Rethinking Metadata: Semantic Proposal for Future

Rajarshi Das ***Vikas Bhushan*** ***Sudipta Shee*** ***Santu Mondal***

Abstract

This paper starts with describing Metadata and subsequently proposes a linked open metadata model which can unearth vast firsthand knowledge hidden in the deep layer of various libraries of the world. The benefit of the model can be multipronged. It can help generate new knowledge as in many cases of cultural artifacts we do not have any account which describe the same except for the metadata that attaches with it. Metadata can also be tailor made to serve the purpose of distant reading of large corpus of literary resources of human civilization. Finally vernacular metadata schema can help in spreading knowledge while serving community information requirement.

Keywords: Distant Reading, Linked Data, Linked Open Metadata, Metadata, Semantic Web

1. Introduction

Metadata as we know it is a structured data about data. Today, the world is replete with information. In digital domain another set of information have come into being in order to retrieve the required information and more precisely to locate it in its rightful context. This is commonly known as metadata.

A User Guide For Simple Dublin Core describes metadata in following manner. Metadata describes an information resource. The word "Meta" comes from a Greek word that denotes something of a higher and fundamental nature. Metadata is then data about data. It is the internet age term for information that librarians have put into catalogue and most commonly known as descriptive information about web resources (El-sharbini,2014).

Metadata is divided according to the purpose it serves. They are in the following (Haynes,2009).

Administrative Metadata is generally used in managing and administering information resources as a whole, acquisition of information is the particular case in point.

Descriptive Metadata, as the name suggests is used to describe or identify information sources and cataloging records are one such example.

Preservation Metadata is related also with information sources along with information resources, for example any documentation of actions of preservation of all sorts of physical and digital versions of resources likes data refreshing and migration

Technical Metadata describes how this whole system of data about data thing functions in real time scenario or in simple terms it records the behavior of metadata. For example any firsthand documents created after testing and successfully implementing any hardware and software and digitization of information are some case among many other cases.

There are also some other uses of metadata based on different types and levels of information resources. The above mentioned are the basics of

10th International CALIBER-2015
HP University and IIAS, Shimla, Himachal Pradesh, India
March 12-14, 2015

metadata usage, described to explore metadata by showing what it does.

1.1 Scope and Coverage

The scope and purpose of this paper is to derive an idea and propose a model for metadata which is open and democratic in nature. This model takes into consideration huge cultural artifacts that are lying idle in the various libraries, museums and other repositories of the world and attempts to devise a metadata model for linking and making possible distant readings of these artifacts. It also proposes a vernacular model of metadata which can eventually facilitate in spreading community information resources.

2. Related Works

Tillett, in his work explored the opportunity for library wisdom and practices to contribute in global semantic web endeavor. AACR2 has reflected on creating standards for metadata. New opportunities for using these records in the digital world are described (interoperability), including mapping with Dublin Core metadata (Tillett, 2003). Semantic layer can be applied to an existing group of resources. Moreover the integration and interoperability with the other technologies related to the semantic web, such as Dublin Core and RDF (Meschini, 2005).

Significantly increase in activity over the past few years to integrate library metadata with the Semantic Web along with the development of controlled vocabularies as "linked data" (Dunsire & Willer, 2011). Semantic Web activity in a W3C project whose goal is to enable a 'cooperative' World Wide Web where machines and humans can exchange electronic content that has clear-cut, unambiguous meaning. (Heery & Wagner, 2002). Taniguchi described and explain the current situation of metadata (Taniguchi, 2010).

Metadata of library catalogues can stand autonomously, providing valuable information detached from the resources they point to and therefore could be used as data in the context of the Semantic Web (Peponakis, 2013). Automatic global interpretation must be facilitated without recourse to a universal semantic schema (Niederee, 2003). The library cataloguing systems are for non-restricted and completely free to access. Libraries of Bavaria, Berlin and Brandenburg which decided in 2011 to publish their shared network catalogue with nearly 23 million records as open data and as linked open data. (Messmer, 2013).

3. Semantic Web and Linked Open Data

Linked data is about using the Web to create typed links between data from different sources. It is based on a set of principles to publish structured data on the Web so that it can be interlinked and become more useful. These principles allow data to be published on the Web in order to be machine-readable, have the meaning explicitly defined, be linked to other sets of external data, and in turn with the possibility of being linked from external data sets (Bizer, 2009). According to Tim Berners-Lee, a set of best practices have been set for the publication of data on the Web in a way that all published data becomes part of a single global data space. That set of rules - known as the "principles of Linked Data" - claim that every piece of data must has an associated URI (Uniform Resource Identifier) that, when looked up, should provide useful information, using the standards RDF and SPARQL (Simple Protocol and RDF Query Language). Farther, the data should include links to other URIs so that more things can be discovered either by people or by machines. While the current Web is based on HTML to describe un-typed documents connected by hyperlinks, the Linked Data relies on documents containing data in RDF to make typed statements that connect arbitrary things in the world (Berners-Lee, 2006).

Summarizing, the principles of Linked Data provide the basic mechanism for publishing and connecting data using the infrastructure of the Web, taking advantage of its architecture and standards (Bizer, 2009) and thus forming the Web of Data [Eckert, 2013].

According to Tom Heath - Not all Linked Data will be open, and not all Open Data will be linked (Design issues, 2009). So there is an important difference between these two terms; Linked Data is the data which is linked with other related datasets but is not open to reuse whereas, the Linked Open Data is the data which is linked as well as published under an Open License to reuse (Bhushan, 2014). From figure 2 it's possible to see the graph of Linked Open Data as of April 2014 (Schmachtenberg et al. 2014).

3.1 RDF

The RDF is a standard defined by the World Wide Web Consortium (W3C) for making statements to describe information resources. The purpose is to enable applications to process Web content in a standard way that is machine-readable, therefore simplifying the operation at Web scale. This technology is an essential foundation for the development of the Semantic Web (Hayes, 2004). In RDF, data is divided into individual statements about resources (Eckert, 2013). Here the word "resource" is treated as a synonym of entity, i.e., as a generic term for anything in any domain. Resources are described by RDF statements, also known as RDF triples which are composed by a subject, a predicate and an object. The collection of such triples describing resources is called the RDF graph. Intuitively, graphs can be viewed considering statements as subjects and objects as nodes that are connected by lines labeled by the declaration of the predicate (Hayes, 2004). RDF Graph can be described by the Figure 1.

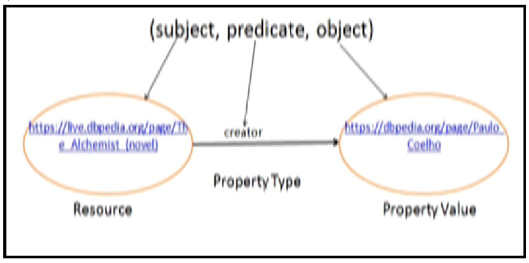
(subject, predicate, object)
https://live.dbpedia.org/page/The_Alchemist_(novel)
creator
https://dbpedia.org/page/Paulo_Coelho
Property Type
Resource
Property Value

Figure 1: Graphical Representation of a RDF Statement

The collection of RDF graphs is called RDF dataset and is used to organize collections of RDF graphs. Furthermore, URIs are the basis mechanism to identify subjects, predicates and objects in RDF statements, due to its generic nature. The subject of a triple is the URI identifying the described resource; the object can either be a simple literal value or a URI of another resource that is somehow related to the subject; the predicate indicates what kind of relation exists between a subject and an object [Heath, 2011]. In order to represent RDF statements in a machine-processable way, RDF defines a specific Extensible Markup Language (XML), referred to as RDF/XML.

4. The Idea of Linked Open Metadata: A Proposal

Linked Open data by definition emphasizes on the fact of describing methods of exposing and connecting data on the web from different sources (www.dbpedia.org) available on the Linked Open Data Cloud (LOD Cloud). The figure 2 below depicts this LOD cloud as of April 2014 (Schmachtenberg et al. 2014).

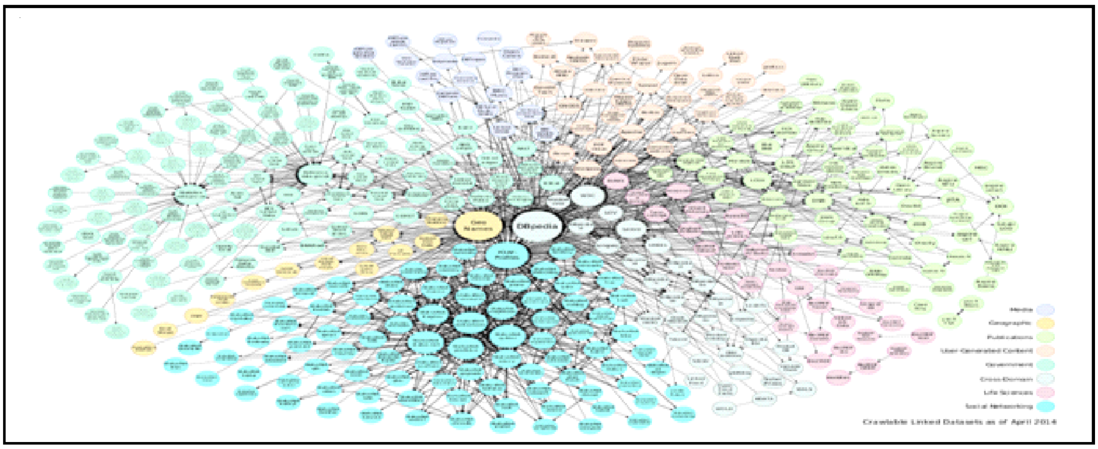

Figure 2: Linked Open Data Cloud Diagram

Currently the web uses hypertext links that allow people to move from one document to another (www.wikipedia.com). The shift is now to linking the open data to provide user with seamless experience for reading. Throughout the ages library catalogues or metadata of the past have created a vast pool of information about information. In many cases it served as only traceable historical information about some cultural object of its time.

The aim is to extend the idea of linked metadata, if practically implemented can open up a new vista facilitating not only in information retrieval but also knowledge assimilation process and new researches of the cultural preserves of the past.

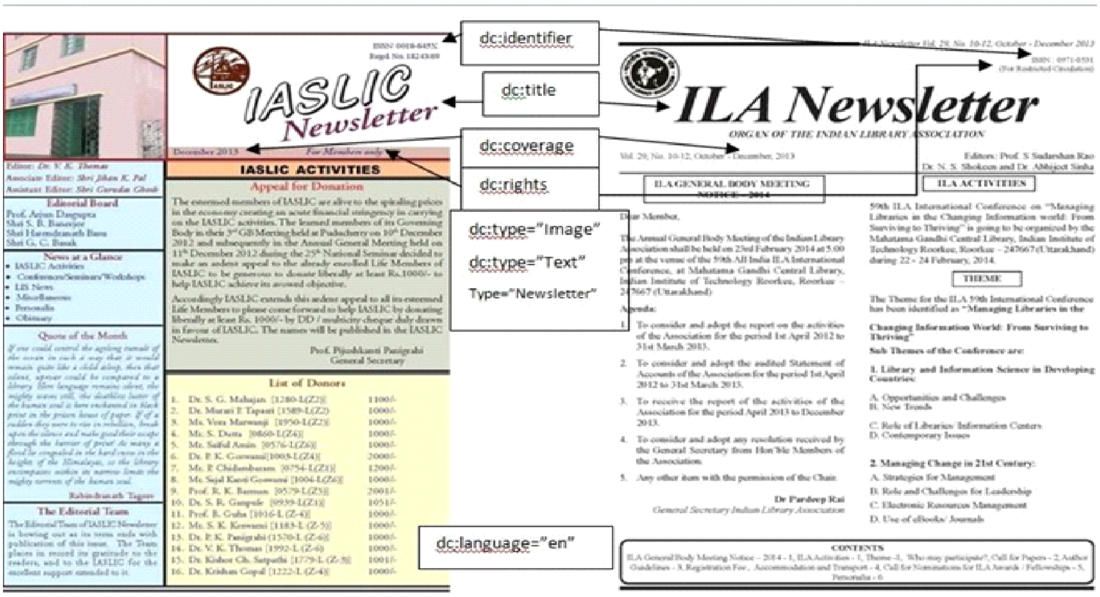

Figure 3 : Documents mapped using Standard Metadata Elements

In this Figure 3, we have considered Newsletters from two renowned Organizations in the field of Library and Information Science i.e. Indian Association of Special Libraries and Information Centers (IASLIC) and Indian library association (ILA) for finding suitable metadata elements to describe both the documents. But using elements from Dublin core or any metadata elements for describing content only leads to retrieval of that particular document but if we wish to provide our user/searcher with all the related information or contents instead of a particular item in context, we need to extend this present set of metadata elements with semantics in order to extract other linked up contents. This will help in filling the gap between actual need of information and query term used to search it, thus patrons may also be able to discover and satisfy their latent need.

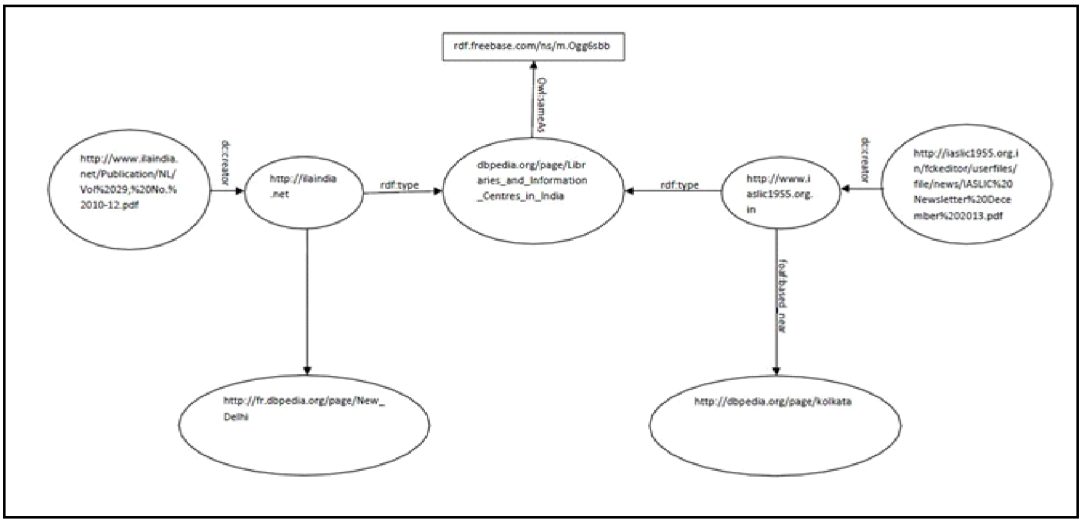
rdf.freebase.com/ns/m.Ogg6sbb
Owl:sameAs
http://www.ilaindia.net/Publication/NL/Vol%2029,%20No.%2010-12.pdf
dc:creator
http://ilaindia.net
rdf:type
dbpedia.org/page/Libraries_and_Information_Centres_in_India
rdf:type
http://www.iaslic1955.org.in
dc:creator
http://iaslic1955.org.in/fckeditor/userfiles/file/news/IASLIC%20Newsletter%20December%202013.pdf
foaf:based_near
http://fr.dbpedia.org/page/New_Delhi
http://dbpedia.org/page/kolkata

Figure 4: RDF Graph linking two documents with Metadata Elements

To put it into the perspective, we may start by using Ontology Web Language (OWL) vocabulary set which is an ontology language developed by the W3C for the Semantic Web which was designed for use by applications that need to process the content of information instead of just presenting information to humans. It facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDFS is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalization-hierarchies of such properties and classes) by providing additional vocabulary along with a formal semantics (McGuinness, 2004). OWL is a vocabulary extension of RDF that enables the definition of domain ontologies and sharing of domain vocabularies. It is modeled through an object-oriented approach and the structure of a domain is described in terms of classes and properties [Wang et al., 2004]. Now picking up 'sameAs' element from therein. (as shown in Figure 4). Thereafter 'creator' set in previous IASLIC example can be linked with other but semantically related set found elsewhere. Consequently, we may found library associations in any geographically defined space are grouped together (by again applying elements like 'basednear' from the FOAF vocabulary set). Then it is also possible to group together publications/proceedings from these organizations in one place.

This RDF graph can be further extended with Dublin core elements on both sides with details present in Figure 3.

5. Distant Reading Purpose

Distant reading is a concept which attempts to redefine old form of leisurely reading practice in a way which may revolutionise the whole reading culture as necessitated by the information deluge. The idea is based on processing content in its various facets for example any text in literary sense can at least generate subjects, themes, persons, places etc. or also some information like publication date, place, author, title. Therefore, involving in such practice may open up large number of textual items without actually engaging reader in the reading of the whole text. The data that could be extracted by this model of mining essentially make metadata of the text. (www.dh101.humanities.ucla.edu)

By this very description we can find the conception of distant reading comes close to the general idea of metadata. There are large corpuses of historical documents, volumes of literary texts, manuscripts and other cultural outputs of ancient time that can be found in the knowledge and information centers of the world. The sheer volumes make it almost humanly impossible to go through all these. But at the same time this is very crucial for the furtherance of researches on cultural studies and its understanding. In this context this model of creating a metadata map for the processing and analyzing of these texts are very useful.

It's possible to apply principles like P(property),M(matter),E(entity),S(space),T(time) of S.R. Rangnathan or even new D(domain), E(entity),P(property),A(attributes) version for basic classification and structurisation of the thought contents embedded in these works. Therefore a map may emerge in due course which relate all the vital factors of the works and looking at it can give quick understanding of the cultural works which is under lens. The same logic, if translated in digital domain can automate whole process, thus saving lots of time and money.

6. Vernacular Metadata

Along with discharging its traditional role, metadata in vernacular can create wide access of knowledge among common people. And if it's combined with the software which can automatically generate metadata like date and time as in images captured in digital camera, it may also create new knowledge and fulfill the aspiration of community and go long way to serve the cause of the society.

7. Metadata as New Knowledge

As it's evident from many previous instances, for many cultural objects, metadata acts as sole description or knowledge of that object that is available as no other description is to be found for that object anywhere. For example many photographs that are found in the book called 'People of India', eight volume study compiled by John Forbs Watson and John William Kaye is woefully devoid of any pinpointed and detailed description of the subjects in the photographs as it was the 'common sense' that informs British colonizers only limited and remote point of view to their Indian subjects from some distance and readily made them subject for stereotyping and essentialism along their trade, caste or class lines. In these cases metadata provide, albeit limited but vital and only source of firsthand knowledge about these images.

8. Conclusion

To conclude it can be said that metadata schema of the future must accommodate all the cultures and variety that emanates from it. If we may historicize the cataloguing and later metadata schema development process, we will see the evolution from AACR to Dublin Core metadata schema is a progress in right direction. AACR model of cataloguing practice was, as its known, devoid of local variation and uphold dominatingly the cultural reality that existed in Anglo American geo-political sphere. The libraries of global south which is also the theater of great diversity are also rich repository of metadata harvested over many years. If these vast resources can be digitized and linked in an open manner it will create new knowledge as many images, sketches, recordings of oral tradition, masks and other handicrafts kept in these libraries are variously described in its basic metadata form alone.

Therefore, liberating this knowledge from deep and dark forgotten layers of library shelves can unleash indigenous spirit from many untold stories hitherto unknown. Further researches may be undertaken in order to scientifically elaborate basic model for creating interlinked and open metadata map for cultural objects in a way so that all the facets of any particular object gets reflected through these metadata.

There remain however, some problems which require further study to go deep into it. The phenomena of social networking sites which have witnessed huge information creation and sharing activities in recent years necessitated metadata creation of these cultural outputs that found its way into web. The smart phones are aiding these phenomena with the advent of Selfie culture. Selfies have now caught popular imaginations with many such phonographs are captured with places of historical importance adorning its backdrop. These images captured and subsequently put on some social networking sites soon find its way towards oblivion. If we can use smart phones with its smart metadata ingesting system, these images may add very interesting twist to the reading of popular and micro history.

References

1. BERNERS-LEE, T. Linked Data. Available at http://www.w3.org/DesignIssues/LinkedData.html. (Accessed on19/10/2014).
2. BHUSHAN, V (2014). Linked Data: Emblematic applications on Legacy Data in Libraries. In: Proceedings of the 17th National Convention on Knowledge, Library and Information Networking (NACLIN 2014), India (pp. 8-20).
3. BIZER, C, HEATH, T and BERNERS-LEE, T.(Mar'2009) Linked Data: The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3),1–22.
4. "Design issues." Linked Data. Available at http://www.w3.org/DesignIssues/LinkedData.html. (Accessed on 18/06/2009)
5. DUNSIRE , G. and WILLER, M. (2011) Standard library metadata models and structures for the semantic web, Library Hi Tech News . 28(3),1-12 available at http://dx.doi.org/10.1108/07419051111145118 (Accessed on 27/09/2014).
6. ECKERT, K.(2013) . Provenance and annotations for linked data.

7. EL-SHARBINI, M. (2001). Metadata and future of Cataloging, library review 50, 116-27 available at http://www.emrald-library.com/ft (Accessed on 27/09/2014)

8. HAYES, J. A. (2004). Graph model for RDF.

9. HEATH, T and BIZER, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool,

10. HEERY, R and WAGNER, H. (2002). A metadata registry for the Semantic Web, D-Lib Magazine .8(5) Available at http://search.proquest.com/docview/57559910?accountid=27563 (Accessed on 27/09/2014)

11. HAYNES , D.(2009). Metadata for Information Management and Retrieval. Michigan: Facet.

12. IASLIC.(2013). IASLIC Newsletter available at http://iaslic1955.org.in/fckeditor/userfiles/file/news/IASLIC%20Newsletter%20December%20201.pdf (Accessed on 27/09/2014).

13. ILA. (2013). ILA Newsletter available at http://www.ilaindia.net/Publication/NL/Vol%2029,%20No.%2010-12.pdf (Accessed on 27/09/2014)

14. KOUTSOMITROPOULOS, D. A. et al. (2010). The Use of Metadata for Educational Resources in Digital Repositories: Practices and Perspectives, D-Lib Magazine .16,1-2. Available at http://search.proquest.com/docview/742896565?accountid=27563 (Accessed on 27/09/2014).

15. MCGUINNESS, L.D, & HARMELEN, V. F.(2014). OWL Web Ontology Language. Available at http://www.w3.org/TR/2004/REC-owl-features-20040210/ (Accessed on 28/09/2014).

16. MESCHINI, F. (2005). Topic maps: How I learned to stop worrying and love metadata, Bollettino AIB .45(1), 59-73, available at , <http://search.proquest.com/docview/57669197?accountid=27563> (Accessed on 27/09/2014).

17. MESSMER, G.(2013). Linking library metadata to the web: The german experiences, Italian Journal of Library and Information Science 4(1), 391-401, Available at http://dx.doi.org/10.4403/jlis.it-5507 (Accessed on 27/09/2014).

18. NIEDEREE, C. (2013).Metadata as modules of the semantic web, Zeitschrift Fur Bibliothekswesen Und Bibliographie. 50(4), 193-198, Available at http://search.proquest.com/docview/57599057?accountid=27563 (Accessed on 27/09/2014).

19. PEPONAKIS, M.(2013). Libraries' metadata as data in the era of the semantic web: Modeling a repository of master theses and PhD dissertations for the web of data, Journal of Library Metadata .13(4), 330-348, Available at http://dx.doi.org/10.1080/19386389.2013.846618 (Accessed on 27/09/2014).

20. TILLETT, B. B.(2003). AACR2 and metadata: Library opportunities in the global semantic web, Cataloging and Classification Quarterly .36(3),101-119, Available ar http://search.proquest.com/docview/57544931?accountid=27563 (Accessed on 27/09/2014).

21. TANIGUCHI, S.(2010). Current Situation of Metadata: Major Topics, Dublin Core and Semantic Web, The Journal of Information Science and Technology Association .60(12).

22. http://dh101.humanities.ucla.edu/?page_id=62 (Accessed on 27/09/2014)

23. http://lod-cloud.net/(Accessed on 27/09/2014)

24. WANG, X.H.[et.al.].(2004). Ontology based context modeling and reasoning using owl. In Pervasive Computing and Communications Workshops. Proceedings of the Second IEEE Annual Conference on, pp.18– 22.

25. Wikipedia available at http://en.wikipedia.org/wiki/Linked_data (Accessed on 27/09/2014).

About Authors

Mr. Rajarshi Das, Student, MPhil, Library & Information science, University of Calcutta, Kolkata.
Email: rajrishid@gmail.com

Mr. Vikas Bhushan, Ph.D. Student, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore.
Email: vikas@drtc.isibang.ac.in

Ms. Sudipta Shee, Assistant Librarian, Budge Budge College, Kolkata.
Email: sudi.art09@gmail.com

Mr. Santu Mandal, Project Assistant, Indian Statistical Institute, Calcutta Library.
Email: santu.mandalcu@gmail.com