girona, 10-12 marzo 2010).* http://prezi.com/f8k9_9rnxlap/a-free-software-model-for-open-knowledge/