Integration of Complementary Archaeological Sources

Martin Doerr[1], Kurt Schaller[2] and Maria Theodoridou[1]

[1]Institute of Computer Science – Foundation for Research and Technology – Hellas
{martin, maria}@ics.forth.gr
[2]Forschungsgesellschaft Wiener Stadtarchäologie, Vienna, Austria
k.schaller@ubi-erat-lupa.org

Abstract. We are developing a knowledge base that integrates complementary archaeological information sources. Our source data comprise complementary scientific databases and corpora describing finds with inscriptions and iconography of the Roman era. The integration of such complementary information is innovative and of immense potential value for the cultural heritage domain. Integration is achieved by intellectually interpreting each source schema in terms of the CIDOC CRM model and storing it in an RDF knowledge base, thus creating a body of unique archaeological knowledge in digital form. Our main objective is to provide procedures for information extraction and global querying over all the contents of the complementary resources. Additionally we aim at performing reliable statistical evaluation of the integrated data. In order to ensure that the methods used converge towards the best state of knowledge available and that the results are of high quality, we apply data cleaning procedures both at the individual sources and at the integrated knowledge base.
Keywords: knowledge base, complementary information, integration

1. Introduction

When useable relational database applications – not longer restricted to be created and employed exclusively by esoteric 'code cracks' using cabalistic ciphers – appeared on the market in the early eighties of the last century, they soon turned out to become an indispensable tool for archaeologists. The typological structure of archaeological data perfectly matched the possibilities of storing and exploring large amounts of information offered by tables and arrays. Since the internet revolution of the second half of the nineties many of these digital data collections are also becoming available as on-line resources. A constantly growing number of easily accessible archaeological web-databases actually provides a substantial increase of available knowledge but unfortunately Mr. Bill Gates and his competitors did not yet get around to establishing industry standards for the description of Roman, Greek or Hebrew inscriptions, Sumerian vessels or Germanic metal fittings. In practice many data sources contain partly complementary, partly overlapping and partly contradictory information held in heterogeneously structured databases that are using a multitude of data formats. In order to prevent information overflow (e.g. by allowing cross domain searches) and trying to turn the chippy data into consumable information that can be statistically evaluated, there is a high demand for the integration of these data while archaeologists on the other hand still tend towards establishing private terminologies to describe their material, backed up by divergent traditions of the numerous archaeological disciplines and their national peculiarities.

1.1 Problem Statement

Archaeology has a huge amount of well elaborated corpora of highly interrelated, overlapping and complementary character. Current data base technology provides sophisticated tools for continuously updating and searching vast amounts of information. The Semantic Web technology and activities provide new opportunities in better enabling archaeologists to integrate and exploit these data. The integration of such complementary information is innovative and of immense potential value for the cultural heritage domain.

Traditional archaeological corpora are comprehensive, organized collections of ancient data, collected and described by hundreds of scholars over the preceding centuries. They have been and continue to be of fundamental importance as an authoritative source for the study of classical antiquity, providing very high quality of information. However, their maintenance is extremely difficult since it is a lasting, centralized process, which is not any longer supported by the current research policies. Moreover, their paper form increases the difficulty of updating and searching the contents while it is almost impossible to correlate the information with other complementary or overlapping resources. On the other hand, current database projects provide quick access to rapidly growing data of varying quality. Neither those are integrated, and the contents are increasingly overlapping. We would like to combine the quality of the traditional corpora with the ease of access of modern electronic data management.

1.2 Working Context

The VBI-ERAT-LVPA database project (LUPA 2000) has started in the early nineties as an integrative study of the iconographic and epigraphic aspects of Roman stone monuments. Drawing together these aspects, that are traditionally covered by different academic disciplines and published in specialized printed corpora like Corpus Signorum Imperii Romani (CSIR) and Corpus Inscriptionum Latinarum (CIL 2004) soon evoked the demand for linking and integrating LUPA with other, more specialized data

sources like ARACHNE (ARACHNE 2004), the Epigraphic Data Bank Clauss / Slaby (Clauss 2003) or the ONOMASTICON – a corpus of Roman names, currently only available as printed publication (OPEL 2003). The criteria for such a common knowledge base were concertedly formulated by all partners and associates of the LUPA project, the realization and implementation was overtaken by ICS FORTH in Crete, a project partner with the necessary expertise and experience in this field.

1.3 Objectives

We propose an information integration scenario, which aims at bringing out the value of cultural heritage domain information by creating a body of unique archaeological knowledge in digital form, out of the aggregation of complementary archaeological sources of overwhelming detail and volume. Our main objective is to provide procedures for information extraction and global querying over all contents of the complementary resources and to perform reliable statistical evaluation of the integrated data. We would like to ensure that the methods used converge towards the best state of knowledge available and that the results are of high quality.

On the Semantic Web applications we cannot make the assumption that databases are maintained appropriately or modified according to the needs of the integration. In our application environment, our partners are willing to share and improve their databases while at the same time they want to keep their autonomy. Taking this into account, we aimed at developing data cleaning procedures that ensure quality improvement both at the individual, autonomous sources and at the integrated knowledge base.

2. Approach

In this section we will describe in detail the approach we followed in order to integrate complementary archaeological sources. Our work was motivated by the information integration scenario described in (Calvanese, De Giacomo, Lenzerini, Nardi and Rosati 1998).

Central to our approach is the notion of a domain model, which is a conceptual representation of the global concepts and relationships that are of interest to our application. For our application domain, we chose as our domain model the CIDOC CRM model, a high-level ontology which enables information integration for cultural heritage data and their correlation with library and archive information (CIDOC 2004, Doerr 2003). CIDOC CRM provides the basis for the integration while thesauri, digital gazetteers and possibly other background knowledge also contribute in the creation of an integrated knowledge repository based on RDF descriptions, as shown in Fig. 1. RDF (RDF 2004) was chosen because it is a universal format for data on the Web that abstracts knowledge from the documentation units and from the perspective in which they were produced. For example, it allows for analysing knowledge independent from if it were

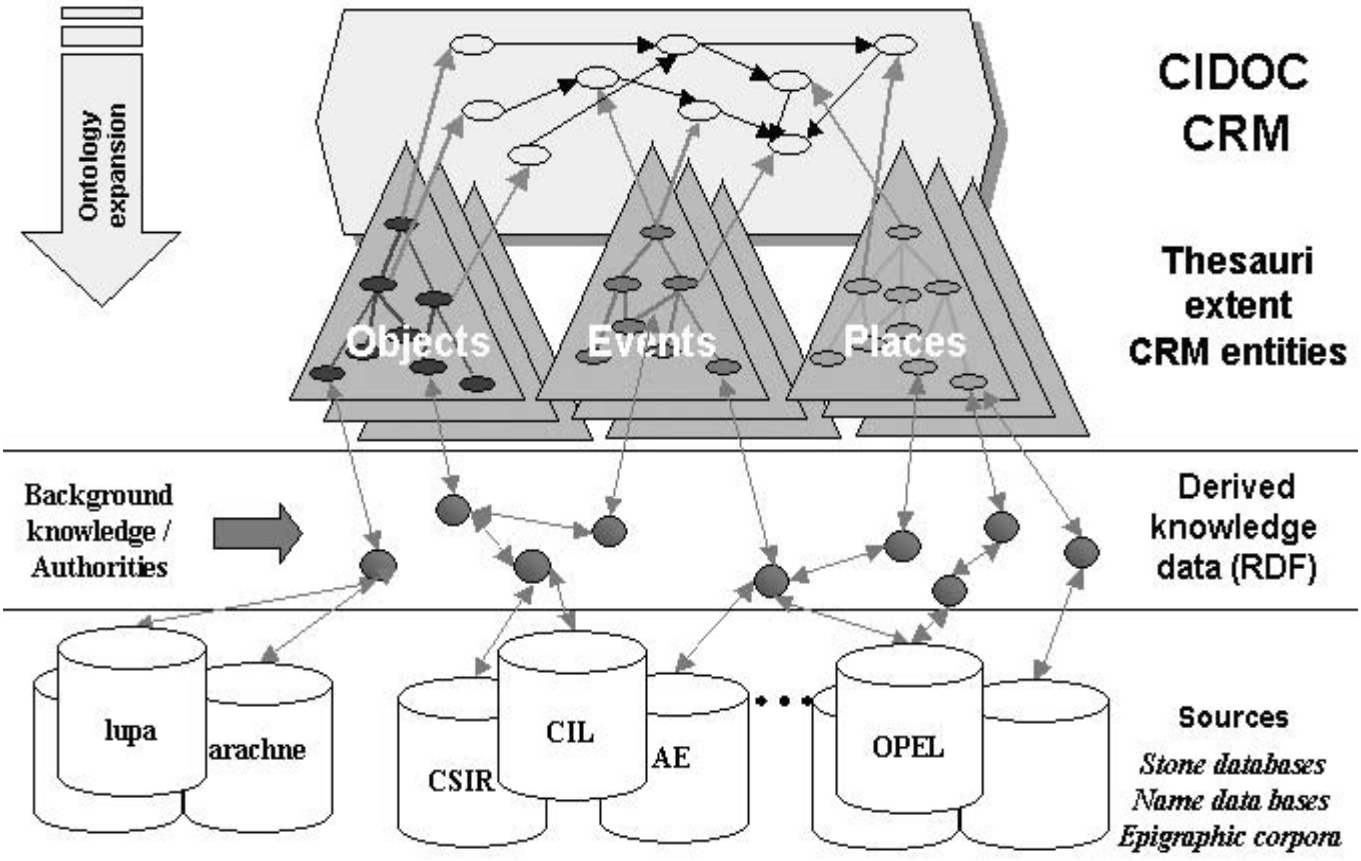

Fig. 1. The CIDOC CRM – VBI-ERAT-LVPA Repository Indexing.

initially part of the description of a stone, an image, an inscription, a dictionary etc.

The mapping is possible due to:

- The establishment of source models that provide conceptual representations for each category of source data. These models are not necessarily complete (representing all knowledge in the source data) but are sufficient for the application demands on the domain model.
- The establishment of strict rules for the format of the source data. These rules include spelling of placenames, disambiguation of broader or narrower placenames, format of citations, data separators etc.
- The definition of an algorithm that expresses the transformation from the source data to the domain model. The algorithm is implemented using commercial conversion tools, Java and JavaCC (JavaCC 2004).

One consumer of such an integrated model might be a cross-domain search service. If the information integration is correct then a query to the domain model should produce the same answer as that yielded by applying the respective query over each of the sources. Additionally, due to the complementarity of the source data, it should be possible to draw inferences from the combined knowledge of the individual sources. For example, the query "In which coordinates were found tombstones sawing a specific name" cannot be answered by any of the individual sources. However, the integrated repository, where the individual overlapping and complementary information has been combined into a network of integrated knowledge, can give an answer to such a query. The information regarding the tombstones has been integrated with the one of the epigraphic corpora (CIL 2004, AE 2000), the corpus of Roman names (OPEL 2003) and digital gazetteers (ADL 2004) as shown in Fig. 2 and thus it is possible to infer the answer to the query. The integrated repository will be continuously augmented by new sources.

A second consumer might be a service that performs statistical evaluation of the data. The information integration has to ensure that the methods used converge towards the best state of knowledge available and that the results are of high quality. We aimed at developing a variety of data cleaning procedures that ensure quality improvement both at the individual, autonomous sources and at the integrated knowledge base allowing incremental updates without loss of information.

Archaeologists improve the quality of the individual archaeological sources continuously. We foresee both semiautomatic and manual data cleaning procedures which will facilitate their task. We follow strict rules regarding the format of the data, in particular unstructured data such as references and citations. If during the transformation from the source data to the domain model we identify any "non canonical" data we report them to the respective source and so mistakes can be removed by the respective partner and the quality of its source be improved.

In the integrated repository, duplicates that can be automatically or manually detected are removed from the integrated base, thus improving the state of the knowledge built. In the future, we would also like to provide a mechanism to report possible duplicate objects. These reports have to be evaluated by an expert that will validate the duplicates as such and proceed with manually cleaning the source. When the sources are changed we want to make sure that information will not be lost during the incremental update of the integrated knowledge base over time.

3. Technical Description

We created a central RDF repository that receives information from a number of satellite applications. The repository provides the integrated information access while the satellite applications, keeping their autonomy, continue to function as channels for data collection. Transfer of data from the satellite applications to the central repository is being provided in a standardised "migration" format. Integrating incompatible data sources requires some sort of "bridge" to be constructed between them. The physical separation and logical autonomy of data sources could thus be overcome: the migration format

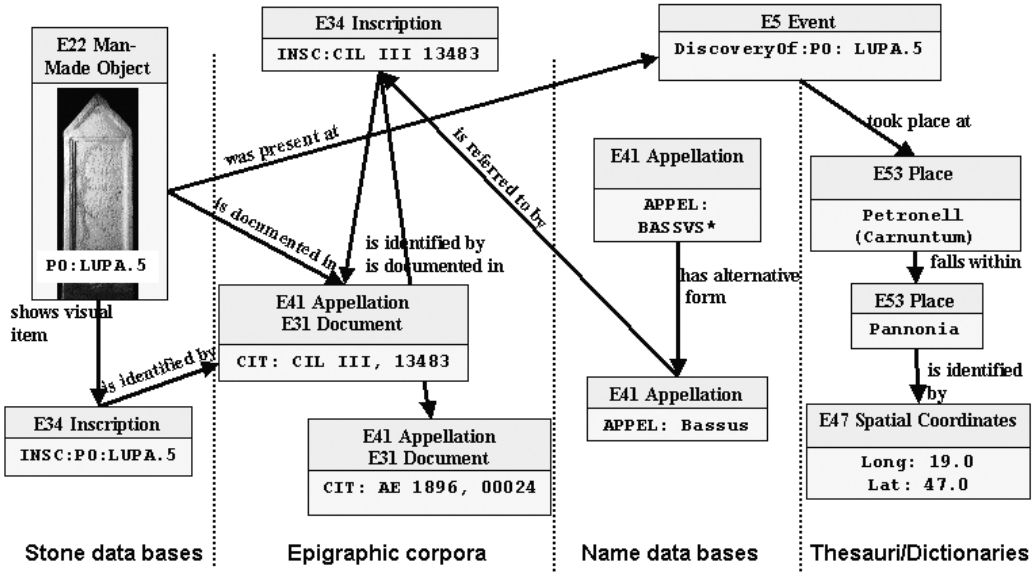

Fig. 2. Integrated knowledge repository.

provides a conceptual and technical "target" for data export from satellite programmes so that software incompatibilities and differences between data schemas can be ironed out. It was decided that the transport format itself will use XML, so that existing applications would require only minor technical modification in order to become effective "satellites" of the central repository. The central repository is capable of reading any data set provided in this XML format, so the number of satellite applications can be progressively increased without the need for further modifications to the central database.
We identified three steps needed to put this architecture into place:

- Ontological and semantic analyse of existing data sources in order to arrive at a common conceptual schema
- Design of the migration format based on this schema
- Transformation of the data into the migration format

We developed a procedure, which transforms database files to xml files in a semiautomatic way. Since CIDOC CRM is the formal ontology that we use as a common conceptual schema, the XML files that we produce during the transformations are compatible with CIDOC CRM.

3.1 Data Quality

Ensuring the quality of data in information systems is crucial for decision-support and research-oriented applications. Data quality concerns arise in three different contexts: when one wants to correct anomalies in a single data source (e.g. inconsistent use of field separators in bibliographic references); when poorly structured or unstructured data is migrated into structured data (e.g. dates in a data field of a place); or when someone wants to integrate data coming from multiple sources into a single new data source (e.g. duplicate elimination). The main goal of a data cleaning process is to eliminate anomalous data in each of these situations (Galhardas, Florescu, Simon and Shasha 2000).
In our application domain, the owners of the individual sources are willing to share and improve their data bases while at the same time they want to keep their autonomy. We assume that for each individual source there exists a 1–1 relationship between objects and identifiers. Additionally, references to other sources (e.g. place names, citations) follow specific, strict rules. In the integrated repository, two local identifiers from two different data bases may denote the same object.
We propose a proactive data cleaning procedure which reports mistakes, non-canonical data, and other detected errors to the respective source data bases where cleaning will be done manually by the archaeologists after they evaluate the reported information.
One of the most problematic issues that any data integration system has to confront with, is the existence of multiple records without a common identifier for the same object. Cleaning data coming from multiple sources needs to identify overlapping data, in particular matching records referring to the same real-world entity. This problem is also referred to as the object identity problem, duplicate elimination or the merge/purge problem (Galhardas, Florescu, Simon and Shasha 2000). Frequently, the information is only partially redundant and the sources may complement each other by providing additional information about an entity. Thus duplicate information should be purged out and complementing information should be consolidated and merged in order to achieve a consistent view of real world entities (Rahm, and Do, 2000).
There exist two possible ways to approach the identity problem. In the first, we try to find global names with a high chance to match. There is a risk of overmatching, with the result that the merged properties of the matched objects cannot be separated afterwards. Although this approach has a better recall, there exists information that is lost. In the second, two objects are regarded different unless proven differently. Objects are identified by their initial source data base identifiers where uniqueness is guaranteed and the autonomy of each source is preserved. The integrated knowledge base preserves the initial source information ensuring better precision.
For example, two stones for which we know only their size and where they were found cannot be compared or regarded as the same object. If, however, they both have the same reference of an inscription then we can assume that they are the same object even if the place found is not the same. In this case there is a possible mistake that has to be evaluated by an expert and corrected in the respective sources. It is theoretically impossible to find all duplicates and to make sure that an identified duplicate is not in reality two different things. Duplicate removal mechanisms are a cost-benefit optimization of over- and under-identification, and manual intervention is inevitable.

3.2 Data Cleaning Procedure

Consequently we propose a reactive data cleaning process which removes as many duplicates as can be (semi-) automatically detected. On removing a duplicate we maintain in addition the object identifiers from the individual sources, such that updates from the sources can be directly matched with the ultimate identifier, even after manual duplicate removal.
We will present in this section an example. Let us consider the source data base of LUPA that contains archaeological records regarding roman stones. Each stone has a unique serial number identifier ll. Respectively, a stone in the ARACHNE data base has also a unique serial number identifier aa.
In the integrated repository, both objects are mapped according to CIDOC CRM as 'E22 Man-Made Object'. A global index is constructed where each object is assigned a serial number based on its source data base and its initial serial number. With this approach objects preserve their original source data base serial number identifier as a partial identifier, while at the same time it is possible to identify objects from different source databases. So, the LUPA object will get the identifier 'P.O.: lupa.ll' while the ARACHNE object will get identifier 'P.O.: arachne.aa'. Source data bases reference external third party sources such as inscription corpora, dictionaries, name data bases etc. Two objects coming from two different data bases can be identified as one through a common property which is identified from a third shared context.

As it can be seen in Fig. 3. object 'P.O.: lupa.2849' and object 'P.O.: arachne.80581' are both referencing the inscription 'INSC: CIL III 10514'.

Thus, the two objects are the same and the system will automatically extent the knowledge built in the integrated repository by adding the properties of the second object to those of the first as it can be seen in Fig. 4.

Following the elimination of the duplicate object 'P.O.: arachne.80581', the objects 'OID: arachne.80581' and "Title: Grabstele des Nertus" are linked to the object 'P.O.: lupa.2849' since we want to be able to support incremental updates. We maintain all the local identifiers in the global index as valid names and remove detected duplicates continuously.

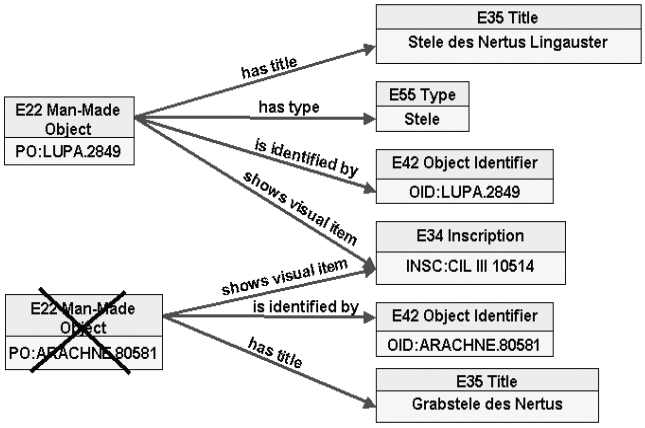

Fig. 3. Reactive Data Cleaning Initial Data.

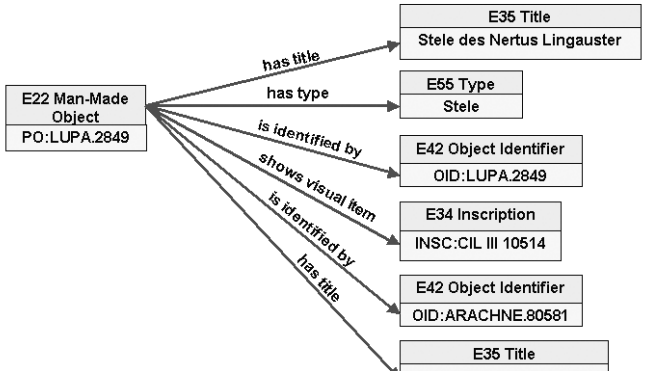

Fig. 4. Reactive Data Cleaning Result.

4. State of Work and Experience

The implementation of our system is based on three sets of tools. The first includes the set of transformation/mapping tools. Its goal is to convert the data of various formats and sources into a common XML format compatible with the CIDOC Conceptual Reference Model. The input data come from the diverse archaeological sources and might be databases, text files, spreadsheet files etc. The results of the transformation are XML files, compatible with the CIDOC Conceptual Reference Model. The set of classes and properties of the CIDOC CRM that are used for the needs of the UBI-ERAT-LVPA project are shown in Fig. 5.

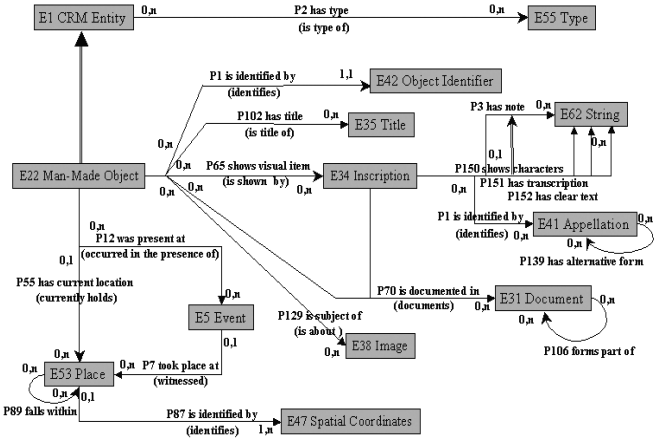

Fig. 5. CIDOC CRM classes and properties used by VBI-ERAT-LVPA.

It became apparent that approximately 10% of the CIDOC CRM model was sufficient for the needs of the application domain (13 classes out of 84 and 13 properties out of 139) while three new properties (P150F shows characters, P151F has transcription and P152F has clear text) had to be defined as subproperties of the property P3F_has_note of the class E34_Inscription in order to cover the specific needs of the archaeologists regarding the information of the text of the inscriptions.

Most of the transformations are done using DataJunction 7.5 (DataJunction 2004), an integration/migration tool, market leader in its field, designed to convert structured data from one format to another. It is also designed to clean and restructure the data to fit the new format. Additionally, Java and JavaCC (JavaCC 2004) were also used for specific transformations, particularly for dictionary entries, such as OPEL (OPEL 2003). Finally, XML files are mapped to RDF descriptions through a specialized converter program.

The second set of tools is the RDF Suite (RDFSuite 2003) which allows for effective and efficient management of large volumes of RDF descriptions. The produced RDF descriptions are validated with a Validating RDF Parser (VRP) and then loaded in the RDF Schema-Specific Data Base (RSSDB), a persistent RDF store that is used for the integrated knowledge repository.

Finally, we are modifying the SWPG, Semantic web portal generator (Athanasi 2004) in order to provide a Web-based easy to use by archaeologists user interface.

The user interface will allow the formulation of three types of queries:

1 Data cleaning queries, which will either produce reports that will be mailed to the respective source in order to improve and/or correct the data manually or activate procedures that will clean the integrated repository automatically. For example:

"Find the inscriptions of Lupa that have transcriptions that differ from the respective transcriptions of CIL"

The answer to such a query indicates a mispelling and will be reported to the interested parties.

"Find the inscriptions that are referenced by two distinct stones"

This query will trace duplicates that will be eliminated from the integrated base.

2 Queries of archaeological content to the integrated knowledge repository. With these queries we can detect contextual relationships that cannot be derived from interpreting the sources in isolation. For example:
"Which names appear in a specific region?"
"Name X which appears on a stone belongs to an important Roman person. In which other stones do we have the same name?"
"In which coordinates were found tombstones sawing a specific name"
3 Statistical queries. For example:
"How often, and in which regions appears a specific name?"

5. Conclusions

The main result of this work is the development of a method and tools for the integration of diverse archaeological information on the Roman stone monuments, such as the Corpus Inscriptionum Latinarum, the Onomasticon of Roman personal names, the VBI-ERAT-LVPA archaeological database in Vienna, the archaeological database ARACHNE in Cologne. The system will continue to be expanded with the addition of new sources in the future and we are investigating ways to support an automatic mapping process so that archaeologists will be able to maintain the system.
The highlights of this work are summarized in the following:

- creation of a global index about a set of semi-autonomous archaeological bases and corpora on the Roman stone monuments, for global access to the unified knowledge
- integration of complementary information under the common CIDOC CRM ontology/schema and identification of common elements in different sources
- development of an integration algorithm that converges to the best state of knowledge and continuous update
- creation of a research tool for formulating queries and drawing conclusions of archaeological content to detect contextual relationships that cannot be derived from interpreting the sources in isolation
- development of a method for identifying epigraphic references and finds
- development of an efficient way for place name recognition
- a very good test bed for the CIDOC CRM model that proved its adequacy. This work demonstartes that CIDOC suites very well the needs of applications that handle any kind of cultural heritage material although it was not designed for these specific data. This suggests that there is no need for every project to develop its own schema as is mostly assumed.

To our knowledge, it is the first large-scale integration project in the cultural heritage domain that creates a global index of multiple complementary resources.

References

ADL 2004. Alexandria Digital Library Gazetteer home page http://www.alexandria.ucsb.edu/gazetteer/

AE 2000. L'annee epigraphique, 1902 onwards, annual review of publications on Roman epigraphy http://www.anneeepigraphique.msh-paris.fr/

ARACHNE, 2004. Die computergestützten Datenbanken des Forschungsarchivs für Antike Plastik Köln, www.arachne.uni-koeln.de

Athanasis, N., 2004. SWPG: Semantic web portal generator, Master's Thesis, Computer Science Department, University of Crete, available in Greek: http://www.ics.forth.gr/isl/publications/paperlink/athanasis.pdf (visited 7/7/2004)

Calvanese, D., De Giacomo, D., Lenzerini, M., Nardi, D. and Rosati, R., 1998. Description logic framework for information integration. In *Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR'98)*. 2–13.

Clauss, M. 2003. Epigraphic Data Bank Clauss / Slaby www.uni-frankfurt.de/~clauss

CIDOC, 2004. The CIDOC Conceptual Reference Model, Official Web Site, 2004 http://cidoc.ics.forth.gr/

CIL 2004. Corpus Inscriptionum Latinarum a Th. Mommsen a. 1853, http://www.bbaw.de/forschung/cil/

DataJunction, 2004. DataJunction Migration Toolkit, http://www.pervasive.com/migrationtoolkit

Doerr, M., 2003. The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata, *AI Magazine*. 24(3).

Galhardas, H., Florescu, D., Simon, E. and Shasha, D., 2000. Declaratively Cleaning your Data using AJAX, *Journées Bases de Données Avancées (BDA), October, 2000*.

JavaCC 2004. Java Compiler Compiler – The Java Parser Generator home page https://javacc.dev.java.net/

LUPA 2000. VBI-ERAT-LVPA project home page: www.ubi-erat-lupa.org

OPEL 2003. ONOMASTICON PROVINCIARVM EVROPAE LATINARVM, COMPOSVIT ET CORREXIT BARNABAS LOERINCZ, EDITIO NOVA AVCTA ET EMENDATA. Wien, 2003.

Rahm, E. and Do, H.H., 2000. Data Cleaning: Problems and Current Approaches, *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol. 23 No. 4, December 2000.

RDF 2004. Resource Description Framework (RDF) http://www.w3.org/RDF/ (visited 05/06/2004)

RDF Suite, 2003. The ICS-FORTH RDF Suite: High-level Scalable Tools for the Semantic Web home page http://athena.ics.forth.gr:9090/RDF/

Sheth, A., Thacker, S. and Patel, S., 2002. Complex Relationship and Knowledge Discovery Support in the InfoQuilt System, *VLDB Journal*, September 25, 2002 (on-line publication; ISSN: 0949-877X), 12 (1) 2003 (print publication; ISSN: 1060-8888).