IFLA WLIC 2013
SINGAPORE

Submitted on: June 30, 2013

Publishing Persian Linked Data of National library and Archive of IRAN

Saeedeh Eslami
Information Technology & Communication Department, National Library and Archive of IRAN, Tehran, IRAN.
E-mail address: s-eslami@nlai.ir

Mohammad Hossein Vaghefzadeh
Deputy of Research, Planning and Information Technology, National Library and Archive of IRAN, Tehran, IRAN.
E-mail address: m-vaghefzadeh@nlai.ir

Abstract:

Nowadays institutions try to publish, share and interlink their data by using semantic web technologies specially Linked Data. This technology has considerable potential for libraries and cultural institutions and helps them to complement their data by linking it to other external data sources. Library linked data derived from bibliographic records which based on international standards will be of high. Thus, National Library and Archive of IRAN(NLAI) is planning to transform its data to RDF following the Linked Data principles proposed by Tim Berners Lee. This paper is the first experiment in the realm of publishing linked data at NLAI. Using IFLA FR models, it has been stating to define a model to generate linked data version of authority files of NLAI which is based on IRANMARC. It discusses the challenges we met during the experience particularly Persian language problems. We outline how this process can be facilitated.

Keywords: Library Linked Data, libraries, Authority files, Ontologies, RDF, URI.

1 BACKGROUND

Libraries are producing larger and more complex data than ever before. It is imperative that these data outputs are effectively managed and shared. Better data – better described, more connected, more integrated and organised, more accessible, more easily. Unfortunately, library data is not yet an integral part of the web. Jan Hannemann (2010) noticed that this is mostly due to the poor level of linking between library datasets and data from other domains, but also due to the current data collection processes and data formats, which – naturally – focus on classical usage scenarios for libraries. To achieve this, linked data (LD) would be helpful.

Linked Data is one of the Semantic Web technologies. Berners-Lee,et. al(2009) stated that the term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web”. It encourages the institutions to publish, share and interlink their data Linked data In Linked data, data is expressed as simple statements using Resource Description Framework(a generic method of describing links between structured data in a graph-based data model), and connected using machine-readable Web-addressable identifiers to identify entities, Uniform Resource Identifier(URI). An RDF statement is a three-part subject-predicate-object structure commonly known as a "triple". The basis of a triple is its predicate, which is represented as an RDF property, and the specific subject and object of a triple are represented as members of RDF classes. Classes describe things, and properties describe the relationship between those things; classes and properties are the basic types of element in RDF. Dunsire & Willer (2009) stated that using current library standards as the bases of new triples and the extraction of triples from legacy records requires the representation of such standards in RDF, either by creating appropriate RDF elements or mapping to existing elements. This will not just allow the Semantic Web to benefit from library metadata; it should also improve interoperability between bibliographic entities, attributes, and relationships described by different, but related, standards. RDF properties can be chosen from different standards and mixed within a single application to meet its functional requirements (G. Dunsire, 2011).

Whereas LD help libraries to increase their presence and discoverability on the Web, National Library and Archive of IRAN (NLAI) are going to enrich its data by linking it to other external data sources. NLAI as mother library in IRAN is responsible for distributing and sharing library collection data. Therefore, we decided to publish authority file data as LD. The main contributions of this paper are: 1) presenting the first experiment of publishing authority files of NLAI as LD 2) classifying the main challenges of publishing Persian library linked data (LLD) 3) proposing potential solutions for those challenges and 4) would be a pioneer in mapping IRANMARC format to existing ontologies and vocabulary, choosing collections for external link regarding to Persian languages.

2 RELATED WOKRS

Tim Berners-Lee (2006), director of the World Wide Web Consortium, coined the term in a design note discussing issues around the Semantic Web project. LD supports the integration between data by connecting resources on related topics. We classify related works in two categories: researches on LD in general and researchers on LLD. LD in general refers to the studies which exploit LD usage in context of non-librarian applications such as spamming, question answering systems, content-based recommender systems, etc.

In the area of library linked data, The World Wide Web Consortium addresses library data as a central pillar of the Web of Data with the Library Linked Data Incubator Group[1] stressing the central role of library LD for the Semantic Web as a whole(Baker, et al. 2011). On the other hand library of congress (LOC) Bibliographic Framework Initiative General Plan demonstrated that the semantic web and related LD model hold interesting possibilities for libraries and cultural heritage institutions. Hence, Keßler, et al.(2009) express that libraries recognized that they are at the forefront of the LD movement. During these years,

[1] See: http://www.w3.org/2005/Incubator/lld/

high number of libraries published their catalogues as Linked Open Data and in a comparably high degree of standardization with accepted vocabularies.

DBpedia is the core of LD approach, other LD projects developed and connected to DBpedia gradually. In the realm of National library LD, National library of France (BNF), Hungarian National Library(NSZL), Spanish National Library(BNE), National library of Germany(DNB)(Hauser, 2012), National Library of Sweden(LIBRIS) , Library of Congress(LCSH), British National library (BNB) published their authority/bibliographic data as LD. Furthermore, other institution cooperates in library LD improvements; for instance, OCLC has been experimenting with LD for quite some time. For example, in August of 2012, bibliographic linked data for nearly 1.2 million WorldCat resources was published. The Australian National Data Service (ANDS) created a national collection of research resources in Australia.

Authoring LD is difficult because it demands the use of exact vocabularies and rules for making RDF data and linking RDF dataset. Some works discussed the facilities which ease the process. For instance, Jung & Park (2011) presented a system that helps non-expert users create RDF documents and LD easily. Using this system, a user can easily generate RDF documents and add new links between RDF entities without the complete knowledge about RDF grammar and vocabularies. However, publishing library LD is dependant to library standards and data itself. Library bibliographic data is more intricate than to make a comprehensive plan for publishing it as LD so ontology modelling and vocabulary selection are the most significant point. Various libraries have been using various vocabulary and ontologies to describe their resources as LD. For instance DNB uses Bibo[2] ISBD[3], Dublin Core[4], FOAF[5], BNE exploit FRBR[6], ISBD, Dublin Core.

The IFLA FRBR Review Group and ISBD/XML Task Group have worked in the development of representations of IFLA standards in Resource RDF. Consequently, IFLA Namespaces for the Functional Requirements (FR[7]) family of bibliographic metadata models have been published in RDF. The models include FRBR, Functional Requirements for Authority Data (FRAD), and Functional Requirements for Subject Authority Data (FRSAD). The FR element set vocabularies include RDF classes and properties corresponding to FR entities, attributes, and relationships. Each class and property has a URI for use in Semantic Web data triples. Philippe Le Pape (2011), member of the PUC[8] has stated that UNIMARC fully supports the FRBR structure. With the FRBR model the library community has been introduced to a new conceptual framework for bibliographic data. Aalberg, et al.(2011) express that alignment of the UNIMARC standard with FRBR is important to facilitate the implementation of LD publishing and the exchange of bibliographic data based on the FRBR model for a broad range of semantic aware applications.

Being developed RDF version of IFLA standards, libraries commenced to use it in core ontology modelling. Projects which focus on the extraction of FRBR entities and relationships from MARC-based records have demonstrated many of the possibilities and

[2] Bibliographic Ontology: http://purl.org/ontology/bibo/
[3] International Standard Bibliographic Description
[4] http://purl.org/dc/terms/
[5] http://xmlns.com/foaf/spec/
[6] Functional Requirements for Bibliographic Records
[7] See: http://iflastandards.info/ns/fr/
[8] Permanent UNIMARC Committee

problems of using MARC formats in this new context(Dunsire, 2011). Dunsire discussed some of the problems that exist when coding FRBR in UNIMARC .The problem of identity and attribute assignment mainly originates in cataloguing rules and is caused by insufficient information (such as lack of titles that properly identifies the works) or data fields that are ambiguous in the context of FRBRR. Aside from ontologies and vocabularies used in LD, to support identifying different repositories, meta-level descriptors of repositories should be represented using the VoiD vocabulary. K.Alexander,et al(2009) discuss the design and usecase for void (Vocabulary Of Interlinked Datasets) to describe linked datasets. They have released the voiD vocabulary to linked data communities in January, 2008.

3 OUR APPROACH

According aforementioned related works, we decided to exploit IFLA RDF models namely, FRBR, FRAD, ISBD in order to transform IRANMARC records to RDF linked data. NLAI dataset based on IRANMARC, its infrastructure is ISBD and represented as XML. Regarding to this research, NLAI aimed at creating a cohesive national collection of Iranian bibliographic resources which is a long-term plan. This paper outlines what we plan to do to achieve this goal. At present, NLAI database includes 982892 Authority records which are used in 3175125 bibliographic records; both are coded according to IRANMARC.

In March of 2013, a research plan," Publication of NLAI Collections as LD in order to Join Web of Data", is submitted to Research, Planning and Technology deputy at NLAI. However it has been verifying in research council, we have been starting our modelling. In the first level of implementation an incremental approach has been selected, so we intentionally considered a small scope by opting authority data for the beginning. . Our choice is due to the absence of Persian data in LOD and Persian authority files are the most referable data in publishing Persian LD. Publishing authority files would grantee a better persistence for publishing Persian bibliographic data in future. Thus, selecting name authority worksheet as starting point. Currently, authority files at NLAI are stored in oracle relational database (RASA) in IRANMARC format. Figure 1 shows architecture overview of publishing Persian LD process.

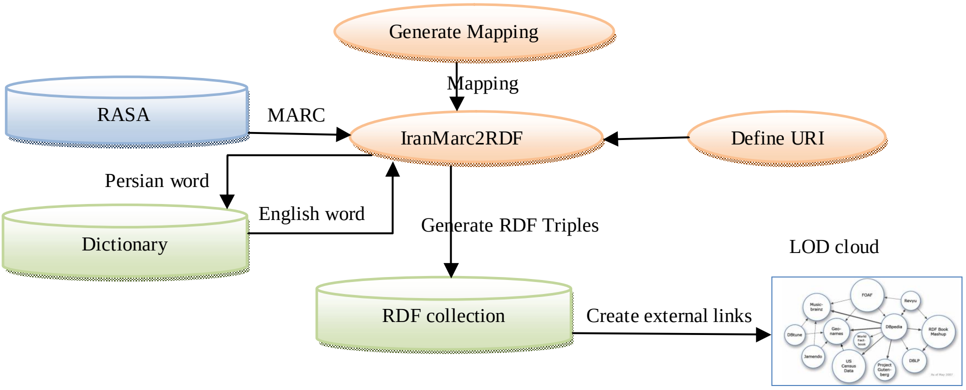

Figure 1 : Architecture overview shows the publishing Persian LD process. It uses IRANMARC Authority files of NLAI.

For each type of authority files a mapping is generated, authority file records are extracted from RASA relational database and then in IranMarc2RDF phase, according to the mapping which is created by specialist librarian, a RDF file is created and sorted in triple store (RDF collection). During the mapping process different vocabularies and ontologies are used. An important step in publishing a dataset as LD is designing a URl schema for addressing entities that are to be published. These URIs are used for creating external and internal links. Internal links between our data is set during RDF file generation. Links generation will be discussed further.

For data modelling we used various existing ontologies which match our authority files worksheets. We generated a manual mapping which defines the corresponding ontology for each part of IRANMARC authority files. In processing each IRANMARC record the mapping indicates the entities of IRANMARC records, the attributes of those entities and how these entities are connected to each other and an equivalent for each one in FRBR model. Therefore, NLAI bibliographic and authority records would consist of entities, attributes of those entities (properties), the entities relationships.

According to IFLA namespaces, we plan to start with these important properties which are shown in Figure 2. and then expand the scope incrementally.

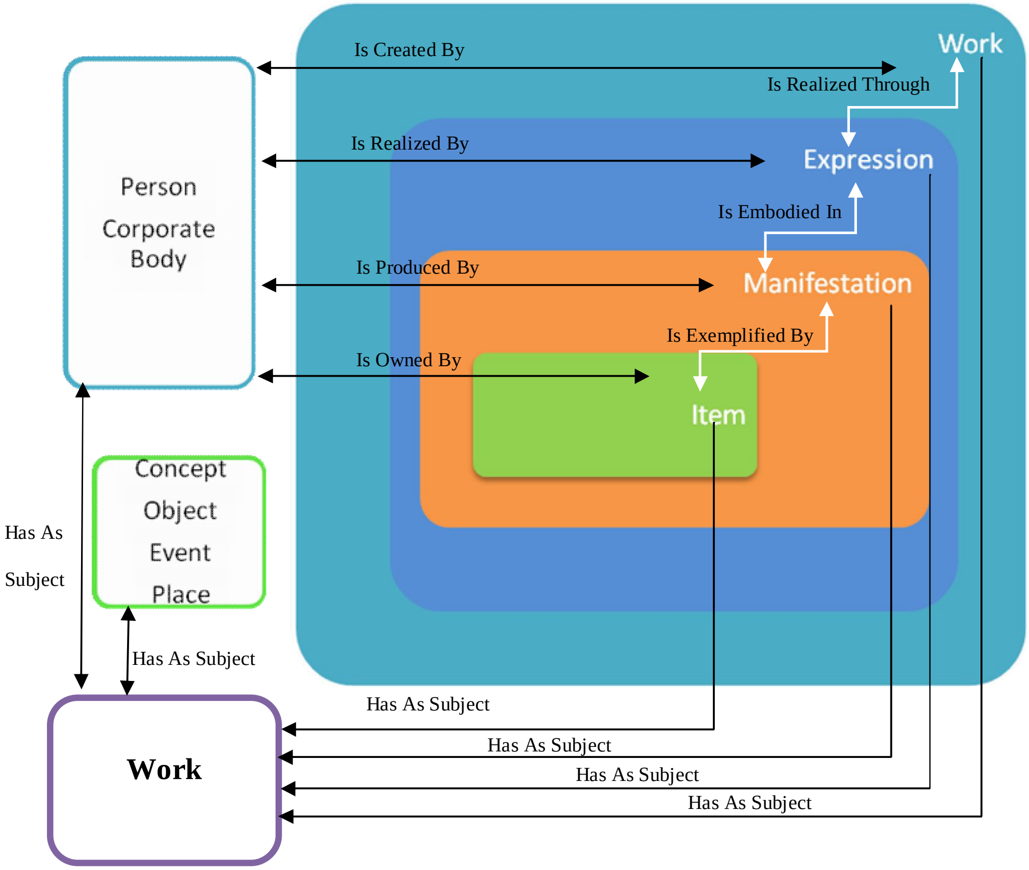

Figure 2 – FRBR model and object properties uses in our mapping

For authority data, each record mapped to one RDF instance of type Person, Corporate Body, Work or Expression. For bibliographic records, each record mapped to one RDF instance of type Manifestation. Then IFLA RDF properties employed to establish relationships between classes and their instances. A work IsCreated by Person/Corporate Body and IsRealized through an Expression which IsEmodied in a Manifestation that IsExampelified by an Item. To establish relationship between entities FRBR properties used (e.g. مولانا is creator of "مثنوي"). In order to describe the instances of each authority record, ISBD RDF properties used (e.g. resource hastitleproper شرح مثنوي). In all, FBBR RDF properties are used in mapping process to describe works, expressions entities from group1, persons and corporate bodies from group2 and ISBD RDF properties are used to describe manifestations. In manual mapping process these concept should be considered which don't lead to find mapping for all IRANMARC fields. Considering Figure 2 the main challenge is in what manner we could figure out entities from IRANMARC resources. Now it is done manually and not for all IRANMARC fields.

According to LD principles, each RDF instance should be identified by a single URI. For NLAI RDF resource identification we defined http://linkeddata.nlai.ir/rasa/identifier scheme for URI. General processing data field (001) is used as local part in the specified URI (e.g. http://linkeddata.nlai.ir /rasa/ 148350 is allocated as URI for "صادق هدايت" which is an instance of FRBR Person. Table 1 shows main classes (i.e. entities) URI from IFLA namespaces which we plan to use.

Table 1 – Classes according to IFLA standards

Entity (Class)	Namespace (URI)
Work	http://iflastandards.info/ns/fr/frbr/frbrer/C1001
Expression	http://iflastandards.info/ns/fr/frbr/frbrer/C1002
Manifestation	http://iflastandards.info/ns/fr/frbr/frbrer/C1003
Person	http://iflastandards.info/ns/fr/frbr/frbrer/C1005
Corporate Body	http://iflastandards.info/ns/fr/frbr/frbrer/C1006

To exemplify the mapping process, Table 2 shows some IRANMARC name authority fields and the equivalent entity it mapped to and also the corresponding URI for representing relationship.

Table 2 – IRANMARC tags mapped to appropriate property and classes

IRANMARC Name Authority Field		Mapped To	
001	General Processing Data	Used as local part of URI of a person	
101$a	Language of Entity	Property	http://purl.org/dc/terms/language
200$a 200$b	Entry element Part of name other than entry element	Entity Type	Person
		Same as	Owl:sameAs
		Property	hasNameOfPerson
		URL	http://iflastandards.info/ns/fr/frbr/frbrer/P3039
200$f	Date	Property	hasDatesOfPerson
		URL	http://iflastandards.info/ns/fr/frbr/frbrer/P3040
400$a	Variant Access Point – Personal Name	Property	http://iflastandards.info/ns/fr/frbr/frbrer/P4031
IRANMARC Bibliographic Field		Mapped To	
001	Record Identifier	Used as local part of URI of a manifestation	
200$a	Title Proper	Property	http://iflastandards.info/ns/isbd/elements/P1099

Regarding the external connectivity; number of links to external datasets; we considered the owl:sameAs object property in order to refer the equivalence links. In the terms of linking to external data sets, because persian external data sources are rare we decided to use popular existing data sources such as VIAF, LCSH, DBPedia then using a dictionary in order to link persian data to its correspondent english term.

As mentioned in other projects and studies, implementing FRBR in a library system means more than simply changing the format or the underlying data model. So we have a long way to go. For assessment, similar works have mainly focused on measures such as the number of RDF triples, or the number of links to external datasets, to indicate the quality of their work. We estimate that after uploading beta version of NLAI linked data set.

4 CHALLENGES AND SOLUTIONS

In this process IFLA RDF models help us to take advantage of the semantics behind FR models. However, there are still issues in generating overall solution for transforming NLAI IRANMARC records to LD. However previous works proved that IFLA RDF models facilitate both the process of data transformation and the development of user interfaces for navigation through data**Error! Reference source not found.**, nevertheless different problems and challenges are identified during this project. Here, we discuss these problems and recommend some solutions. Particularly, a common issue is the lack of experience for establishing persian LD.

- Linking Challenges

An important issue in publishing LD is to decide which ontologies should be used to describe the resources. Mostly ontology selection depends on its popularity. There are some ontologies which became standard in specific domains (es. Dublin Code, FOAF). But, popurality is not adequate for ontology selection. For instance there is no well-known ontology for specific domains. In order to obtain efficient interoperability we must choose accepted vocabularies for describing our data. We need methods that assist LD publishers to determine suitable ontologies. In library domains choosing ontologies must be considered the most importance step because it decreases the accuracy and quality of the results, especially for large, dynamic, and complex datasets such as library bibliographic data. Therefore, one of the challenges of LD is lack of a standard approach for choosing ontologies. On solution is to apply other institutions experiences.

- Data Interlinks

The task of linking data to external resources can be accomplished by tools. Some tools use matching techniques to detect semantic relationship between two entities. BNE[9] developed MARiMbA for this purpose(Daniel Vila-Suero,et al. 2012). Unfortunately toward LD publishing NLAI is still an infant, however we plan to ease these process by developing such applications.

- Persian Data challenges

Lack of data, incomplete or incorrect data in original dataset is another challenge in publishing LD. Some IRANMARC records are incomplete or is expressed in different in writing, or in different format type. Such problems lead to challenge in linking to external resources. To address this challenge, original data should be analyzed precisely to reveal existing problems. We could employ cleansing techniques to emend it. For instance, it is

[9] Biblioteca Nacional de España

possible to implement an algorithm to convert different formats of some fields to a unique one. Since most data on LOD cloud is published in English, it is hard for us to link a persian dataset to external datasets. In multi-language systems where data is generated by end users. some users choose their mother tongue language while others use English for entering their data. For instance, instead of ‘John’one could enter ‘جان’. As another example, identical Persian terms exist in different English forms in the database, e.g. a single Persian name "اسلامي" is entered both as "Eslami" and "Islami". Such problems caused by multi-lingual data, introduce challenges when searching external datasets for related resources to be linked, and decrease the quality of the published dataset. Therefore, there should be a mechanism for finding English equivalents of the Persian terms. As a solution we plan to use a local dictionary to find appropriate equivalent of the required term. Thus, all equivalent of ‘اسلامي’ (ie. Islami, Eslami) would be checked when searching external datasets.

- Link Maintenance

It is significant to maintain the links and also data quality in LD. Updating interlinked datasets may cause invalid links. It may need updating existing links. If external datasets defines their last modification, we could decide when to update out own. On the other hand, as the original dataset changed it is necessary to update LD. Thus, information such as ‘time of creation’, ‘time of modification’ should be published along with dataset. To address this requirement we use dcterms:created and dcterms:modified in voiD(Vocabulary Of Interlinked Datasets) specification of our dataset.

5 CONCLUSION AND FUTURE WORKS

This work was done in accordance with the principles of LD. For future we would make knowledge bases available as RDF from library bibliographic and authority records. It needs time to investigate on the alignment of IRANMARC and standards such as FRBR, ISBD, FRAD. NLAI has planned to use Drupal content management system for representing LD. Exposing NLAI LD sets will provide efficient services for end users and researchers. The implementation of an interface for searching NLAI linked data dataset is also considered as our coming works. Furthermore, these data would be interrelated with other existing knowledge bases in the world such as DBPedia/Wikipedia other linked data datasets exist in the world. This is pretty easily done once you have the data in RDF by the way. On the other hand we try to gradually define a comprehensive data model which includes all the important entities for publishing IRANMARC bibliographic records in data web. Some reasons that forces NLAI to keep going are: others could link to our data more easily, we could build new services based on our content, and finally we could become a "canonical reference point" locally and globally.

6 REFRENCES

Hyosook Jung,Seongbin Park,(2011).A System for Linked Data Creation,Studies in Informatics and Control,vol. 20 . Issue 4.Avaiable at ;http://sic.ici.ro/sic2011_4/art04.php

Jan Hannemann & Jürgen Kett (2010). Linked Data for Libraries, in world library and information congress: 76th IFLA general conference and assembly, Gothenburg,,, p. 12.

Bizer, Christian; Tom Heath, Tim Berners-Lee (2009). Linked data: the story so far ,International Journal on Semantic Web and Information Systems (IJSWIS), vol. 5, issue 3. Available at: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Dunsire, Gordon; Mirna Willer (2011). UNIMARC and Linked Data, IFLA Journal 37, 4, 314-326, Available at: http://conference.ifla.org/past/ifla77/187-dunsire-en.pdf

Thomas Baker, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, Jodi Schneider, Ross Singer, Ed Summers, William Waites, Jeff Young, and Marcia Zeng. Library Linked Data Incubator Group Final Report. W3C Incubator Group Report, available at http://www.w3.org/2005/Incubator/ lld/XGR-lld-20111025.

Carsten Keßler, Mathieu d'Aquin, and Stefan Dietze (2012). Linked Data for Science and Education. Semantic Web Journal vol. 4, no1.

Julia Hauser(2012).Linked Data Service of the German National Library. Available from http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

Library of Congress.Subject Headings. Available from http://id.loc.gov/authorities/subjects.html. Last accessed2012-98-26.

Trond Aalberg, Jan Pisanski and Maja Žumer(2011).UNIMARC and FRBR - can we have both? Advancing UNIMARC: alignment and innovation, IFLA.UNIMARC and linked data

Philippe Le Pape (2011).Expressing FRBR in UNIMARC Yes we can! Advancing UNIMARC: alignment and innovation — IFLA.

Daniel Vila-Suero, Boris Villazón-Terrazas, Asunción Gómez-P,(2012). datos.bne.es: a Library Linked Data Dataset. In Semantic Web Interoperability, Usability, Applicability an IOS Press Journal

Tim Berners-Lee (2006-07-27). "Linked Data—Design Issues". W3C. Retrieved 2013-05-18.

Keith Alexander, Michael Hausenblas (2009).Describing linked datasets - on the design and usage of void, the ’vocabulary of interlinked datasets,In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference.

*_**Saeedeh Eslami**_ : Master of Science
Technology and Communication Department, National Library and Archive of IRAN(NLAI)
s-eslami@nlai.ir
+98 21 81622440 phone

Saeedeh Eslami was born in 1983 in Tehran, Iran, she holds BA degree in Computer Software Engineering in 2006 and then graduated in MA in Computer Software Engineering in 2010. She is Fellow member of staff NLAI and has been working as software specialist since 2005 and has participated prominently in NALI software projects. She is software analyst and programmer of Software Architected and Design Group. She is a university lecturer, teaches at Islamic Azad University and NLAI. Her research interest lies around Semantic Web with specific interest in developing Linked Data, Ontologies and Topic Maps, Free/Libre Open Source Software development and interoperability and so on. She has published several Papers on such areas.

Mohammad Hossein Vaghefzadeh : Master of Science
Deputy of Research and Development, Planning and Information Technology , National Library and Archive of IRAN(NLAI)
m-vaghefzadeh@nlai.ir
+98 21 81623270 phone

Mohammad Hossein Vaghefzadeh was born in 1967 in Tehran, Iran. He holds BA/MA degree in Industrial engineering. He was the General Director of ICT Department till May 2013 and now the deputy of Research and Development, Planning and Information Technology, and also CEO of Rahyab, Company in Ports Designing. He is a university lecturer, teaches at several University. Her research interest lies around Simulation, Engineering Economy and Strategic Planning. He has published several papers and book titled "Programming with Turbo Pascal" in 199.

Language of Presentation: English

The corresponding author can be contacted at: eslami.saeedeh@gmail.com , s-eslami@nlai.ir