Sharing Data on the Aquileia Heritage: Proposals for a Research Project

Vito Roberto and Paolo Omero
Department of Informatics, University of Udine, Italy
vito.roberto@uniud.it, paolo.omero@uniud.it

Abstract. Basic ideas are presented of a multi-national research project to share data about the Roman city of Aquileia employing the Information and Communication Technologies (ICT). A Consortium is proposed to manage the project, adopting the Open-source approach to the software design. The Consortium proposes a common vision of the data which we detail in the paper. There results a shared vocabulary of terms and meanings, as well as standard metadata formats to encode, classify and exchange data from whatever source. A federated system of computer resources realizes and supports the project. We also discuss the results to be realistically expected in short time of a low-cost, joint research effort.

Keywords: Distributed Systems, Open Source, Web Services, Data Repositories, Ontologies, Metadata.

1 Introduction and motivations

The extraordinary archaeological, historical, artistic heritage of Aquileia has attracted scholars from several Countries, whose research efforts have contributed to the studies since at least two centuries. There resulted another, invaluable heritage: the data originating from surveys, excavations, maps, photos of the city and its large commercial basin, including the Adriatic areas now pertaining to Italy, Slovenia and Croatia, as well as districts of Austria, Croatia, Germany, Hungary, Slovenia.

Such data heritage is currently dispersed among numerous Institutions, which are generally not aware of the data owned by the other ones. In addition, data are recorded on different supports - manual, mechanic, electronic analog or digital ones, which raises the problem of making them interoperable among the Institutions.

As a consequence, an invaluable amount of data is not adequately exploited, with waste of human resources and research funds.

We believe it's time to consider the opportunities offered by the Information and Communication Technologies (ICT) to promote the exchange, interoperability and re-use of data resources, namely: establishing a dedicated infrastructure based on the

Internet and the Web; gathering data sets into shared repositories; proposing common standards of digital encoding, indexing and database archiving.

In the domains of Cultural Heritage, data remain of interest for long time provided they are accessible to a large number of users. They are validated by the association and correlation with other data, so that pooling them in a consistent way allows to re-use and better exploit their potentialities.

The paper presents a few proposals towards an international collaboration project. In Section 2 we propose some basic technological principles of a federation agreement among Institutions; Section 3 presents a federated architecture, i.e., how the agreement can be put into practice by means of a distributed system of computer resources; Section 4 addresses the issue of how to share a common vision of the data; Section 5 contains our conclusions.

2 A federation agreement

The interested Institutions should agree on a few basic principles, under an ICT perspective, inspiring a collaboration within a Consortium. Let us discuss some main issues.

- The Institution is willing to make available a part of its own data to other Partners, to the extent dictated by its own policies;
- The Institution remains the owner of the same data; the responsible of their integrity, maintenance, updating and security;
- The Institution is willing to adopt a unique semantic data model to be shared with the Partners – i.e., a vocabulary of terms, concepts, taxonomies; a set of metadata enabling the indexing and exchange over the Internet;
- The Institution is willing to adopt Open Source software technologies in accordance with its own policies, to support possible joint projects;
- The Institution is willing to adopt technological standards of data encoding and exchange over the Net.

The autonomy of each Institution is apparent: to preserve its investments on hardware/software resources for data acquisition, storage, processing; to benefit from the access to the partners' data; to enable the better exploitation of its own data.

It should be emphasized the federated nature of the agreement. The Institutions are peer partners; no one 'rules' the Consortium, nor has the power to prescribe technical choices, if not agreed with the partners themselves.

It is also apparent that the federation needs a governance, e.g., a central authority in charge to propose the joint projects; define and maintain the agreed data models; propose novel technological standards; manage the facilities eventually owned by the Consortium as a whole. Open source software technologies help in keeping minimal the cost of the shared resources.

3 Federated computer resources

The *Distributed Architectures (DA)* were proposed in the domain of Computer Systems already in the eighties of the past century, i.e., when the digital networks were still being explored.

The idea is to assign the computations not to a single computer, but to a number of connected, co-operating units called *nodes.* The nodes may be hardware components (processors); sensors, like video-cameras or remotely-operated instruments; database subsystems; entire computers; local computer networks,…..which gives room to a wide range of real-life situations.

The advantage is the opportunity to manage huge amounts of information, not affordable by a single units: large archives, just like the client accounts of a bank; data flows, e.g., from a user to her/his personal account; computing speed of operations,….

Other chances are exploiting the decentralization, e.g., suggested by the logistics of an enterprise; the operational independence, e.g., local teams manage local nodes without relying on the intervention of externals.

Conflicting issues are the autonomy of the nodes vs. data sharing, and the co-operation to achieve common goals. Therefore, the distributed architectures range from the *tightly-coupled* ones -- lower autonomy of nodes, higher computational efficiency, as in multi-processor systems -- up to the *loosely-coupled* systems, with higher autonomy of the individual nodes and the need to coordinate them. The advent of the Internet has boosted dramatically such architectures, and distributed computer systems are now ubiquitous over the Net.

3.1 Federated architectures

Among the loosely-coupled ones, the *federated architectures* originated the *federated systems [1].* They foster the independence of the component nodes, even with heterogeneous technological platforms, provided that adhere to an abstract vision of how to share the information.

The federated framework applies to large enterprises or organizations, possibly on a large spatial scale. They preserve the highest possible autonomy of each unit, and achieve high flexibility, i.e., local resources and peculiarities are maintained, and problems solved locally. Of crucial importance is achieving the *interoperability of data*, by which we mean that data originally encoded and stored on a hardware-software (HW-SW) platform, can be read, interpreted and processed elsewhere on a different HW-SW platform.

As much as in a federation agreement, a governance is necessary for a federated technological system, according to a vision of how to share information, as we shall detail in the sequel.

3.1 Semantic data models and exchanges

Sharing data is not merely accessing a site via the download/upload procedures: the Net basic services would do the job with no need of complex architectures. Instead, a crucial problem is sharing the *meaning (semantics)* of the data themselves, which is a profound topic to be addressed.

First: the data are to be named according to a shared vocabulary of terms and associated meanings. The Institutions should agree on such a vocabulary: concepts, physical objects, relations among them,... Fortunately, this task is constrained by the operational context – i.e., the archaeological research on Aquileia, in our case.

Second point: how do we describe each piece of data? We need to define – and agree on – a description scheme, i.e., a structured record of *metadata:* data about acquisition parameters; their technical features; classification; ownership; administration.

Therefore, two layers of a semantic data model are envisaged:

- A basic, conceptual one fixing the concepts, objects and terms to denote them;
- An upper layer, built on top of the former, containing structured records of data descriptions, aimed at indexing them for efficient storage and retrieval.

Each federated Institution should adhere to the semantic model, in order to read, interpret and process the data consistently with the other partners: it's a fundamental step towards the interoperability of the partner resources.

Once a semantic basis has been established, a main issue is *how to exchange data*. Problems to be faced are the heterogeneous technological platforms of the nodes. On general grounds, data are exchanged as messages by means of protocols, i.e., standard formatting rules and procedures to transmit/receive bit packets over the Net.

Adequate technologies are the *Web Services,* international standards for data exchange in distributed environments. *Services* are units on a computer to accomplish definite tasks -- such as placing an on-line booking of an airline ticket. A Web service is a scheme for organizing and using such units, that in principle can be controlled by different technological and ownership domains.

Importing and exporting data from one domain to another is one such service: therefore, Web services are candidate components of a federated architecture. An amount of software design is required in order to develop such services.

Alternative solutions exist for exchanging data. In case a node hosts a database, *partial views* of the latter can be defined in order to enable an external user to access a portion of the local data. Another, simple-minded solution is uploading data files from a local to a centralized repository: it should be operated under the responsibility of the owner administration, and perhaps scheduled on a regular time basis, in order to keep mutually aligned the two repositories.

Last - not least! - component of a federated system is the interface providing partners with a unique access to the shared resources, according to the agreed data model, which means, access to the available databases and exchanges with the partner data repositories.

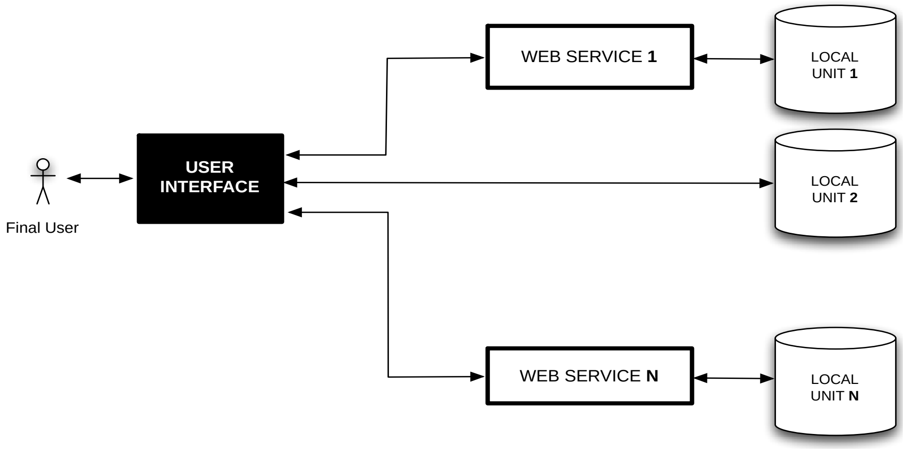


Fig. 1. Abstract scheme of a federated computer architecture. N local units are connected to a user interface via the Internet. The Units 1 and N exchange data via web services; Unit 2 is connected simply by the basic Internet services.

4 Proposals for Aquileia

Specific technological solutions are needed for an integrated access to the Aquileia data heritage. A large amount of information is made of spatial data, i.e., directly referred to the terrain – georeferenced - by means of international standard coordinate systems, like the UTM (Universal Transverse Mercator) or the GPS (Global Positioning Satellite).

An appropriate choice is the GIS (Geographic Information System), i.e., a complex software system – not merely an interface -- that displays georeferenced data; processes them and extracts thematic layers; classifies and archives them; allows for data query/retrieval by means of a search engine as in ordinary database systems.

4.1 Integrating sparse data

Three solutions support the effective data sharing among the partners:

- The georeferencing mechanism: data referred to a spatial coordinate system can be correlated and superimposed by the GIS into a unique working framework. As an example, the home page of ANTEO, an open-source GIS operating on the Web has been reported in Figure 3.
- The central repository, i.e., a large disk hosting data that Institutions have made available, that could not export from their own site. For example, historical maps to be digitized, or data recorded on analogical supports. The central repository ensures compliance to an agreed digital encoding standard, as well as a standard

metadata classification scheme. Consequently, it guarantees a uniform access to the data by all the partner nodes;

- The query engine ensures the query/retrieval of data: either directly, via the central repository (previous point), or via the other solutions (Section 3.2) to import/export data from each Institution.

Stated differently, the technological governance of a federated system is grounded on a GIS and a repository, that provide uniform access, indexing, processing, query and retrieval of the shared data sources.

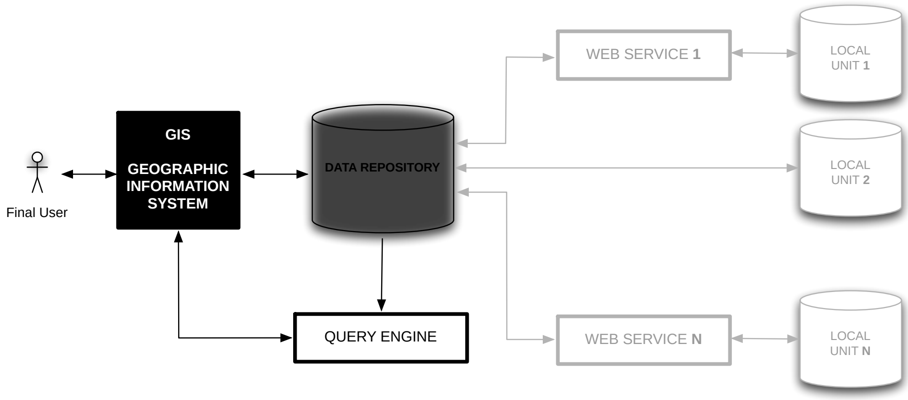
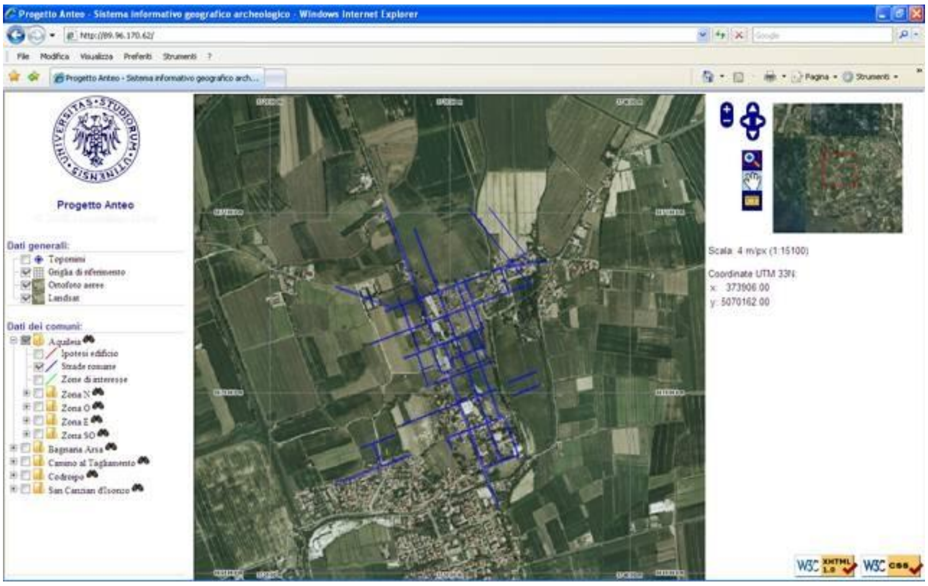


Fig. 2. Proposal of a federated architecture for sharing the Aquileia data heritage, with reference to Fig.1. A centralized data repository queries and retrieves data from each local unit. It also directly hosts data after digitization and format conversion, in order to make them available to all units. The access to the system is through a GIS.


Progetto Anteo - Sistema informativo geografico archeologico - Windows Internet Explorer
http://89.96.170.62/
File Modifica Visualizza Preferiti Strumenti ?
Progetto Anteo - Sistema informativo geografico arch...
Pagina
Strumenti
UNIVERSITAS • STUDIORUM • UTINENSIS
Progetto Anteo

Dati generali:

- □ Toponimi
- ☑ Griglia di riferimento
- ☑ Ortofoto aeree
- ☑ Landsat

Dati dei comuni:

- ☑ Aquileia
  - □ Ipotesi edificio
  - ☑ Strade romane
  - □ Zone di interesse
  - □ Zona N
  - □ Zona O
  - □ Zona E
  - □ Zona SO
- □ Bagnaria Arsa
- □ Camino al Tagliamento
- □ Codroipo
- □ San Canzian d'Isonzo

Scala: 4 m/px (1:15100)

Coordinate UTM 33N:
x: 373906.00
y: 5070162.00

Fig. 3. The ANTEO Open Geographic Information System [3]. Centre window: an aerial image of the Aquileia (july 2007), with superimposed the grid of roads of the Roman city.

4.2 Encoding, archiving and beyond…

The heterogeneous, non-standard data on Aquileia are to be analyzed, and their current status fixed. Problem is selecting international standard technologies for encoding, indexing, archiving them, in order to ensure the interoperability of the shared data sets.

Supposing that the digital encoding can be addressed by widespread technologies – just a working hypothesis for the purposes of this paper – the indexing and archiving problems deserve a major design effort.

Both issues pertain to what has been called ‘semantic data model ‘ in Section 3.2. In order to share data, we must agree on a unique vocabulary of terms denoting the data themselves along with their physical attributes and properties; the relations among them; their purposes: what in the ICT is called an *ontology [4]*.

Ontologies address the need of a semantic data model. They suggest descriptions of the data themselves in terms of properties and parameters: the *metadata,* which address the issue of classifying and indexing, to archive and retrieve data in a database – just like the book cards in a public library.

Once more, problem is defining such ‘data card’ formats according to an international standard scheme. A practical solution is given by the *Dublin Core* initiative [5].

SIMPLE DUBLIN CORE METADATA ELEMENT SET (DCMES)

1. Title	9. Format
2. Creator	10. Identifier
3. Subject	11. Source
4. Description	12. Language
5. Publisher	13. Relation
6. Contributor	14. Coverage
7. Date	15. Rights
8. Type	

Fig. 4. A basic set of data descriptions according to the Dublin Core international initiative. The Metadata Element Set is a vocabulary of fifteen properties for use in resource description. Its elements are broad and generic, usable for describing a wide range of data resources.

We should keep in mind that local databases - with their own data description cards – are likely to conflict with the common data model. Mapping the local format into the shared one, and vice-versa, is a non-trivial issue to be addressed at a later time, for each single unit, after a careful data analysis.

5 Conclusions

We have proposed some basic ideas of a project to share data on Aquileia and overcome the current dispersions; to enable a better exploitation of the data heritage and foster international collaborations.

A federation agreement among the interested Institutions should support such perspectives, and inspire a research Consortium, grounded on the autonomy of each Institution; the protection of the investments and technological choices of each; the willingness to share a vision of the data, as well as exploit the opportunities of the Open software and international standard solutions. A governance of the Consortium ensures the proposal and maintenance of the agreed software standards.

On this basis, we propose a federation of computer systems, as a kind of distributed, specialized computer network. A set of nodes, each pertaining to an Institution, are connected via the Internet and the Web, and share a semantic data model. Accordingly, well-experimented technologies ensure the exchange of data over the Net with minimum cost. The technological governance based on Open-source products ensures the uniform access to the federated resources through a GIS platform.

An exploratory international project can perform a feasibility study, and a few results are under reach with low cost in a reasonable time:

- An inventory of the data made available by the Institutions, along with a preliminary analysis of their features;
- Proposals towards a semantic data model: an analysis of concepts and terms towards a shared ontology;
- Proposal of metadata formats;
- An Open GIS platform for a uniform access to georeferenced data;
- Experimental Web services to explore effective data exchanges among Institutions through the Internet.

We believe that integrating local databases within more complex schemes – e.g., retrieving multiple pieces of data from different archives by joint queries, as in the Online Public Access bibliographic Catalogs (Meta-OPAC) – is a difficult task, not under reach until a careful analysis of the data resources has been carried out.

However, we are confident that the ICT solutions we have proposed open new perspectives towards an international effort of data sharing on Aquileia.
The Institutions will be aware of the data hosted elsewhere. They will possibly access pieces of data collected by other Institutions with minimum cost. Data will be correlated within a unique GIS platform, enabling their joint interpretation. Novel data clusters will be available and bring to the synthesis of new knowledge.

Another perspective opened by a federated system is the fast and consistent data archiving: excavation, geophysical, aerial survey data can be registered and archived by *mobile devices* according to standard formats. In this way, they will be made available in short time by the owners to the users – researchers, schoolboys and girls, tourists, citizens of any Country - upon authorization.

References

1. Heimbigner, D.M., McLeod, D.: A Federated Architecture for Information Management, ACM Trans. On Information Systems (TOIS), Vol.3, 3, The ACM Press, New York, NY, USA (1985)
2. Geographic Information Systems (GIS), http://library.stanford.edu/depts/gis/index.html
3. Buora, M., Roberto, V.: New Work on the Plan of Aquileia based on Aerial Photographs and a GIS Platform, Journal of Roman Archaeology, Vol.23, pp.320--334, USA (2010)
4. Gruber,T.: A Translation Approach to Portable Ontology Specification, http://tomgruber.org/writing/ontolingua-kaj-1993.pdf (1993)
5. The Dublin Core Metadata Initiative, http://dublincore.org/documents/dces/