This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

EIAH Data Model: Semantic Interoperability between Distributed Digital Repositories

Emad Khazraee, Azade Sanjari, Shadi Shakeri, Saeed Moaddeli
Affiliation: Encyclopedia of Iranian Architectural History

Abstract:

The encyclopedia of Iranian architectural history was established with the goal of increasing the accessibility of the widespread resources and documents related to Iranian architectural history and to provide a better and more productive space for collaboration of researchers and scholars, enabling them to expand and improve this encyclopedia. The information architecture which started to get implemented is aimed to achieve three goals. First, increase the accessibility of the documents related to topics; second, the relation between concepts; third, the relation between concepts and documents. A three-layer architecture is designed to achieve the mentioned goals (EIAH cake). The underlying layer is a pool of information which is an integration of distributed digital repositories in our case. The top level is the knowledge representation level, an ontology of Iranian architectural history and the last layer which sits in the heart of this architecture is the mediator level which is responsible for establishing the relation between concepts and documents and enhancing search and semantic interoperability. The metadata model for describing resources in distributed digital repositories is customized based on Dublin Core with refinements. All documents in distributed repositories get their metadata according to this model and a detector agent (the mediator level) harvest metadata to interpret them by the ontology (the top layer). The results of this process will be presented in a semantic portal or might be used for complex search queries by end users. When this happens on a federation of distributed digital repositories, the ocean of separated documents becomes much meaningful and interpretable by human scholars.

Keywords: Digital Repository, Metadata, Topic Map, Semantic Interoperability, Federated Search, DC Application Profile
Classification: Case Study

1. Introduction to EIAH

The Iranian World contains a collection of some oldest human settlements dating back to ten millennia ago. Few of these properties have remained today. This remaining heritage amounts to approximately one million historic buildings and sites. The properties within this region are not only numerous, but also extremely diversified. This is due to the varied natural environment of the region. The history of Iranian architecture has mostly been limited to the western fringes of the desert. Thus studies conducted reflect neither the environmental nor ethnic diversity of the Iranian world. The reason might be the vast diversity of these sites, as well as limited access to the scattered resources and documents in this field. Due to the lack of any integrated network to collaborate, plenty of redundant and parallel works and studies have been carried out. The efforts to integrate different data centers was unsuccessful yet, due to shortage of infrastructure for knowledge exchange and interoperability, the problem which led to a recess in Iranian architectural studies.

The Encyclopedia of Iranian Architectural History (EIAH) established and aimed to ease the road for scholars and remove the aforementioned obstacles. The main objective of this encyclopedia is to increase the quantity and improve the quality of information on Iranian

culture in order to facilitate the recovery of vernacular identity, as well as presenting the joint heritage of the countries in this region for further interaction and focusing on cultural unity. Representation of resources in history of Iranian architecture, identifying, collecting and preserving them for long term use are tasks which EIAH tends to accomplish during the project's lifetime. According to the above goals, the purpose of this project is to establish a center for managing information on history of architecture and urban development[i].

2. EIAH information architecture

Two core concepts in EIAH information architecture are Entry and Document. An entry is a topic or concept which information accumulates around it. The topic or concept can be a person, or a term which is used in the field of architectural history. Seven categories have been defined for entries which covers all concepts of Iranian architectural history: Terms, Monuments, Geographical name, Person, Historical period, primary resource and movable object. A document is any kind of resource, published or unpublished, which provides information regarding history of Iranian architecture. A document can be in various formats, from text, photo and drawing to audio or multimedia (Khazraee et al., 2008).

The information architecture must meet three main criteria. First it must facilitate the access to the resources and documents. The recall and precision factors are evaluative factors in information retrieval (Van Rijsbergen, 1979). The second factor is the ability to represent the conceptual relations between topics of Iranian architecture domain; (here the topics could be mentioned as entries in this paper. The last requirement of information architecture for EIAH is the ability to represent and develop relations between topics and resources associated with them.

A semantic inter-related network of entries enables us to have a big picture of Iranian architectural history and therefore deduction of new ideas would be feasible. One can find more complex relations which are not simply recognized in the first look. Additionally, this model can provide a holistic view to the domain from different perspectives. This made it possible to examine the accuracy of hypotheses regarding architectural history of Iran, and their alignment to the evidences extracted from historical resources.

Due to the aforementioned points, a three-layer architecture is designed to achieve objective of EIAH. This three-layer architecture is a service-oriented architecture. It is consisted of distinct components and services which work both on their own and in the grid. Services in this architecture communicate and interchange data with each other. This architecture is also semantic enabled. The modules and services in this architecture pass data and objects to each other while they are aware of the semantic identity of that data or object (Newcomer and Lomow, 2005). The modularity of the architecture not only makes it easier and faster to develop and extend but also enhances its performance, since there are different customizable components already available. This architecture is consisting of a foundation layer and three core layers as follow:

2.1 Standards and policies (foundation layer)

All processes and work-flows in this project must follow open and international standards and guidelines, so all the products in different phases of the project could be homogenized and optimized. These guidelines are known as EIAH's standards and policies. As the days of one standard fits all, is over due to the needs of different cultural centers in Iran adopting an already available international standard will not be sufficient. Therefore, EIAH developed standards based on mostly used international standards taking into account the special attention to the local and organizational requirements. The base standards range from different ISO/IEC standards to W3C and semantic web standards.

The compliance of EIAH standards with open and international standards enables the project to be smoothly interoperable with other projects and services in this domain and eases the exchange of data using standard and widely-used protocols. EIAH has adopted seven standards which are as follows:

- Software Standard Policies
- Hardware and Network Standard Policies
- Technical Tracking Standard Policies
- Information Storage and Exchange Standard Policies
- Content Legal and licensing Standard Policies
- Security Standard Policies
- Resource Description and metadata Standard Policies

2.2 Information pool

In the three core layer architecture (which is also known as EIAH cake) the underlying level is the information pool. A network of digital repositories, containing various types of resources related to Iranian architecture, sits at the bottom of this architecture. To reach homogenized and interoperable data, all information in the different digital repositories should be aligned to the mentioned standards. On the other hand, to establish the grid of digital repositories a powerful Open source solution which can easily shared with different institutions was necessary. As a result, Dspace institutional repository platform was chosen for this part after evaluating the platform and reviewing twenty other solutions. Dspace is an Open source software and developed in a joint effort by MIT and HP company[ii]. Dspace is now the most popular institutional repository in the world (Markey et al., 2007). Dspace supports different types of metadata, can exchange metadata and accepts crosswalk plug-ins. Dspace uses strong Apache Lucene search engine[iii]. EIAH customized and localized Dspace for the institution's needs. These modifications include right to left text rendering, full support of Persian and Arabic scripts, Persian Calendar, fully translated user interface.

Different types of documents in digital repository are all annotated with standard metadata values. These documents are described with relational metadata elements which are according to the EIAH metadata and tagging standard and EIAH controlled vocabulary (which is a developing work in progress). EIAH policy is to support different distributed digital repositories due to the numerous cultural heritage centers in Iran. Therefore, the information pool is consisted of multiple digital repositories in a grid. Each contains their own resources but share a metadata schema. This enables high semantic interoperability among different services using this information architecture. This grid of digital repositories is not the only source of information and documents. Since the architecture is service oriented, every other available service can communicate with other parts and join the grid just by talking with the same language and following the laid down guidelines. One of such services is Aratta, a collaborative research tool, which is a semantic note taking tool for scholars of Iranian architecture. Aratta is being developed as a web-based research tool which allows note taking along with establishment of semantic relations between notes as well as providing reference management services (Khazraee et al., 2008). Aratta deploys the conceptual model of the EIAH and defines its relational tags based on this model. Meanwhile the accumulation of notes entered in Aratta by scholars could be used to improve and enrich EIAH controlled vocabulary.

2.3 Ontology – knowledge representation level

An ontology is a specification of a conceptualization and a formal representation of a set of concepts within a domain and the relationships between those concepts (Gruber, 1992). Ontologies in different domains are the product of collaboration between ontology engineers and domain experts. Having such a model for Iranian architectural history, one

can deduct hypotheses and figure out relations which was never obvious without an ontology. In Iranian architecture domain, the ontology gives us an overall picture of Iranian architectural history with all its concepts and all their relations. For example, A geographical name object is an object in the ontology and *the city of Kangâvâr* is an instance of it as an entry. There is another object like monument which has an instance like Â*nâhitâ temple* which is located near the city of *Kangâvâr*. This monument belongs to historical periods of Parthians and Sassanids. Such relations between entry types are described in the ontology and based on these representation and relations, the information related to a topic or entry can be explored and discovered. On the other hand, the ontology can be deployed to improve the search system as well as interpretation.

2.4 The mediator level

The mediator level contains a set of tools which interrelate entries with their occurrences in digital repository. In its advanced performance, in response to a received query, the mediator level looks into the ontology and by tracing relations between topics, suggests other items that might relate to the query's answer as results. For example if a query is looking for documents or information about *Shâh Abbâs,* the mediator level suggests items like the monuments which have been built in the time of his reign, or by his family members, the works which belong to his ruling period and etc.

The mediator level can even give more interesting results by using some rules like weighting the results and identifying the nearest and best results based on users' rating. The gadgets here can index different web resource and use the ontology to inter connect different data pieces not only the information available in the EIAH information pool. This information architecture could simply compare with Topic Map model, an ontology of topics (entries here) which connected to their occurrences (documents here). However, in EIAH model there are differences that it could be called a dynamic topic map model, since it does not use static occurrence connections, while there are agents which discover and interconnect the occurrences to the topics automatically. On the other hand, in this architecture as we use an independent ontology it does not limited to the restrictions of topic map expressiveness and it can be a more expressive ontology.

3. Metadata Model

Since metadata plays a key role in establishment of the mentioned information architecture, the EIAH metadata model should comply with the requirements of this architecture. EIAH has customized its metadata model based on the Dublin Core (simple & qualified) in order to describe and facilitate the discovery of the electronic resources in the area of Iranian architectural history. According to the characteristics of Dublin Core Standard, namely, simplicity, Semantic interoperability, Extensibility and compliance with the Resource Description Framework (RDF) Dublin Core metadata standard is used by EIAH (CEN, 2003). To completely comply with the EIAH needs, an application profile developed for the special domain-specific focus of EIAH. This application profile designed to be aligned with the need to interrelate resources to the topics.

EIAH relational elements are adopted as refinements of Dublin Core Element, Subject, because they narrow down the subject of the resources. These refinements design to comply with the EIAH ontology. EIAH ontology includes three major classes of entities: temporal, spatial, and human (actors and actions). These are superclasses which are consisted of several subclasses themselves. Each class has abstracts and instances. Since the major categorization may seem too vague for the end user, these relational tags are classified in six groups which are more tangible: Persons (as subclass of human entities), Works (monuments and sites) and Geographical Names (as subclass of spatial entities), Historical Periods and Events (as subclass of temporal entities), as well as (Architectural) Terms (which is the abstract level of all classes). Since these selected

relational qualifiers of EIAH descriptive Element Set are based on EIAH conceptual model, they would help to develop more specific semantic relations between resources and EIAH ontology (Khazraee et al., 2008).

EIAH Controlled Vocabulary has been developed, with the aim of developing a terminology in the domain of Iranian architectural history and it has classified according to the EIAH six major concept (entries), works, persons, historical periods, geographical names, and events, and implemented in the EIAH Archival Repository (Dspace) to enhance the quality of the metadata in the process of resource description. The usage of EIAH Vocabularies has resulted in improvement of searching precision and recall of EIAH resources and also has enabled automated interoperability.

The development process of EIAH Controlled Vocabulary includes the following steps: Defining different classifications for vocabularies according to the six major topics in the field of Iranian Architectural history (works; person, geographical names, period, events and terms), the vocabularies in these classification are used as the value of subject qualifiers developed by EIAH; Identifying the resources and specifying their preferences; extracting the vocabularies (terms) from the preferred resources; compiling and editing all the extracted vocabularies; and recording the finalized entries. At the moment the process has led to the accumulation of 5900 geographical names, 2000 work names, 750 person names, and 6000 architectural terms. However, the research and development in this area has not been yet stopped and the number of vocabularies in each classification is being increased.

3.1 EIAH Application Profile

For efficiency, specificity, and localization within the context of our community, EIAH has designed its particular application profile based on the Dublin Core Singapore Framework. In this framework the application profile is consist of Functional Requirement, Data Model, Description Set Profile, Syntax Guidelines and data formats, and Usage Guidelines (Nilson et al., 2008). In addition to the application of Dublin Core terms, other domain-specific qualifiers have been developed by EIAH for its special needs. Therefore, to make it possible for EIAH to represent all its restricted elements and qualifiers, schemes, values and guidelines, and introduction of the new qualifiers, and values, EIAH defined its new namespace (http://www.eiah.org). Moreover, EIAH Description Set Profile designed based on the Dublin Core Description Set Profile [DC-DSP]. All the terms used in the EIAH Application Profile has been specified by Description Templates which contain all the constraints on the described resources (entities of EIAH Data Model) and by Statement Template which contain all the constraints that apply on the metadata of EIAH domain (Table 1).

4. Semantic portal

The outcome of EIAH's information architecture will be presented to the end user through a web portal. Web information portals are very successful but today there are several ways in which their current performance can be improved. In particular, the emergence of the semantic web[iv] brings a new set of tools and techniques that could radically change the way portals are built, integrated and used. In this portal two main services would be available, Entries and Documents. Faceted browse on the entries based on their properties as well as visual browse of semantic network of entries. On the other hand, integrated access to the distributed repositories is possible. Additionally, all the retrieved data can be visualized on geographical maps or timeline whenever they have geo-spatial or temporal properties.

Term Name: Is related to Event		Term Name: Is related to Term	
URI	Http://eiah.org/en/Entries#Event	URI	Http://eiah.org/en/Entries#Term
Label	Is related to Event	Label:	Is related to Event
Definition	An entity responsible for correlating the an event to the resource	Definition:	An entity responsible for correlating a term to the resource
Type of term	Property	Type of term:	Property
Refines:	http://purl.org/dc/elements/1.1/subject	Refines:	http://purl.org/dc/elements/1.1/subject

Term Name: Is related to Geographical Name		Term Name: Is related to Historical Period	
URI	Http://eiah.org/en/Entries#Geographical_Name	URI	Http://eiah.org/en/Entries#Historical_Period
Label:	Is related to Geographical Name	Label:	Is related to Historical
Definition:	An entity responsible for correlating a geographical name to the resource	Definition:	An entity responsible for correlating a historical period to the resource
Types of term:	Property	Types of term:	Property
Refines:	http://purl.org/dc/elements/1.1/subject	Refines:	http://purl.org/dc/elements/1.1/subject

Term Name: Is related to Person		Term Name: Is related to Work	
URI	Http://eiah.org/en/Entries#Person	URI	Http://eiah.org/en/Entries#Work
Label:	Is related to Person	Label:	Is related to Work
Definition:	An entity responsible for correlating a person to the resource	Definition:	An entity responsible for correlating a work to the resource
Types of term:	Property	Types of term:	Property
Refines:	http://purl.org/dc/elements/1.1/subject	Refines:	http://purl.org/dc/elements/1.1/subject

Table 1: EIAH 'DC' Subject Refinements

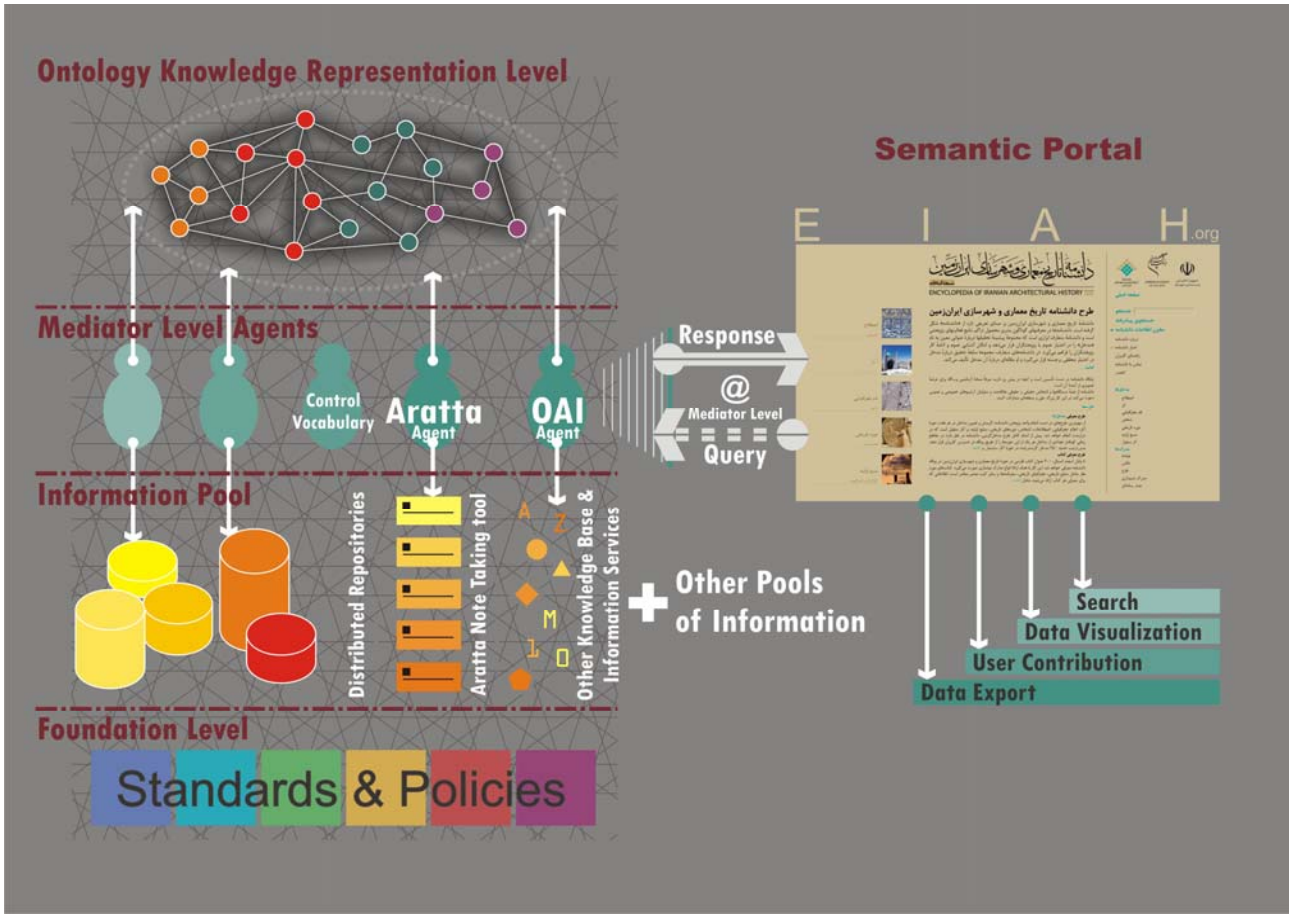

Figure 1EIAH Information Architecture

EIAH used this approach and therefore chose to adopt Semantic Mediawiki[v] as the content management system for current phase of development. Although the current outcome based on Mediawiki somehow differs from the final portal, but it resembles most of the functionality which is expected from a semantic portal in the near future (Figure 1).

5. Distributed repositories

EIAH federates distributed repositories in widespread cultural heritage centers. Documents in centers are stored in digital repositories and are described using standard metadata guidelines based on Dublin Core Metadata. The single application profile used to describe documents in distributed repositories provides uniform metadata which is harvestable and therefore processable by web services in this grid. A connection from the semantic portal based on Semantic Mediawiki to Digital repositories is already established using the OAI protocol for metadata harvesting (OAI-PMH)[vi]. OAI-PMH is a lightweight protocol for gathering metadata from a number of distributed repositories into integrated data storage for sharing them between services. It gives data providers an option to make their metadata available to service providers. The protocol is based on open standards HTTP and XML. Dspace has a built-in OAI interface and allows harvesting metadata records of items in the repository. Through Dspace OAI interface, one can send OAI queries in URL, using HTTP, and receive the response in XML format. This XML responses which includes metadata records of all items available in the digital repository, will be processed and metadata fields and their associated values will be stored in a database (aggregation). On the portal side, an extension is prepared for Semantic Mediawiki which searches the database of harvested metadata records for related values to the entries on the portal. The documents related to a specific entry could be found by different fields of their metadata. This extension periodically scans distributed repositories

for their new items and adds the new harvested metadata records to the database.

6. The Current Implementation

The current implementation of EIAH cake is a simplified version of the architecture reviewed in this paper. The repository level is deployed using Dspace. These repositories contain resources described with EIAH metadata application profile based on Dublin Core application profile guidelines. The knowledge representation level (the ontology level) is now implemented in a very simple form using Semantic Mediawiki tools. By using categories and properties in Semantic Mediawiki, it is possible to simulate the functionality of an ontology. The mediator level is implemented as some independent extensions to the Semantic Mediawiki portal. They look into the digital repositories and Aratta and links the related resources to the specific entries. SIMILE data visualization tool, Exhibit, has been used for semantic data visualization right now[vii].

7. Future Works

There are a few steps remained to reach the first vision of EIAH project:

- Deploying more digital repositories in other cultural heritage centers;
- Development of EIAH ontology;
- Development of EIAH controlled vocabulary;
- Implementing of Dspace XML UI framework (Manakin) to increase adaptability;
- Enhancement of EIAH application profile based on DCAP Singapore framework;
- Development of more data visualization tools;

8. References

Van Rijsbergen, C.J. (1979), *Information retrieval (2nd. edition)*, Butterworths, London, Boston.

Khazraee, E., Malek, H., Forghani, H., (2008)"Introduction Of Aratta As A Collaborative Research Tool For Iranian Architectural History " in *Digital Heritage: Proceedings of the 14th International Conference on Virtual Systems and Multimedia in Limassol, Cyprus, 2008*, pp. 281-286

Newcomer, Eric, Lomow, Greg (2005), *Understanding SOA with Web Services*, Addison Wesley.

Markey, K. et al. (2007), "Census of Institutional Repositories in the United States, MIRACLE Project Research Findings", available at: http://www.clir.org/pubs/abstract/pub140abst.html (accessed 15 April 2009)

Gruber, T., (1992) "What is an ontology?" available at: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html (accessed 15 April 2009)

Nilson M., Baker, T., Johnston, P. (2008), "The Singapore Framework for Dublin Core Application Profiles", available at: http://dublincore.org/documents/singapore-framework (accessed on 25 April 2009)

CEN, (2003), "Guidance information for the use of Dublin Core in Europe, CWA 13988" available at: http://www.cen.eu/cenorm/businessdomains/businessdomains/isss/activity/wsmmi.asp (accessed on 15 April 2009)

Notes

i "Digital Encyclopedia of Architectural History of the Iranian World, Objectives and methods" http://eiah.org/en/About_Us/Vision (accessed 15 April 2009)

ii "Dspace Homepage" http://www.dspace.org (accessed 25 April 2009)

iii "Lucene Homepage" http://lucene.apache.org (accessed 25 April 2009)

iv "W3C Semantic Web Activity" http://www.w3.org/2001/sw/ (accessed 25 April 2009)

v "Semantic Mediawiki Hompage" http://www.semantic-mediawiki.org (accessed 25 April 2009)

vi "The Open Archives Initiative Protocol for Metadata Harvesting" http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed 25 April 2009)

vii "SIMILE project" http://simile.mit.edu (accessed 25 April 2009)