June 2021

Classifying the Data: A Comparative Analysis of Traditional Library Classification and Linked Data Classification Systems

Saima Hanif
saimasherali@yahoo.com

saima hanif
University of the Punjab Lahore Pakistan, nosheen.im@pu.edu.pk

Classifying the Data: A Comparative Analysis of Traditional Library Classification and Linked Data Classification Systems

Saima Hanif
PhD Scholar, Department of Information Management
University of the Punjab, Pakistan
Saimahanif2015@yahoo.com

Dr. Nosheen Fatima Warraich
Associate Professor, Department of Information Management
University of the Punjab, Pakistan
nosheen.im@pu.edu.pk

ABSTRACT

Background. With the evolution of information technology, library classification schemes have transformed to effectively manage the information in electronic environment. The availability of library data on the web makes it challenging to devise a classification scheme that meet the need of Linked Data (LD) technologies. Objectives. This study aims to survey the library classification system along with automated classification system. It also highlights the links between the library classification systems and web of document classification system as a joint venture of LD.
Methods. For achieving the objectives of the study, available literature related to the traditional classification schemes, automated classification schemes and Linked Data classification systems were consulted. Different classification formats at different ages, Components of traditional classification, and components of Linked data RDF triples are described through figures. Comparison for the principles of library classification and linked Data and the types of classification are given through tables. Results. The results of this study show that Linked Data classification methods such as subject, predicate and object have the foundations on the traditional and machine readable classification systems. It is found that LD technologies for linking and sharing structured data on the web like, Uniform Resource identifier (URI) and Resource Description Framework (RDF) are based upon previous classification schemes.
Contributions. This study provides a precise picture of renowned traditional, online and LD classification schemes. This will be helpful to develop new RDF triple based ontologies for library LD organization.

Key Words: Linked Data, Classification, Data classification, Resource Description Framework

INTRODUCTION

Classification has been part of human life since the beginning of time. In the early ages, it was accomplished in different ways and consequently, different classification schemes emerged in different epochs.

There was an increased use for the classification schemes in the libraries; therefore, a good variety for classification schemes was also there. Moreover, with the passage of time Universal, National, Subject specific and Indigenous schemes were also developed. All these schemes were developed for the organization of knowledge.

Now we are entering in such an era, where latest technologies are growing day by day. The Linked Data technology is one of them. This technology is also classifying the knowledge and interestingly having the same bases of electronic and manual classification schemes.

The resemblance of the basic principles of traditional and Linked Data classification is given in this article. The precise picture of different classification schemes will be helpful for the development of new RDF triple based ontologies.

CLASSIFICATION SYSTEMS

Related literature in all forms was consulted and some categories were also developed for the literature.

Library Classification

Classification is a first step to organize and arrange any kind of objects. In libraries, it is related to books, subjects, documents, shelves and records. In everyday life, it is related to all forms of life ranging from our homes to shopping malls. The huge online and digital born resources are a challenge to classification schemes. These resources are available on the World Wide Web in structured and unstructured forms.

The traditional classification schemes were specifically developed and designed for the organization of print records in a physical environment. Though for digital documents, there is no need of physical aspects and environments. As said by (Broughton, 2006) ‘There is no shelf.’ The extent of online available documents advocates the need for an automated approach.

Conversion from Historical Classification Formats to Automatic Classifying Systems

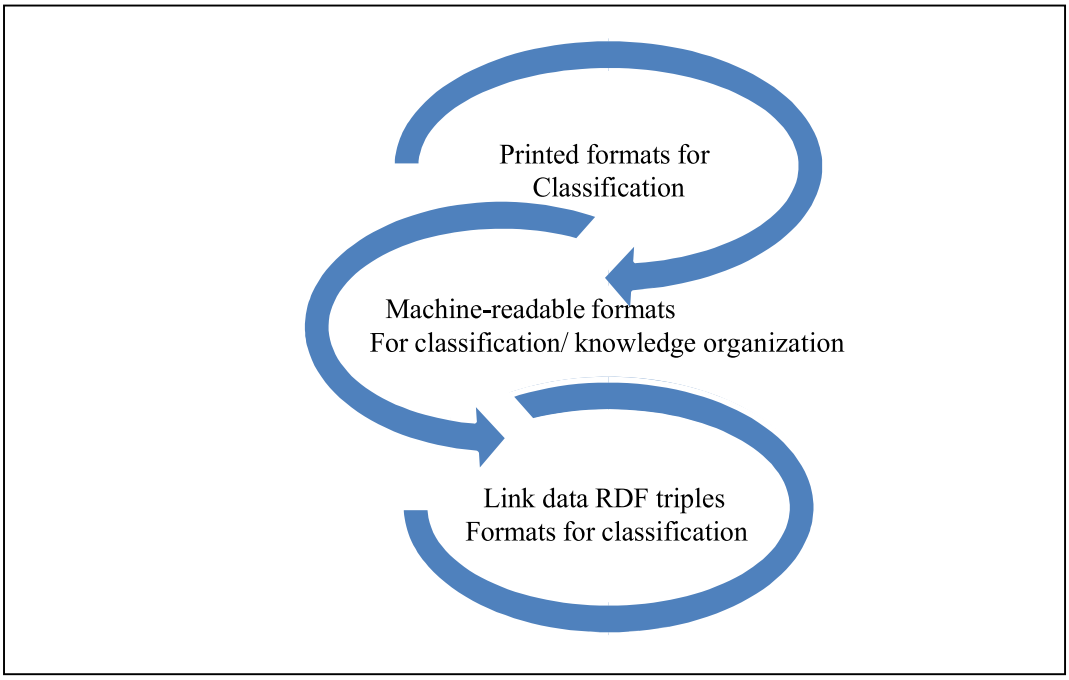
The interesting point is that, classification schemes have been developing and growing with the development and growth of libraries. In the past, these schemes were available in printed forms, then the need for machine-readable format arose and now the Linked Data technology is automatically classifying the data on the web through various softwares.

System of Library Classifications

The Process of classification is related to the knowledge association and the way people recall, remember and comprehend their world (Satija, M.P, 2015) in our everyday life,

sometimes we are not able to distinguish very clearly in classification and categorization. However, in the field of Library and Information Science these differences are very clear. Categorization may be seen as formless; whereas classification can be seen as a well-structured organization of information resources on linear shelves (Taylor & Joudrey, 2009).

If we will start looking specifically in the context of library science classification there are schemes which came in the late 1800s and in the beginning of 1900s for the handling of the early stages of the printing revolution, for example, for organizing and retrieving the bibliographic material. Different classification styles at different ages can be well defined through the following model.



Printed Formats for Classification

Printed formats for library classification were introduced in 18th century. There are many classification schemes in different years introduced by different renowned individuals. Some examples are

- Dewey Decimal Classification Scheme (DDC) by Melvil Dewey
- Universal Decimal Classification Scheme (UDC) was also known as FID Federation for Information and Documentation famous as Universal Decimal Classification consortium now.
- Expansive Classification Scheme (EC) by C.A Cutter
- Library of Congress Classification (LCC) by Library of Congress
- Subject Classification (SC) introduced by J. D. Brown
- Bibliographic Classification (BC) introduced by H.E. Bliss

- Colon Classification (CC) by S. R. Ranganathan
- Bibliothecal Bibligraphical Klassificaton (BBK) By VINITI for Russians, widely used in all over Russia
- Rider's International classification (RIC) by I. dhahlberg
- Broad system of Ordering (BSO) introduced by Eric Coates

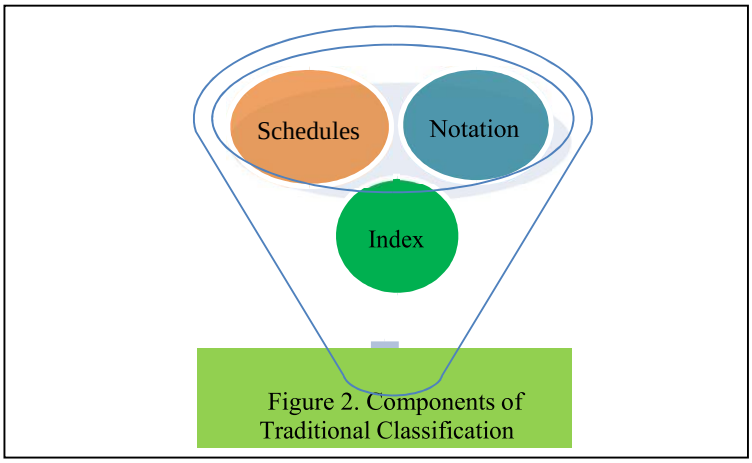


Figure 2. Components of Traditional Classification

Figure.2 is giving a picture about the classification components. The schedule is a list of classes, subdivision of classes and arranged in a logical way. Notation is providing codes using numbers/letters and has an easy to understand order. This order guides about the arrangements of subjects and documents and the Index is an alphabetical list for searching the terms within schedules.

Classification of libraries can be general or specific. General classification is a wide ranging taxonomy covers all subjects in the creation of knowledge. Though, special classification is a particular taxonomy covers a narrower range of topics. It may include the specific services. Moreover, it covers government reports, music, maps, etc.

Koch al. (1997) has stretched and systemizes the different types of classification as follows:

Table 1. Types of Classification

Universal Schemes	National Schemes	Specific subjects Schemes	Indigenous Schemes
Classifying the entire universe and for the use of anyone	universal in subject coverage but are used in a single country	Developed for the special subject community	Reader interest classification

Classification of Knowledge Organization

The contextual change in the library classification is not new. In 1933 Henry E. Bliss wrote a

book “Organization of knowledge in Libraries.” The same term Knowledge Organization is used today for document descriptions, indexing and classification but not only for the libraries but also for archives and databases for other similar types of institutions and environments. Then in 2000, Hodge provided a taxonomy and divided the knowledge organization into three categories

1. Tem lists; A list of terms providing definitions
2. Categories and Classification; construction of subject sets 3. List of Relationship; connections between terms and concepts in 2002, Linda Hill, et al, revised this list.

1. Classification and categorization; classification and categorization schemes
2. Metadata like models; Directories, Gazetteers and Geo-spatial dictionaries of places.
3. Relationship models; Ontologies, Semantic networks and Thesauri
4. Term list; Authority files, Dictionaries and Glossaries
   So from 1933 to 2002 the term knowledge organization has also been changed and revised in their context.

Machine-Readable Formats for Classifications

The creations of digital born content and digital resources have made it difficult to organize and classify knowledge manually. So, there was a need to re-examine methods of classification. Now the framework in which knowledge organization tasks have to be performed are also developed and expanded. Some of the Knowledge organizations tasks are being performed in a conventional way such as emails are classified in a spam and non-spam, classifying Web Pages and extracting metadata for textual e-resources.

Renowned classification systems have also changed and are viable in machine-readable databases e.g., Decimal Classification, Library of Congress Classification and Universal Decimal Classification are available in printed and machine-readable formats. These living classification schemes are converted from print to machine-readable forms and are available in both forms.

According to Slavic (2008) MARC-21 is having following functions.

- “Classification Search by notation Schedule display in different layouts Auto link tracking.
- Steering between Subjects areas, facets and tables.
- Chronological data tracing”

We can conclude that online classification system is not rationally and logically different from its print version. No doubt it has a variety of extra functions.

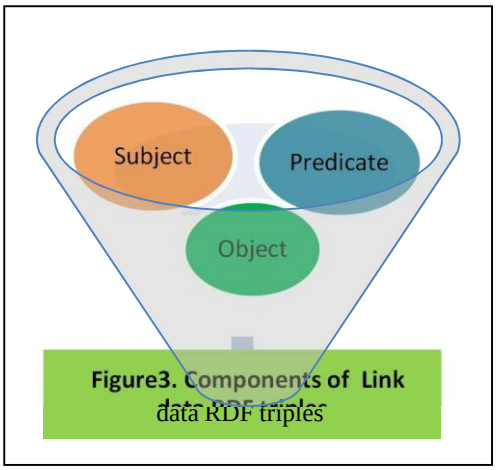
Linked Data Formats to Classify Web Documents

The Principles of both approaches have the resemblance. The Linked data follows the following four principles.

1. Using URIs.
2. Using HTTP URIs so that everyone can look up those names
3. Representation of all resources identified by those URLs
4. Makes sure that data contains links to other data allowing software agents to look up related information.

Latest Linked Data classification technologies like URI and RDF have a potential to generate the automatic metadata and semantic layer making it more precise. A current idea of document classification is using semantic rather than syntactic approach. These techniques have come forward by rapidly increasing projects of Semantic Web. Making all the data machine-readable is having resemblance with the Colon Classification Scheme of Ranganathan.

Linked data narrates the publishing methods of structured data to make it more useful and interlinked. On the web, there are many documents that can be read by semantic layer and humans where they function as an interaction between the readers and language of the document. With the addition of metadata to these documents, it becomes understandable for computers as well. For linked data, RDF (Resource Description Framework) triples and URIs (Uniform Resource Identifier) do this classification. The RDF provides a tripartite expression of the link between Subject, Predicate and Object.




```
1. @prefix dbpedia: <http://dbpedia.org/resource/>
2. @prefix dbpedia-owl: < http://dbpedia.org/Ontology/>
3. @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntaxns#>
4. @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
5. http://dbpedia.org/page/United_States
6. rdf:type dbpedia-owl : Country;
7. rdfs:label "USA"@en;
8. dbpedia-owl : capital dbpedia:Washington_D.C.;
9. dbpedia-owl:areaSqMi "3794101"^^xsd:integer.
```

Figure 4. Snapshot of RDF Graph Representing United States of America (From DBPedia.org)

As an example/ the subject at (line 5) ("http://dbpedia.org/page/United States") is stated (at line 6) to belong to the category "Country". Using the predicate "rdf: type". This subject is also given a tag ("rdfs: label" via predicate; on 7 line) of "USA". The diagram states that "Washington D.C." is its capital (via the predicate "dbpedia- owl: capital"; on 8 number line) and the total area square miles is "3794101" (via the predicate "dbpedia-owl: areaSqMi; on 9 number line). In the same way Thinkpedia is also classifying different relevant classes.

The URIs provides a fixed description of the subject and object and it is the exposure of Linked data. It can easily create link with some other data. This interaction of data sets develops the hyper data (Idehen, 2009): a direct link to hypertext.

A Comparison between principles of traditional library classification and Linked data classification is given below.

Table 2. Comparison for the principles of Library Classification and Linked Data

Principles of Library Classification	Principles of Linked Data
Clear notes, examples and instructions are provided	• Using URIs
Planned for wide representation	Using HTTP URIs so that people can find those names.
Coverage of all fields of knowledge	Representation of the resources identified by those specific names
Available in interoperable version for various sizes of libraries	Makes sure that data contains links to other data allowing software agents to look up related information.

CONCLUSION

Organizing information without any classification system is not easy. This is the reason that many classification schemes are available and these classification schemes were getting mature day by day. The Linked Data classification methods are having the same foundations of traditional and machine-readable classifications systems. A comparison for the principles of library classification and Linked Data classification given in Table.2 is showing this resemblance. The E-formats of classification schemes were accepted widely and quickly because that was the need of the time. The libraries will also accept the Linked Data classifying methods, as this is also an important development of the technological age. The librarians and information professionals are taking initiatives for making their systems more advance and linked globally.

REFERENCES

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings: New Information Perspectives, 58 (1/2), 49-72.

Fripp, D. (2010, November). Using linked data to classify web documents. In Aslib Proceedings (Vol. 62, No. 6, pp. 585-595). Emerald Group Publishing Limited.

Hodge, G. Systems of knowledge organization for digital libraries: Beyond traditional authority files. The Council on Library and Information Resources, Washington DC, 2000. http://www.clir.org/pubs/reports/pub91/contents.html/.

Idehen, K. (2009), "Five very important things to note about HTTP based linked data", Kingsley Idehen's Blog Data Space, 2010, available at: www.openlinksw.com/dataspace/ kidehen@openlinksw.com/weblog/kidehen@openlinksw.com%27s%20BLOG%20% 5B127%5D/1591

Koch, T., Brümmer, A., Hiom, D., Peereboom, M., Poulter, A., & Worsfold, E. (1997). DESIRE: Project Deliverable. Retrieved from http://cordis.europa.eu/cybercafe/frames/25.htm

Satija, M. P., & Martínez-Ávila, D. (2015). Features, Functions and Components of a Library Classification System in the LIS tradition for the e-Environment. Journal of Information Science Theory and Practice, 3(4), 62-77

Semantic Proxy – part of the Thomson Reuters Calais initiative, currently at Beta testing stage, available at: www.semanticproxy.com

Slavic, Aida. (2008). Faceted classification: Management and use. Axiomathes, 18(2), DOI: 10.1007/s10516- 007-9030-z

Smiraglia, Richard P. Universes, dimensions, domains, intensions and extensions: Knowledge organization for the 21st century. In Categories, contexts and relations in knowledge organization, edited by A. Neelameghan and K.S. Raghavan. Ergon Verlag, Wurzburg, 2012.

Prote´ge´ – Ontology builder used to expose the main class fields and class relations in the DBpedia database, available at: http://protege.stanford.edu/

Raghavan, K. S., & Krishnamurthy, M. (2013). An overview of classification technologies. DESIDOC Journal of Library & Information

Technology, 33(4). Taylor, A.G., & Joudrey, D.N. (2009). The organization of information (3rd ed.). Westport, CT: Libraries Unlimited.

The Organization Of Knowledge In Libraries by Henry E. Bliss Published in 1933.https://archive.org/details/organizationofkn007596mbp