Publishing MARC 21 Format for Bibliographic Data as Linked Data With Provenance

Ujjal Marjit Kumar Sharma Utpal Biswas

Abstract

Publishing Bibliographic data as Linked Data into the Web of Data provides information that are rich in semantics and syntax. Linked Data approach allows machine to analyse the data without human intervention. Data becomes reusable and publicly revealed into the web. Such data are always shareable among different data sources. Data that are published into the Web of Data can be increased in size. In spite of having meaningful and structured data into the web, their metadata or provenance are available at a minimum rate. In order to provide the utility and the glimpse of the data, the enquiry such as, what a dataset will contribute, who published it, their access mechanisms and licensing information are essential for the data consumers. In this paper, we present an approach for publishing MARC 21 Format for Bibliographic data as Linked Data with Provenance. We believe that the provenance information will help data consumers, mainly in library domains, to determine the quality & uniformity of the bibliographic information.

Keywords: Linked Data, MARC 21, Semantic Web, Provenance, VoID

1. Introduction

Machine readable catalogue has brought a new era in library & information science. People in various libraries were able to communicate the bibliographic information relying on a particular format. In the present day, MARC 21 (MAchine Readable Catalogue) format serves as the basis for communicating bibliographic data among libraries. Several national libraries today are based on MARC 21 format. However, these formats barely talks about the semantics of data being communicated. Since the data is being served for humans, the machine is aware of only the syntax & format of the data. Besides, the issue of sharing, exchanging, interoperability among the data is always there. However, today these limitations are overcome by using Linked Data. Linked Data allows data publishers to publish the structured data into the web. Data is modelled using Resource Description Framework (RDF) and the knowledge is shared among data using Web Ontology Language (OWL).

Several data presented in other legacy formats such as RDBMS, CSV, XML are published into Web of Data. The size is doubled every year, as provided by (CYGANIAK Richard and Jentzsch Anja), the statistics of the datasets published into the Linking Open Data Community[1]. These datasets are available at the Linked Data Community Project[2]. Provenance information should be presented for the end users along with the original dataset to observe the data quality, the terms and conditions of data usability and the reliability of the data. By viewing provenance information, people can easily find their desired information. Any user or librarian searching for bibliographic data may want to know the availability, accessibility, the licensing information and other related sources of the original data. By knowing these relevant information one might take further decisions.

In this paper, we present our work of publishing MACR 21 Format for Bibliographic Data as Linked Data with Data Provenance. Our work to generate provenance information is based on both manual and automatic support. The provenance information is represented using VoID (Vocabulary of Interlinked Datasets) and are attached with the original dataset by adding VoID back-links.

The paper is organised as follows: we discuss related work in Section 2, a brief overview of Linked Data is discussed in Section 3, Section 4 presents the concept of Data Provenance, provenance representaion & storage techniques. In Section 5 we present our proposed work and Section 6 concludes our work.

2. Related Work

Tope Omitola et al. (OMITOLA Tope, Zuo Landong, Gutteridge Christopher, Millard Ian C., Glaser Hugh, Gibbins Nicholas, Shadbolt Nigel, 2011) has presented an approach on publishing the provenance information of linked dataset using VoIDP. VoIDP is a lightweight provenance vocabulary, an extension to the VoID vocabulary, which enables publishers to depict the provenance information of their linked datasets. VoIDP provides rich set of classes and properties related to event generated information along with general information. We can easily integrate the VoIDP vocabulary into current practice in order to enrich the provenance information. Matias et. al. (FROSTERUS Matias , Hyvonen Eero) have developed a tool and metadata editor for annotating and publishing the metadata about RDF datasets as well as the non-linked datasets. They also have extended the VoID vocabulary to add required classes and properties. Furthermore, the library related data are also published by a number of prominent researchers. Library of Congress Controlled Vocabularies (Corey A. Harper and Barbara B. Tillett, *2007*) has developed vocabularies for the library domain. Rob Styles et. al. (STYLES Rob, Ayers Danny, Shabir Nadeem, 2008) have presented an approach on converting MARC 21 records to their RDF equivalent by applying mapping rules between MARC 21 and RDF. The resulting RDF library datasets are linked to other sources such as Library of Congress Linked Data sources, DBPedia, Geonames etc. Martin Malmsten (MALMSTEN Martin, 2008, 2009) has presented the implementation of Linked Data of library resources available in Swedish Union Catalog (LIBRIS). Their implementation is based on a server called RDF wrapper around the Integrated Library Systems (ILS) which, upon request, extracts the MARC 21 records in MARC-XML format. The resulting XML records are transformed in the desired format using EXtensible Stylesheet Language (XSL).

A number of community and agencies have also announced the library structured metadata as Open RDF datasets, value vocabularies, and metadata element sets. British National Library (BNB)3, Europeana Linked Open Data4, Cambridge University Library dataset5, Hungarian National Library6, Library of Congress Subject Headings7, Biblioteca Nacional de Espana (BNE, Spanish National Library) have equally contributed their library data into Web of Data.

3. Linked Data

Linked Data is a method for publishing and interlinking structured data on the web. It provides best practices for making data more transparent, robust and efficient. Since there is no interlinking between data in the

existing web, to alleviate the same Linked Data allows each data to be linked and shared the knowledge among them. Data is represented using RDF (Resource description Framework) and identified by using URIs (Uniform Resource Identifier). The HTTP access mechanism is used for designing URIs so that every resource described is globally accessible on the web. Each resource, defined in a dataset, carries a particular knowledge which is assisted by the Web Ontology Languages (OWL). Ontologies provide knowledge concerning the semantics of the data. Thus, the data presented in RDF, which is highly structured, is readable by the machine. The data publishers also provide the data in HTML representation which is mandatory in the context of Linked Data. In this way, the data presented in the web are available to be consumed by both machine as well as human without any intervention of user. Sir Tim Berners-Lee (HEATH Tom, Bizer Christian, 2011) has proposed four principles for publishing the structured data on the web which are as follows:

- Use URIs for naming things.
- Use HTTP URIs, so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Include links to other URIs, so that they can discover more things.

By applying above mentioned rules we can have RDF representation of the library resources such as a book, its author, place of publication, name of the publisher, etc. Each such resource will have unique HTTP URI and are discoverable on the web. Hence, they can be published as web of data rather than web of documents. Eventually, the library resources could be linked, reused and integrated with the data from other sources. Furthermore, ontologies also facilitates the task of categorising the library resources as well as optimises the process of knowledge sharing. Christian Bizer et. al. (BIZER Christian, Heath Tom, Cyganiak Richard) have presented the concepts, and best practices regarding Linked Data implementation. They also discussed how to publish Linked Data, the design architecture, approaches on choosing URIs, and setting RDF links to other data sources.

4. Provenance

Data Provenance is a method of collecting relevant metadata about data. It provides information of an object regarding its origin, the method by which it is recorded, the format and the language of the data representation. Buneman et al. (BUNEMAN Peter, Khanna Sanjeev, Tan Wang-Chiew. 2001) define data provenance in the context of database systems as *"the derivation of a piece of data that is in the result of the transformation step or description of the origin of a piece of data and the process by which it arrives in a database"*. Data Provenance also known as tracing data back, which means knowing data about data by tracing its origin. Data Provenance helps in bringing adequate metadata information so that it guarantees the quality, trustworthy, and acceptability of the data. It guarantees that the information that are available on any source are valid, usable, authorised and legal.

4.1. Provenance Representation & Storage

In our previous work (MARJIT Ujjal, Sharma Kumar, Biswas Utpal. 2012), we had reviewed the provenance representation and storage techniques. Basically, the provenance of any data is represented by two methods: *Annotation Method* and *Inversion Method. In Annotation Method* the metadata are pre-computed and stored in a document. In *Inversion Method* the metadata are collected on the fly based on user defined queries. In our practice we follow A*nnotation Method* where we need to store the provenance of the RDF dataset in a separate document.

One way to implement A*nnotation Method* in the context of Semantic Web is by using VoID (Vocabulary of Interlinked Datasets). Detail implementation regarding VoID is discussed by (ALEXANDER K., Cyganiak R., Hausenblas M., and(Zhao J. 2009). VoID allows publishers to define the metadata of their dataset and publish it separately. VoID is also known as the ontology or the vocabulary which provides a collection of classes and properties to define the metadata about RDF datasets. VoID tells everything about the nature and feature of the dataset, its access mechanisms, statistical information, publishers and interlinking of the data sources. Further the information it provides is categorised into General metadata, Access metadata, Structural metadata, and the information about Interlinking between datasets. Two main concepts or classes are found in VoID which are as follows:

1. **Dataset**: Dataset is a collection of RDF statements fully designed based on the Linked Data principles, and is published and maintained by a single data provider. It provides meaningful information on the web and it should be hosted on a particular server. The publisher of the dataset should provide all the relevant information such as SPARQL End-Points and URI of RDF dumps, information about vocabularies used and other general metadata. *void:Dataset* class is used to model the dataset instance.

2. **Linkset**: In a Linked Dataset, there presents so many outgoing links which are called RDF links. These links actually do the job of interlinking between the source and the target datasets so that consumers of the dataset would find more information. This is mandatory in Linked Data as illustrated in 4th Linked Data principle. A linkset is a collection of such RDF links. *void:Linkset* is used to model the link set instance.

To publish linked data along with provenance the data provider should exactly follow the above Linked Data principles as well as the provenance principles introduced by Edoardo Pignotti et al. (PIGNOTTI Edoardo, Corsar David, Edwards Peter, 2011), which are as follows:

- **One Star Provenance**: Publish the provenance of data on the web in whatever format (e.g. plain text).
- **Two Star Provenance**: Publish provenance as structured data (e.g. database, spreadsheet, XML)
- **Three Star Provenance**: Use URIs to identify individual elements within the provenance record.
- **Four Star Provenance**: Link provenance record to other provenance records using RDF.

The above provenance principles are guidelines and recommended way to publish provenance information of linked datasets. We adhere to these principles and attempt to produce Three Star Provenance. In the following section we discuss our framework for publishing MARC 21 Format for Bibliographic data as Linked Data with provenance information.

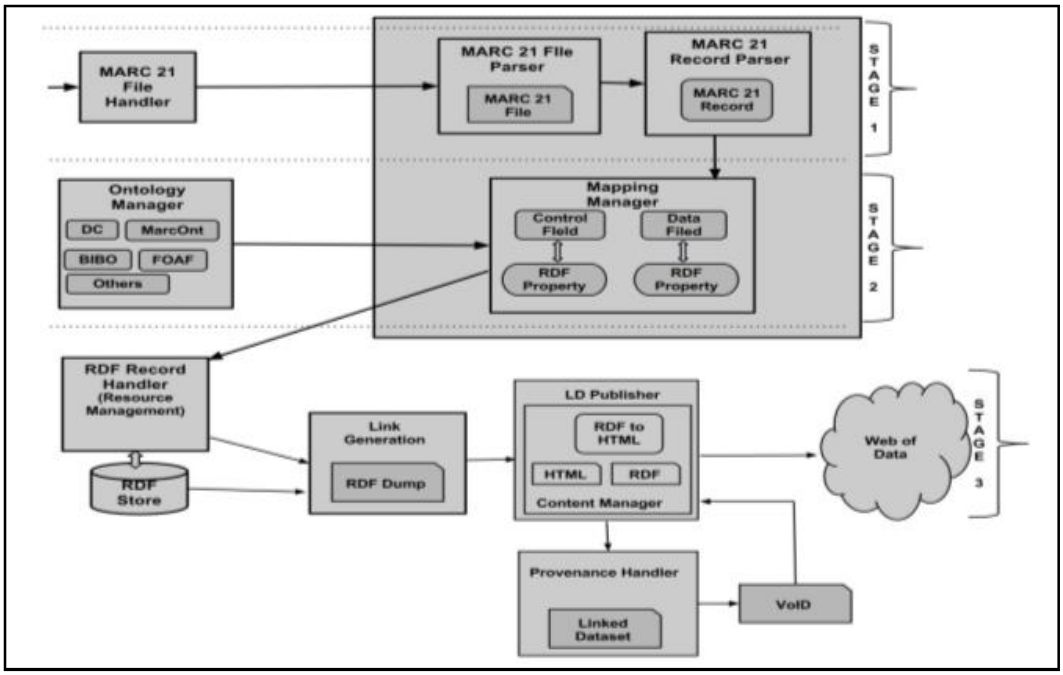

Figure 1: Framework for publishing MARC21 Format for Bibliographic data as Linked Data with Provenance Information.

5. Proposed Framework

Fig. 1 & 2 shows the work flow regarding conversion process from MARC 21 Format for Bibliographic data into Linked Data. In stage 1 the MARC 21 file is taken as the initial input. We have been using *.mrc* or *.bib* MARC 21 file extensions. The MARC 21 file is parsed by the MARC 21 File Parser which results into individual MARC 21 records. MARC 21 record is parsed by MARC 21 Record Parser which results into leader, control fields and the data fields. The leader, individual control fields and the data fields are interpreted carefully. The Mapping Manager is responsible for mapping these information with their equivalent RDF terms. The knowledge to which MARC 21 data is to be mapped with the equivalent RDF term is supplied by the Ontology Manager. Ontology Manager holds the ontologies or the controlled vocabularies. These ontologies or controlled vocabularies provide semantic terms by which the attributes and the relationships are added to the RDF resource. A MARC 21 record is treated as an RDF resource and each of the data within MARC 21 records are treated as property, sub-property or relationship.

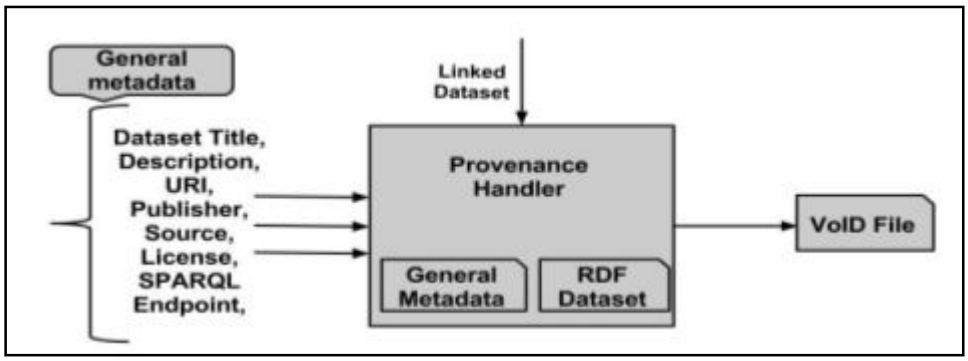

Figure 2: Provenance Handler


```
- <rdf:RDF>
  + <rdf:Description rdf:about="http://localhost:8080/BibliographicLinkedData/BibResources/Manifesto_to_the_Mexican_Republic"></rdf:Description>
  - <rdf:Description rdf:about="http://localhost:8080/BibliographicLinkedData/BibResources/Echoes_of_the_past_:">
      <DC:subject>Bidwell, John,</DC:subject>
      <DC:type>Language material</DC:type>
      <rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Future_Echoes_Of_The_Past"/>
      <rdagroup1elements:titleProper>Echoes of the past :</rdagroup1elements:titleProper>
      <rdagroup1elements:otherTitleInformation>Echoes of the past</rdagroup1elements:otherTitleInformation>
      <marcont:hasTitle>American West (Marlborough, England)</marcont:hasTitle>
      <DC:coverage>Adam Matthew Digital (Firm)</DC:coverage>
      <DC:extent>91 p., [3] p. of plates :</DC:extent>
      <marcont:hasNumber>AC001769</marcont:hasNumber>
      <rdagroup1elements:dateOfPublication>[1914]</rdagroup1elements:dateOfPublication>
      <marcont:hasRecordStatus>New</marcont:hasRecordStatus>
      <rdagroup1elements:placeOfProduction>Marlborough, England :</rdagroup1elements:placeOfProduction>
      <DC:medium>[electronic resource].</DC:medium>
      <DC_11:format>cr bn ---uuuuu</DC_11:format>
      <VCARD:Locality>Overlandjourneys_to_the_Pacific.</VCARD:Locality>
      <rdagroup1elements:placeOfPublication>Chico, Calif. :</rdagroup1elements:placeOfPublication>
      <marcont:hasRecordLength>1776</marcont:hasRecordLength>
      <DC:coverage>Newberry Library.</DC:coverage>
      <marcont:hasDate>20090907111654.5</marcont:hasDate>
      <VCARD:Locality rdf:resource="http://dbpedia.org/resource/California"/>
      <rdagroup1elements:publishersName>Chico Advertiser,</rdagroup1elements:publishersName>
      <marcont:hasCoverage>090617s1914 caua s 000 0deng d</marcont:hasCoverage>
      <VCARD:Locality>Frontierand_pioneer_life</VCARD:Locality>
    - <marcont:hasURL>
        http://www.americanwest.amdigital.co.uk/Contents/Document-Details.aspx?documentid=1258
      </marcont:hasURL>
      <marcont:hasAuthor rdf:resource="http://dbpedia.org/resource/John_Bidwell"/>
    - <rdagroup1elements:otherTitleInformation>
        an account of the first emigrant train to California, Frâemont in the conquest of California, the discovery of gold, and early reminiscences
      </rdagroup1elements:otherTitleInformation>
      <rdagroup1elements:productionMethod>Electronic reproduction.</rdagroup1elements:productionMethod>
      <rdagroup1elements:keyTitle>Echoes of the past about California</rdagroup1elements:keyTitle>
      <DC:type rdf:resource="http://purl.org/dc/terms/BibliographicResource"/>
    </rdf:Description>
  + <rdf:Description rdf:about="http://localhost:8080/BibliographicLinkedData/BibResources/Read_and_ponder!"></rdf:Description>
</rdf:RDF>
```


Figure 3: Linked Dataset

The resulting RDF terms are then stored into the RDF store. The challenging task in this process is the unique assignment of the URIs to each of the resource and their relationship attributes. Each of the resource and their relationship attributes are uniquely identified throughout the system by query and retrieve mechanisms.

In Stage 3 we perform the task of link generation and in the subsequent step we generate provenance information of the dataset being created. If a resource is already defined in other sources, then that resource is dropped out from the local version and put the RDF links. Once we achieve the full Linked Data version of the library data adhering to the four principles of Linked Data as discussed above, our library data is ready to be published and can be shared into the Web. **Fig. 3** shows an excerpt of the generated linked dataset. The generated dataset has few RDF links pointing to the resources of the DBpedia dataset.

5.1. Provenance Generation

Once the linked data version of the dataset is ready, we move to the step to generate the provenance information. **Fig. 2** shows detail about the *Provenance Handler.* Our approach first begins with the general metadata such as dataset title, description, name of the publisher, source of the origin etc. which are manually entered. *Provenance Handler* also takes the linked dataset as the input from *LD Publisher.* Based on the general metadata and linked dataset it generates the provenance information. It automatically processes the linked dataset and results the statistical information, name and number of vocabularies used, and the information about interlinking datasets. The final output is the VoID file which is stored separately. **Listing 1** shows the content of the generated VoID file describing the linked dataset.

```
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
@prefix DC:     <http://purl.org/dc/terms/> .
@prefix rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix void:   <http://rdfs.org/ns/void#> .
@prefix VCARD:  <http://www.w3.org/2001/vcard-rdf/3.0#> .

<link-dataset1>  a      void:Linkset ;
   void:linkPredicate rdfs:seeAlso;
   void:target <dataset1> , <http://localhost:8080/BibliographicLinkedData/> .

<dataset1>  a      void:Dataset ;
   DC:description "DBpedia aims to provide structured information that already
                   present in Wikipedia. The extracted information are modelled using
                   RDF & URI and published into Web of Data." ;

   DC:license "GNU General Public License" ;
   DC:publisher "University of Leipzig, Freie Universität Berlin, OpenLink Software" ;
```

```
  DC:title "DBpedia" ;
  foaf:homepage "http://dbpedia.org/" .

<http://localhost:8080/BibliographicLinkedData/>  a  void:Dataset ;
  DC:title "Bibliographic Linked Data " ;
  DC:description "The bibliographic information provided here are converted from MARC
     21 Format. The bibliographic information are related to Language Material";
  DC:format "RDF" ;
  DC:language "English" ;
  DC:publisher "University of Kalyani" ;
  DC:source "http://www.amdigital.co.uk/librarians-resources/marc-records/" ;
  void:distinctObjects  8437 ;
  void:distinctSubjects  648 ;
  void:sparqlEndpoint "http://localhost:8080/BibliographicLinkedData/sparql" ;
  void:feature "application/rdf+xml" ;
  void:triples 18262 ;
  void:vocabulary "http://www.w3.org/2000/01/rdf-schema#" , "http://purl.org/dc/elements/1.1/" ,
     "http://purl.org/dc/terms/" , "http://www.marcont.org/ontology/2.1#" ,  "http://RDVocab.info/
        elements/" , "http://openlibrary.org/type/edition#" ,  "http://www.w3.org/2001/vcard-rdf/3.0#"
```

Listing 1: Provenance Information (VoID Dataset)

5.2 Publishing Provenance Information

Provenance storage is one of the current research challenge. Generally, provenance can be stored with the original dataset as well as in a separate document. Both of these have certain limitations. In case of former it suffers from scalability issue if the provenance information grow in size. In the later case, once the provenance information is generated and stored separately, it is necessary to keep change management. Because if anything related to statistical information or vocabularies is changed in the original dataset, then the change should be reflected in the provenance information as well. Publishing provenance information of the Linked Datasets is similar to publishing Linked Data. It should be physically deployed in a hosted server and provide a link to the original dataset. *VoID backlink*[1] approach is developed for interlinking the RDF dataset and VoID file using the property *void:inDataset. void:inDataset* is a triple which points to the URI of the VoID dataset.

6. Conclusion

In this paper we have presented an approach on transferring MARC 21 Format for Bibliographic Data into Linked Data with Provenance. We have reviewed different provenance representation methods and implemented an *Annotation Method* of provenance representation using VoID. Further we have integrated this method to generate the provenance information of the linked dataset. We believe that the provenance

information will help data consumers, mainly in library domains, to determine the quality and consistency of the bibliographic data presented on the Web of Data. Our research is still in progress and in future we will be implementing the SPARQL End-point of the linked datasets and also will investigate how the SPARQL queries over provenance information will be benefited to the data consumers.

References

1. CYGANIAK Richard and Jentzsch Anja. The Linking Open Data cloud diagram. Available at http://lod-cloud.net/. (Accessed on 10/02/2013).
2. HEATH Tom, Bizer Christian. (2011). Linked Data Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers, ISBN 978160845431, 2011.
3. BIZER Christian, Heath Tom, Cyganiak Richard. How to Publish Linked Data on the Web. Available at http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/. (Accessed on 10/02/2013).
4. BUNEMAN Peter, Khanna Sanjeev, Tan Wang-Chiew. (2001). Why and Where: A Characterization of Data Provenance. In Proceedings of the 8th International Conference on Database Theory (ICDT). Springer, January 2001.
5. MARJIT Ujjal, Sharma Kumar, Biswas Utpal. (2012). Provenance Representation and Storage Techniques in Linked Data: A State-of-the-art Survey. International Journal of Computer Applications (0975 – 8887) Volume 38– No.9, January 2012.
6. ALEXANDER K., Cyganiak R., Hausenblas M., and(Zhao J. (2009). voiD guide—Using the Vocabulary of Interlinked Datasets. Community Draft, voiD working group, 2009. http://rdfs.org/ns/void-guide/
7. ALEXANDER K., Cyganiak R., Hausenblas M., and(Zhao J. (2009). voiD, the Vocabulary of Interlinked Datasets. Community Draft, voiD working group, 2009. http://rdfs.org/ns/void/
8. PIGNOTTI Edoardo, Corsar David, Edwards Peter. (2011). Provenance Principles for Open Data. In Proceedings of DE2011, November, 2011.
9. OMITOLA Tope, Zuo Landong, Gutteridge Christopher, Millard Ian C., Glaser Hugh, Gibbins Nicholas, Shadbolt Nigel. (2011). Tracing the Provenance of Linked Data using voiD. The International Conference on Web Intelligence, Mining and Semantics (WIMS'11).
10. FROSTERUS Matias , Hyvonen Eero. Creating and Publishing Metadata of Linked Data —Providing Shoes for the Cobbler's Children.
11. Corey A. Harper and Barbara B. Tillett. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. Published in Cataloging & Classification Quarterly, Volume 43, no. 3/4, 2007.

12. MALMSTEN Martin. (2009). Exposing Library Data as Linked Data. Presented at the IFLA satellite preconference sponsored by the Information Technology Section "Emerging trends in technology: libraries between Web 2.0, semantic web and search technology", Florence, 19 20 August 2009.

13. MALMSTEN Martin. (2008), Making a Library Catalogue Part of the Semantic Web. *2008* Proc. Int'l Conf. on Dublin Core and Metadata Applications.

About Authors

Mr. Ujjal Marjit is the System-in-Charge at the C.I.R.M.(Centre for Information Resource Management), University of Kalyani
E-mail: sic@klyuniv.ac.in

Mr. Kumar Sharma is a research scholar of the Department of Computer Science & Engineering, University of Kalyani, India.
E-mail: kumar.asom@gmail.com

Dr. Utpal Biswas received his B.E, M.E and PhD degrees in Computer Science and Engineering from Jadavpur University, India
E-mail: utpal01in@yahoo.com