Incremental, Semi-automatic, Mapping-Based Integration of Heterogeneous Collections into Archaeological Digital Libraries: Megiddo Case Study

Ananth Raghavan[1], Naga Srinivas Vemuri[1], Rao Shen[1], Marcos A. Goncalves[2], Weiguo Fan[1], and Edward A. Fox[1]

[1] Digital Library Research Laboratory, Virginia Tech, Blacksburg, VA 24061
{ananthr,nvemuri,rshen,wfan,fox}@vt.edu
[2] Department of Computer Science,
Federal University of Minas Gerais,
Belo-Horizonte-MB Brazil 31270-901
{mgoncalv@vt.edu}

Abstract. Automation is an important issue when integrating heterogeneous collections into archaeological digital libraries. We propose an incremental approach through intermediary- and mapping-based techniques. A visual schema mapping tool within the 5S framework allows semi-automatic mapping and incremental global schema enrichment. 5S also helped speed up development of a new multi-dimensional browsing service. Our approach helps integrate the Megiddo excavation data into a growing union archaeological DL, ETANA-DL.

1 Introduction

During the past several decades, Archaeology as a discipline and practice has increasingly embraced digital technologies and electronic resources. Vast quantities of heterogeneous data are generated, stored, and processed by customized monolithic information systems. But migration or export of archaeological data from one system to another is a monumental task that is aggravated by peculiar data formats and database schemas. This problem hampers interoperability, long-term preservation, and reuse. The intermediary-based approach is one way to address the interoperability problem [1]. It uses mechanisms like mediators, wrappers, agents, and ontologies. Yet, while many research projects developed semantic mediators and wrappers to address the interoperability issue, few tackled the problem of (partial) automatic production of these mediators and wrappers (through a mapping-based approach).

The mapping-based approach attempts to construct mappings between semantically related information sources. It is usually accomplished by constructing a global schema and by establishing mappings between local and global schemas. However, in archaeological digital libraries it is extremely difficult to construct a global schema that may be applied to every single excavation. Archaeological data classification depends on a number of vaguely defined qualitative characteristics, which are open to

personal interpretation. Different branches of Archaeology have special methods of classification; progress in digs and new types of excavated finds makes it impossible to foresee an ultimate global schema for the description of all excavation data [2]. Accordingly, an "incremental" approach is desired for global schema enrichment.

We explain how all these DL integration requirements can be satisfied, through automatic wrapper generation based on a visual schema mapping tool that simultaneously can improve the global schema. Further, in addition to integrating new collections, we extend access to this newly integrated data. We also enhance browsing through our multi-dimensional browsing component, based on the 5S framework.

We demonstrate the integration process through a case study: integrating Megiddo [3] excavation data into a union archaeological DL, ETANA-DL [4] (see Figure 1). Thin black lines show input whereas thick brown lines show output in the figure. First, we analyze the Megiddo excavation data management system based on a formal archaeological DL model [5]. Next we use the visual mapping service to create a wrapper (thick brown lines show wrapper generation and global schema evolution), which converts data conforming to local (Megiddo) schema to the global ETANA-DL schema (thin black lines show data conversion). Initially, the global schema does not contain specific excavation details, but it is enriched during the mapping process. Finally, the converted data is stored in a union catalog, upon which a multi-dimensional browsing service is built (see rightmost arrows). Thus, we describe the entire largely automated workflow (see Figure 1), from integrating new collections into the union DL, to providing services to access the newly integrated data.

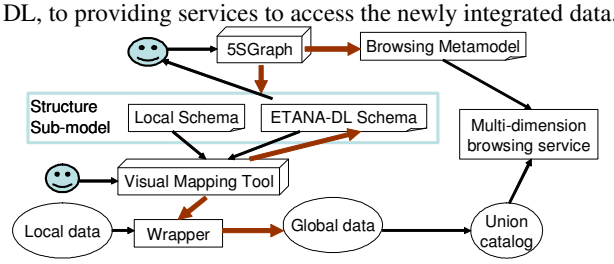

Fig. 1. Process of integrating local archaeological data into ETANA-DL.

The rest of the paper is organized as follows. Section 2 gives an overview of the Megiddo collection. Section 3 describes the visual mapping service provided by ETANA-DL. Section 4 presents the componentized multi-dimensional browsing service module. Section 5 describes the automation achieved through the mapping and browsing services. Conclusions and future work are summarized in Section 6.

2 Megiddo Overview

The Megiddo excavation data integration is used as a case study to demonstrate our approach to archaeological DL integration. Megiddo is widely regarded as the most important archaeological site in Israel from Biblical times, and as one of the most significant sites for the study of the ancient Near East. The excavation data collection

we received from Megiddo is stored in more than 10 database tables containing over 30000 records with 7 different types, namely wall, locus, pottery bucket, flint tool, vessel, lab item and miscellaneous artifact. The Megiddo schema is described in a structure sub-model (see Fig. 1) within the 5S framework (Streams, Structures, Space, Scenarios, and Societies) [6]. Structures represent the way archaeological information is organized along several dimensions; it is spatially organized, temporally sequenced, and highly variable [7]. Consider site organization (see Fig. 2). The structures of sites evidence a containment relationship at every level of detail, from the broadest region of interest to the smallest aspect of an individual find, in a simple and consistent manner. Generally, specific regions are subdivided into sites, normally administered and excavated by different groups. Each site is subdivided into partitions, sub-partitions, and loci, the latter being the nucleus of the excavation. Materials or artifacts found in different loci are organized in containers for further reference and analysis. The locus is the elementary volume unit used for establishing archaeological relationships.

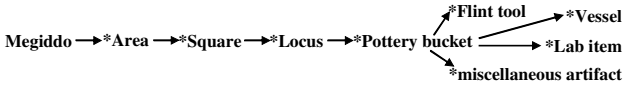

Fig. 2. Megiddo site organization.

3 Visual Mapping Service

In this section we describe our visual mapping service for integrating heterogeneous data describing artifacts from various sites, especially from Megiddo into the union DL. First, we present the architecture and features of the visual mapping component. We then give a scenario-based description of mapping the Megiddo local schema into the ETANA global schema, thus integrating the data describing the various artifacts excavated from the Megiddo site into the union DL. We conclude this section by describing the results of a pilot study comparing Schema Mapper with MapForce [9], for mapping the Megiddo local schema into the global ETANA-DL schema.

3.1 Architecture and Features

Schema mapping is an interesting problem which so far has been addressed from either an algorithmic point of view or from a visualization point of view. In the former case are tools based on a data driven mapping paradigm, like Clio [8] and commercial tools like MapForce [9] and BizTalk Mapper [10]. Based on the latter view, we developed Schema Mapper, the visual mapping component of ETANA-DL, to present local and global schemas using hyperbolic trees [11]. This allows for more nodes to be displayed than with linear representation techniques, and avoids the problem of scrolling. Different colors are assigned to differentiate between root level, leaf, non-leaf, recommended, and mapped nodes (with a color legend present on the GUI – see Fig. 4). A table that contains a list of all the mappings in the current session also is shown.

Schema Mapper recommends matches (global schema nodes) to selections (local schema nodes) made by the user. These recommendations are made using name-based matching algorithms at the schema level and using rules specific to the ETANA-DL domain. The user may or may not choose to accept these recommendations. The recommendations appear in the table shown at the bottom of the screen in Fig. 4.

The other important aspect in integration of data into the union DL is the evolution of the global ETANA-DL schema. Schema Mapper allows global schema editing: deleting nodes, renaming nodes, and adding a local schema sub-tree to the global schema. This has special value for many DLs, e.g., ArchDLs, where it is impossible to predict the final global schema because of its evolutionary nature. Schema Mapper is superior in this respect to commercial mapping tools like MapForce [9] which lack schema editing capabilities. Further, as a global schema evolves, in order to preserve consistency in the naming of semantically similar nodes, Schema Mapper recommends appropriate name changes to global schema nodes, based on the history stored in a mappings database.

Once the local schema has been mapped to the global schema, an XSLT style sheet containing the mappings is produced by Schema Mapper. This style sheet is essentially the wrapper containing the mappings. When applied to a collection of XML files conforming to the local schema, the style sheet transforms it to a set of global XML files, which can be harvested into the union DL. Schema Mapper also saves any changes made to the global schema, and updates the mappings database.

Fig. 3 shows the architecture of Schema Mapper. The Visualization Component contains the logic for generation of hyperbolic trees and other aspects of the GUI. The Recommendation Component makes mapping recommendations as well as recommendations from the Mappings Database for renaming global schema nodes in order to preserve consistency in the naming of similar nodes. The Mapping Component writes the mapping history into the Mappings database and also passes these mappings to the XML Generation Component which generates the style sheet containing the mappings. The global schema is updated by the Global Schema Updating Component.

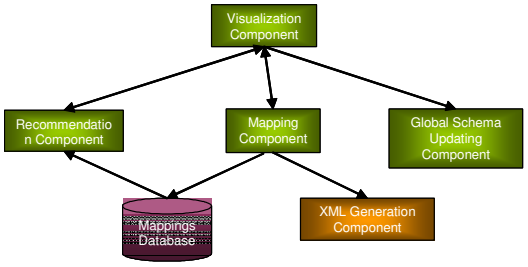

Fig. 3 Architecture of Schema Mapper

3.2 Scenario for mapping Megiddo local schema into ETANA global schema

As described earlier, the Megiddo local schema consists of seven different types of artifacts. For integrating items into the union DL, we produce one style sheet of mappings per item. For the purpose of integration of the Megiddo collection into the global schema, we first consider mapping of "flint tool" and then use the knowledge of these mappings to map "vessel".

Figures 4 and 5 show screenshots before and after the mapping of flint tool to the global schema. The left hand side screen shows the Megiddo local schema, while the right hand side shows the ETANA global schema. Initially, the ETANA global schema shows just those nodes which would be present in all artifacts – namely OWNERID, COLLECTION, OBJECTTYPE, PARTITION, SUBPARTITION, LOCUS, and CONTAINER, along with the root node OBJECT (see Fig. 4).

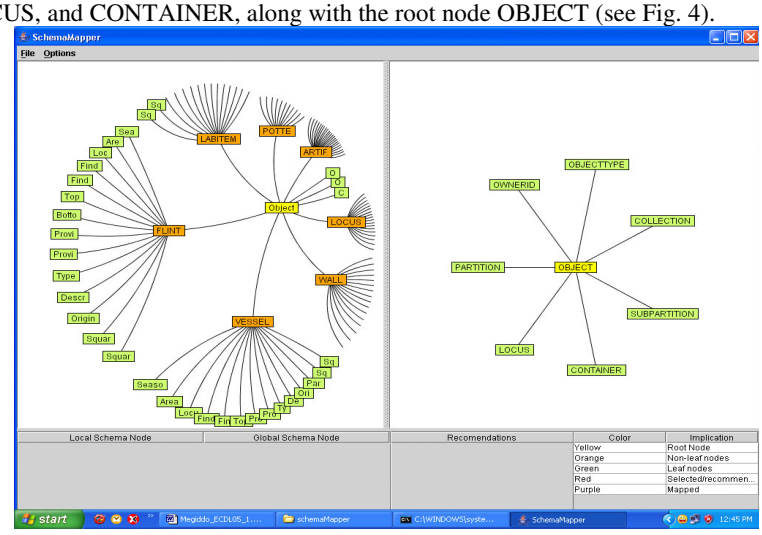

Fig. 4 Before mapping of flint tool in Megiddo to ETANA global schema

Based on rules and name based matching strategies, we recommend mappings: OWNERID->OWNERID, OBJECTTYPE->OBJECTTYPE, COLLECTION->COLLECTION, AREA->PARTITION, Square1->SUBPARTITION, Locus->LOCUS, and OriginalBucket->CONTAINER. The above mapping format has the local schema node on the left hand side of the arrow and the recommended global schema node on the right hand side. We map the nodes according to the recommendations, indicated by coloring these nodes purple (see Fig. 5).

As the remaining nodes in the local schema do not have corresponding global schema nodes, we add the flint tool sub-tree as a child of the OBJECT node in the global schema. This ensures that local schema elements and properties are preserved during the mapping transformation. Schema Mapper determines that some of the nodes (Area, Locus, OriginalBucket, and Square1) are already mapped, deletes these nodes from the global schema sub-tree, and automatically maps the rest with the cor-

responding elements in the local sub-tree (see Fig. 5). The user may decide to rename some nodes in the global schema from within this sub-tree to avoid any local connections with the name. Assume the user renames global schema node Description to DESCRIPTION. With this the mapping process is complete (see Fig. 5). Once the user decides to confirm the mappings, a style sheet is generated, the mappings are stored in the database, and the ETANA global schema is updated with the flint tool schema.

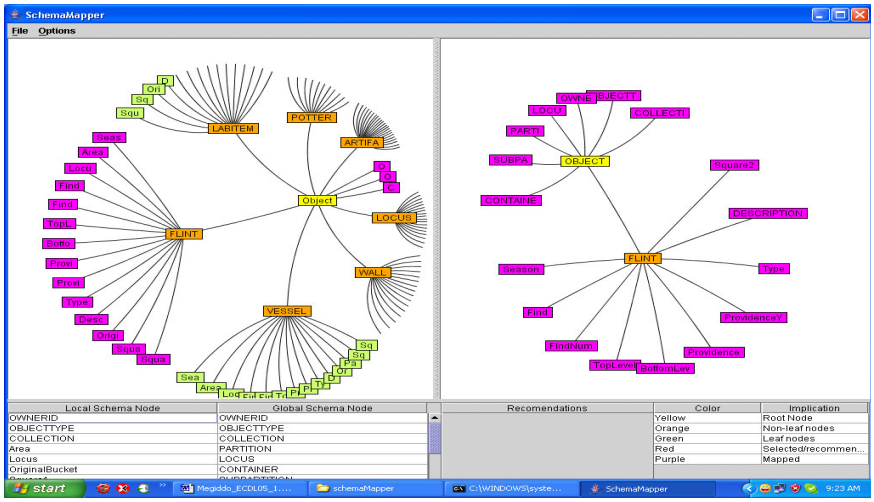

Fig. 5 After mapping of flint tool in Megiddo to ETANA global schema

We next integrate the VESSEL artifact of Megiddo into the ETANA global schema. When we open the global schema for mapping, even though the global schema is now updated with the flint tool schema, we display only the root node and the seven nodes described earlier that are common to all artifacts. This is to avoid erroneous cross mappings from schema nodes of one of the artifacts to similar schema nodes present in other artifacts. This also prevents the user from accidentally modifying a node, from say the flint tool sub-tree in the global schema, and rendering the previously generated XML files inconsistent. Also, this avoids confusing the user by presenting him with only the information he needs to see for mapping. Once again recommendations are made to enable the initial set of seven mappings; after this, the user adds the VESSEL sub-tree to the global schema.

As before, Schema Mapper finds that the Area, Locus, Square1, and Original-Bucket nodes are already mapped – and deletes them in the global sub-tree and maps the remaining nodes to corresponding local schema nodes automatically. Schema Mapper also goes through the mappings history and finds that the Description node in the flint tool sub-tree was mapped to the DESCRIPTION node in the global schema. In order to keep naming consistent, Schema Mapper recommends the user to change the name of the Description node in the VESSEL sub-tree to DESCRIPTION. This is due to the fact that both the DESCRIPTION node in the flint tool sub-branch of global schema and the Description node in the VESSEL sub-branch of the global schema

describe the respective artifact, but as DESCRIPTION has been selected as the global name, all Description elements in the global sub-tree should be renamed as DESCRIPTION. The recommendation as always is not mandatory, but if followed will help keep names consistent. When the user confirms the mappings, the database is updated, the style sheet generated, and the global schema updated with the VESSEL schema. It is important to note that the integration of vessel artifacts into the global schema in no way changed the existing flint global entry. This leads us to the observation that modification of the global schema is simply appending a new local artifact into the global schema without changing the existing global artifacts.

The style sheets generated are applied on local XML files corresponding to particular artifacts, like vessel or flint tool; corresponding global XML files are generated. These are ready for harvest into the union DL, and available for access by services like Searching and Browsing. Detailed screenshots of the mapping process are in [12]. We integrate other artifacts in the Megiddo schema into the union DL similarly.

3.3 Comparison of Schema Mapper with MapForce for integrating flint tool

We did a pilot study by comparing Schema Mapper with MapForce [9], regarding the amount of time required to map the flint tool collection to the ETANA global schema. The pilot tester was an expert in the domain of XML schema mapping and had used MapForce and XML Spy [13] earlier for mapping and editing schemas respectively. Before performing the actual Benchmark task, the pilot tester was given sample tasks to familiarize himself with both Schema Mapper and MapForce.

The actual benchmark task required the user to map the flint tool collection to the ETANA global schema. There were explicit guidelines on how to go about achieving this for both Schema Mapper and MapForce. The metrics for measurement were: time taken to achieve the task, number of errors, and number of times the user scrolled in MapForce vs. number of reorient actions (moving the hyperbolic tree) in Schema Mapper. The pilot tester used Schema Mapper first, and then MapForce, for performing the benchmark task. The results are shown in Table 1:

Table 1. Comparison of Schema Mapper and MapForce [9]

	Schema Mapper	MapForce
Time taken (in minutes)	4:10	9:00
Number of errors in mapping	0	0
Scrolling vs. Re-orient actions	9 (re-orient)	13 (scrolls)

From Table 1 we see that Schema Mapper significantly outperformed MapForce [9] in the amount of time that it took for the user to perform the task. Schema Mapper also required fewer re-orient actions than scroll actions. The user made 0 errors.

Another observation was that MapForce did not help the user to edit the global schema. As a result, whenever the user had to update the global schema he had to open it in XML Spy and edit the schema by hand. For a simple collection like flint tool, the user had to switch between XML Spy and MapForce 4 times, while no switching was required for Schema Mapper as it supports editing.

Thus the pilot study strongly indicates that for simple one-to-one mappings as are found in the Megiddo collection, Schema Mapper significantly outperforms MapForce. Further usability tests will be conducted to expand the comparison study.

4 Integrated Service in Union DLs: Multi-dimensional Browsing

In this section, we consider integrated services for union DLs, specifically the multi-dimensional browsing service. Later, we describe a scenario for extending this browsing service to incorporate the Megiddo collection. From the 5S point of view, it should be noted that this scenario is for the *society* of DL administrators.

4.1 Extending Integrated DL Services: Overview

A digital library is not just about the data but also about the services it provides. ETANA-DL supports various integrated DL services such as Annotation, Browsing, DL object comparison, marking DL objects, multi-dimensional browsing, Recommendation, and Searching. The expressiveness of Archaeology-specific services is defined relative to the collections currently present in ETANA-DL. As new collections are integrated into the union catalog, the ETANA-DL global schema [14] is extended, and maintains the up-to-date state of the DL. In the current scenario, the ETANA-DL global schema is extended to include flint and vessel DL objects for integrating the Megiddo collection into the union catalog.

An integrated DL service can be provided for newly integrated collections in two ways. One is to re-implement the service built upon the union catalog based on the global schema. The other method is to extend the existing service to incorporate new collections. We adopt the second approach since it is more efficient.

The basic idea is to re-engineer domain specific services such that they are updated based on the global schema. So, whenever the global schema is modified, the update routine associated with each of these services automatically updates its internal data structures, if necessary. Then, newly added data collections are harvested from the union catalog into its index, and the service is made available. This approach leads to domain independent, flexible, and reusable components in union digital libraries.

We demonstrate the feasibility of the above idea by developing a prototype of a multi-dimensional browsing component. In contrast to the ODL [15] browsing component, Greenstone [16]'s classification based browsing service, and Sumner et al. [17]'s browsing interface built using dynamically generated components, our focus is to modularize the browsing service and partially automate its development.

4.2 Integrated Multi-dimensional Browsing Service

Digital objects in ETANA-DL are various archaeological data, e.g., figurine images, bone records, locus sheets, and site plans. They are organized by different hierarchical structures (e.g., animal bone records are organized based on: sites where they were

excavated, temporal sequences and animal names). By navigational dimension, we mean a hierarchical structure used to browse digital objects. This hierarchical structure contains one or more hierarchically arranged categories that are determined by the elements of the global schema. In addition to this, a dimension of ETANA-DL can be refined based on taxonomies existing in botany and zoology, or from classification and description of artifacts by archaeologists. Our multi-dimensional browsing component allows the user to browse through multiple dimensions simultaneously. In ETANA-DL, we can browse for DL objects through the OBJECT, SPACE, and TEMPORAL SEQUENCE dimensions.

The three main sub-components of the multi-dimensional browsing component are the *browsing database maintenance module* (for creation and update of the database), the *browsing engine,* and the *browsing interface module*. The last two support browsing interaction. The browsing component maintains a browsing database, i.e., a quick index to browse for DL objects quickly and efficiently.

All three modules work based on input provided by the browsing metamodel – an XML document that encodes the details of all navigational dimensions. This metamodel performs a mapping from elements of the global DL schema to the hierarchical levels in each dimension using XPath expressions. This metamodel, derived from the global schema, is generated by 5SGraph [18]. The browsing metamodel used for ETANA-DL is at http://feathers.dlib.vt.edu/~etana/browse/etanabrowse.xml. The *browsing database maintenance module* created the browsing database from the previously described metamodel.

4.3 Scenario: Extending the Browsing Service to incorporate Megiddo Collection.

We integrated the Megiddo collection into the ETANA-DL union catalog with the help of the XSLT style sheets generated by the Schema Mapper. The union catalog contains flint and vessel DL objects from the Megiddo site, along with other DL objects from various sites. Hence, we consider how to extend the browsing component to provide browsing services for the newly integrated Megiddo collection.

The first step is to determine the elements of the flint and vessel DL objects of the Megiddo collection that map to the hierarchical levels in each browsing dimension. The flint and vessel DL objects support browsing through dimensions SPACE and OBJECT. The elements COLLECTION, PARTITION, SUBPARTITION, LOCUS, and CONTAINER of these two DL objects define hierarchical levels in the SPACE dimension. The element OBJECTTYPE defines the OBJECT dimension. Since these elements are shared by all earlier DL objects present in the union catalog and are already used in generating space and object dimensions, the browsing metamodel need not be modified. The mapping component maps similar elements of local schemas to the same element in the global schema, eliminating the need to modify the browsing metamodel.

The next step is to run the *browsing database maintenance* module to harvest the Megiddo collection from the union catalog into the browsing database. We demonstrate the extended browsing service using screen shots. Once the Megiddo collection

is harvested into the browsing database, the browsing component automatically enables browsing through the Megiddo collection using SPACE and OBJECT dimensions. From Fig. 6, it can be observed that the SPACE dimension of the browsing component contains the Megiddo site (indicated by a red rectangle around it) along with all other sites. Also, the OBJECT dimension has two new objects FLINT and VESSEL (indicated by red rectangles around them) for browsing. Fig. 7 is a screenshot of records that are displayed when we select Megiddo as the site and Flint as the ObjectType (see Fig. 6.) and choose to display results. Thus, we have demonstrated extending the browsing component to incorporate the new DL collection from the Megiddo site. The browsing service that incorporated the Megiddo collection is accessible at http://feathers.dlib.vt.edu:8080/ETANA/servlet/BrowseInterface.

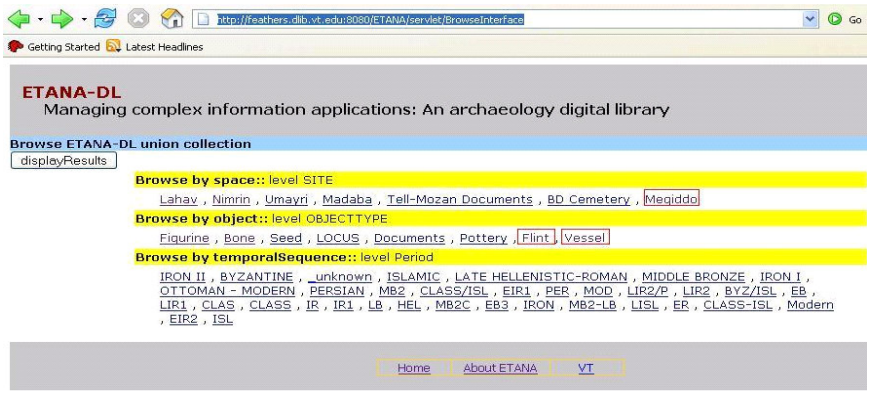

Fig. 6. Browsing component after incorporating the Megiddo Collection.

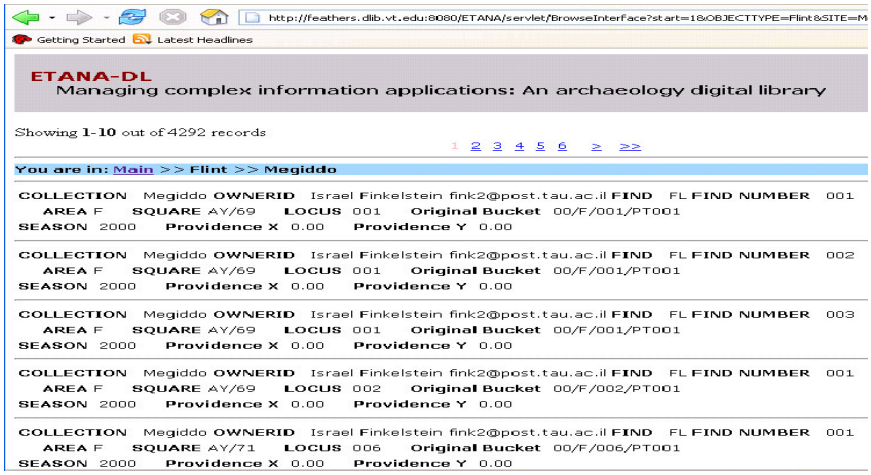

Fig. 7. Records from the flint tool collection of ETANA-DL

5 Savings from Automation

Automation is achieved in the wrapper code generation (style sheet generation), global schema enrichment, and through the extensible multi-dimensional browsing component. We measure this automation in terms of lines of code (LOC) that we saved using Schema Mapper and the Multi-dimensional Browsing Component as compared with the former hard coded approach. Table 2 shows the LOC required in each case to add a collection into the Union-DL. With Schema Mapper, there is no need to write new code when we integrate a new collection, such as Nimrin or Madaba.

Table 2: LOC Comparison between Schema Mapper and Hard Coded Approach

	Hard-Coded Wrapper for Nimrin	Hard-Coded Wrapper for Madaba	Schema Mapper (for Madaba and Nimrin)
Additional LOC required	1605	1770	0

Regarding the Browsing Service, comparison is more difficult, since we earlier hard coded a (less flexible) browsing component. Clearly, we avoid between 50-75 lines of additional code that would be required for a change the old way. But the main advantage of the new component is that it can be plugged into any digital library (not necessarily an Archaeological DL) and be driven by a formal (5S) browsing model.

6 Conclusions and Future work

Through the integration of artifact data from the Megiddo excavation site into the union DL, using the visual mapping component, we have successfully demonstrated a semi-automatic tool which generates a wrapper (XSLT style sheet) using the schema mapping approach. Through the mapping component we also achieve the goal of incrementally enriching the global schema with local schema information from new excavations. Further, through the multi-dimensional browsing component we complete the automation of the workflow for adding a site, from integrating data into the union DL, to extending the browsing service to access all integrated data.

Initial pilot studies of the mapping and browsing components have been positive. We plan extensive usability tests. Also, complex (one to many and many to one) mappings will be explored and the mapping component will be enhanced accordingly. Future work will include enriching mapping recommendations through statistical data analysis, and enhancing the functionality and portability of the browsing component.

Acknowledgements: This work is funded in part by the National Science Foundation (ITR-0325579). We also thank Doug Gorton and members of the DLRL for their support. Marcos Goncalves had an AOL fellowship and has support from CNPq.

References

1. Park, J. and Ram, S. Information systems interoperability: What lies beneath? ACM Transactions on Information Systems (TOIS), 22 (4): 595 – 632, 2004.
2. Finkelstein, S., Ussishkin, D. and Halpern, B. Monograph Series of the Institute of Archaeology, Tel Aviv University, 2000.
3. Megiddo, 2005 http://www.tau.ac.il/humanities/archaeology/megiddo/index.html
4. U. Ravindranathan. Prototyping Digital Libraries Handling Heterogeneous Data Sources - An ETANA-DL Case Study. Masters Thesis. Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, April, 2004, http://scholar.lib.vt.edu/theses/available/etd-04262004-153555/
5. Shen, R. Apply the 5S Framework in Integrating Digital Libraries. Dissertation Proposal, Virginia Tech, 2004
6. Gonçalves, M.A., Fox, E.A., Watson, L.T. and Kipp, N.A. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM Transactions on Information Systems (TOIS), 22 (2): 270 312, 2004.
7. Schloen, J.D. Archaeological Data Models and Web Publication Using XML. Computers and the Humanities, 35 (2). 123-152, 2001
8. L. L. Yan, R. J. Miller, L. M. Haas, and R. Fagin. Data-driven understanding and refinement of schema mappings. In SIGMOD Conference, 2001
9. Altova. *Mapforce,* 2005. http://www.altova.com/products_mapforce.html
10. Microsoft. *BizTalk Mapper*, 2005 http://www.samspublishing.com/articles/article.asp?p=26551&seqNum=5
11. Lamping, J. and Rao, R., Laying Out and Visualizing Large Trees Using a Hyperbolic Space. In Proc. ACM Symp. on User Interface Software and Technology, (1994), 13-14
12. Schema Mapper Screenshots, 2005 http://feathers.dlib.vt.edu/~etana/Papers/Screenshot.doc
13. Altova, XML Spy, 2005. http://www.altova.com/products_ide.html
14. ETANA-DL Global XSD, 2005 http://feathers.dlib.vt.edu/~etana/etana1.1.xsd
15. Suleman, H. Open Digital Libraries, Ph.D. Dissertation, Dept. Comp. Sci., Virginia Tech, 2002, http://scholar.lib.vt.edu/theses/available/etd-11222002-155624
16. Bowman, C.M., Danzig, P.B., Hardy, D.R., Manber, U. and Schwartz, M.F. The Harvest information discovery and access system. Computer Networks and ISDN Systems, 28 (1-2): 119 - 125.
17. Sumner, T., Bhushan, S., Ahmad, F. and Gu, Q. Designing a language for creating conceptual browsing interfaces for digital libraries. In Proc. JCDL, 2003. 258--260.
18. Zhu, Q. 5SGraph: A Modeling Tool for Digital Libraries, Masters Thesis. Dept. Comp. Sci., Virginia Tech, 2002, http://scholar.lib.vt.edu/theses/available/etd-11272002-210531