ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
Journal of Information and Knowledge, Vol 60(3), June 2023, p.143-149
DOI: 10.17821/srels/2023/v60i3/171025

Machine-Generated Colon Class Numbers: Automatic Classification of Indian Literary Works in the Wikidata Environment

Debkanta Halder and Meghna Biswas*

Department of Library and Information Science, University of Kalyani, Kalyani – 741235, Nadia, India;
for3education@gmail.com, meghnabsws@gmail.com

Abstract

Wikidata is emerging rapidly as an omnipresent information space. The main objective of Wikidata is to create a structure to represent all human knowledge in many different languages. It is a free, Linked Open Dataset (LOD) that can interact with humans and machines and act as a bridge between them. Wikidata serves as a central repository for structured data of all other Wiki-based applications (like Wikipedia, Wikibooks, and Wikinews), integrating them into a coherent unit. It is a multilingual, collaboratively editable central repository of structured and linked open data. Interestingly, Wikidata is based on a facet-based linked open data approach like, the theory of facet analysis proposed by Ranganathan. The literary works available in Wikidata are linked with the author of the work, the language of the work, and other important attributes like genre, publication year, and so on. The author of the work has properties like, date of birth, occupation, native language, writing language, and so on. These faceted linked data elements of literary works can be joined together based on the rule base of the Colon Classification (6th edition). The synthesis of the Colon class number is therefore amenable to automation in the Wikidata environment by using a set of suitable JavaScripts (these JavaScripts are called gadgets in the Wikidata environment). One such gadget is CCLitBox, developed by Stefano Bargioni (a librarian) and Carlo Bianchini (a teacher in LIS). This research study explores the applications of CCLitBox in generating CC 6th edition-based class numbers for Indian literary works.

Keywords: Automated Classification, Colon Classification, Faceted Classification, Linked Open Data (LOD), Indian Literature, Wikidata

1. Introduction

Colon Classification (CC) is a systematic scheme of library classification used in many libraries in India and a few libraries in abroad. It was devised by the late Dr. S. R. Ranganathan, who felt a need for a scheme of library classification capable of coping with the multidimensional and dynamic growth of the universe of subjects. Colon classification proceeds differently: Instead of enumerating all possible subjects and their subdivisions, it analyzes the subject into its various components and categorizes these into five fundamental categories *viz*,. Personality, Matter, Energy, Space, and Time. To connect or synthesize the various components of a subject, different indicator digits have been provided. To build a class number, one has to analyze and identify possible isolates belonging to different fundamental categories, which are then put together using appropriate connecting symbols. Colon classification involves analysis and synthesis, which is why it is known as an analytico-synthetic scheme of classification. The number-building process makes the scheme somewhat complicated and difficult to work with, but once understood and followed, it works efficiently and effectively. Colon classification is a general scheme that aims to classify all kinds of documents—books, periodicals, reports, pamphlets, microforms, and electronic media — in all types of libraries by subject. CC is a landmark in modern classification thought and

*Author for correspondence

has greatly influenced modern classification research and developments.

Linked Open Data (LOD) is a web-based technology that enables seamless, machine-supported integration, interplay, and augmentation of all kinds of knowledge into what has been labeled as a huge knowledge graph. Despite decades of web technology and, more recently, the LOD approach, the task of fully exploiting these new technologies in the public domain is only beginning. One specific challenge is to transfer techniques developed in the pre-web era to order our knowledge into the realm of LOD. Wikidata is a dataset that is independent of but highly relevant to the library world. Moreover, due to its openness, anyone can use it to create, publish, use, and reuse data in an automated and programmable way, in a cooperative environment, and with a bottom-up approach. For these reasons, Wikidata has proved to be an appropriate platform for the development of an automated classification process using LOD and faceted classification. This research study utilizes the Wikidata gadget, CCLitBox, for the automated classification of literary authors and works by faceted classification using LOD. The tool reproduces the classification algorithm of class O Literature of the Colon Classification and uses freely available data elements in Wikidata to create CC-based class numbers. CCLitBox is a free tool that enables any user to classify literary authors and their works. It is easily accessible to everyone, uses LOD from Wikidata, and missing data for classification can be added freely if necessary. All data needed for CC author and work class numbers is either already available in a Wikidata item or can be added, like: Language; literary form; author's year of birth; title and date of creation or publication of the work. This study seeks to test and examine how the automated generation of CC-based class numbers is applicable to Indian languages in the Wikidata environment.

2. Review of Related Literature

The paper "Automated Classification Using Linked Open Data. A Case Study on Faceted Classification and Wikidata" by Bianchini and Bargioni (2021) has been the basis of our study. They have mentioned that it is possible to create class numbers and call numbers for literary authors, works, expressions, manifestations, and items using an automated classification process based on a faceted classification system like CC and on data that is typically available as consistent and complete LOD. Moreover, program *configuration* was a requirement for the construction of an automated classification project. In reality, a Wikidata gadget was designed for the study to synthesize and create CC author and work class numbers. The authors thought that the CCLitBox gadget demonstrates that the combination of a faceted classification scheme and a linked open data set that records some characteristics of the entity to be classified in the form of properties or attributes can be transformed into a useful algorithm for an automated classification tool. Just like they used CC as their basis for the project, other classification tools for other disciplines could also be used to generate automatic class numbers. Rao and Gupta (2011) mention that classification is a primary method of information organization, as it meets the most important objective of information organization, which is to bring essentially similar information together and to differentiate what is not exactly alike. Classifications may be used for categorizing and representing entities (or their labels), concepts, and their relationships. The article discusses different terminologies, concepts, and definitions used in the applied classification of web knowledge management and concludes by looking at some ideas for the future of classification on the Internet. Markey (2009), in his paper, examines the forty-year history of online use of classification systems. Enhancing subject access was the rationale for obtaining support to conduct research in classification online and for incorporating classification into online systems. Cataloguers have been the beneficiaries of most of the advances in classification online, and operational online systems are now able to assist them in class number assignment and shelf listing. To this day, the only way in which most end users experience classification online is through their online catalogue's shelf list browsing capability. However, the reasons why online classification never caught on as an end user's tool in online systems are not clear. Both the information industry and the library and information science communities missed the opportunity to lead the charge in the organization of Internet resources; however, OCLC, the publisher of the Dewey Decimal Classification, has made substantial improvements to the scheme that have increased its versatility for organizing Internet resources. Because mass digitization projects such as Google Print may solve the problem of subject access, the author makes recommendations for classification online to solve these vexing problems for end users: staging of access, retrieving the best material in response to user queries, and automatic approaches to finding additional relevant information for an ongoing search. Chudamani (2004), in her paper, pointed out that there has been a lot of research carried out about the models of classification suitable for libraries in a digital environment. UDC, CC, DDC, and LCC are used in libraries all over the world. People are using existing classification models and automating them for use in the digital environment. Though this is a good step, one has to look at the operational requirements of the digital environment. Keeping this fact in view, this paper analyzes the requirements of the digital environment and points out the adaptations required for the traditional schemes to suit the digital environment. Panigrahi, Prasad, and Basu (2003) in their research, made numerous attempts to design powerful automatic classification systems, but those could not bring any well-accepted results. The main problem was the unsuccessful automatic analysis of the titles of documents and finding out subject propositions. Because it is a purely mental process, classification demands human intelligence for analyzing the title to find out its basic subject and other facets, if any, along with its category and also synthesizing those facets according to syntactic principles to construct a classification number. In other words, the document title, which is in natural language, is analyzed carefully to identify relevant words (i.e., subject propositions), and those are synthesized using the classifier's expertise to build the classification number. The emergence of Artificial Intelligence (AI) could bring a solution to this problem. The use of Natural Language Processing (NLP) techniques would help in the automatic analysis of titles, and an expert system could be developed to work exactly in the same way as a classifier does to build classification numbers, guided by canons, principles, and postulates. This paper is based on the research work along this line of thinking. The semantic and syntactic components of an AI-based system for automatic classification are described.

While automated methods for information organization, i.e., classification, have been around for several decades, the exponential growth of the World Wide Web has put them at the forefront of research in several different communities: machine learning (artificial intelligence), document clustering (information retrieval), and weighted string matching against a controlled vocabulary (library and information science). Based on the latter approach, NetLab at Lund University Library explored automated classification based on UDC in Nordic WAIS and WWW as early as 1992 and continued research during the 1990s, testing automatic classification on engineering index classification and DDC. The similarities and differences among the three approaches will be discussed here, and problems with automated classification will be recognized. There is, however, a major issue of evaluation and comparison, largely due to the challenge of identifying the content of documents, which is related to the quality of indexing.

Automated subject classification has been a challenging research issue for several decades. The interest has grown rapidly through the emergence of the World Wide Web (WWW)) and related digital information services with a large number of documents, where the high costs of manual subject classification are a major hindrance. Currently, there are three distinguishable approaches to automated subject classification of text, each taken by a different research community: *Text categorization*, *document clustering*, and *document classification*. They differ in a number of aspects, such as scientific tradition, methodology (including document pre-processing and indexing), test collections, characteristics of categories, evaluation methods, and application. However, all of them deal with the same problem, and similarities among them exist; for example, the selection of the most relevant terms during document pre-processing is common to all the approaches, as is the utilization of specific document characteristics. This leads one to assume that idea exchange and cooperation among the three communities would be beneficial. This goal is to examine the simple bibliometric methods that can be used to investigate to what degree the three communities utilize each other's ideas, methods, and findings.

S. R. Ranganathan developed an elaborate scientific theory of classification after a careful study of the Universe of Subjects and the ways of formation of new subjects. Based on this, he introduced the concept of an analytical-synthetic classification system and devised Colon Classification (CC)) in 1933. These studies obviously have a lot to do with basic ideas and their relations. The parallels can be easily seen between the study of the universe of knowledge and that of Knowledge-Based Computer

Systems (KBCS),, otherwise known as expert systems in the field of *Artificial Intelligence* (AI)). Although both fields have grown independently, they have a lot to offer each other. The application of *natural language processing*, another important area of AI, could solve the problem of designing a powerful automatic classification system.

An AI-based classification system, named Visvamitra, was developed by A.R.D. Prasad for CC 7. The system consists of a number of knowledge bases to represent the schedule of common isolates and special isolates following the frame-based knowledge representation model. A huge lexicon is developed to support the natural language-based analysis of document titles. Cannons, principles, postulates, and different instructions mentioned within schedules are implemented to a large extent through an inference engine developed in Prolog.

3. Objectives

This paper aims to link two information spaces- Wikidata and faceted library classification- with the help of a gadget called CCLitBox. The framework of an automated classification system based on this approach and the availability of LOD creates a platform (CCLitBox) using which class numbers could be formed for literary works.

The objectives of this study are:

- To integrate two information spaces- Wikidata and faceted library classification- using the CCLitBox gadget
- To enable users (or readers) throughout the world to create CC class numbers for authors and their literary works
- To examine the usability of the CCLitBox tool for the classification of Indian literature (in Indian languages) with suitable modifications of the original gadget

4. Research Perspective

The structure of Wikidata can be categorized into three facets: item (any entity), property (attributes of an entity), and value (reference or literal). Each of these facets is being allocated with a unique, consistent identifier: a positive integer to be prefixed with the upper-case letter 'I' for the Item facet, 'P' for the Property facet, and values that are generally strings, numbers, or media files. S.R. Ranganathan's Colon Classification is essentially based on the same notion of the faceting. It shows us how class numbers could be constructed by using an analytico-synthetic or simply a faceted approach.

- To create the class number for the author, some properties must be included. They are:
- The entity must be a literary work in Wikidata
- The work must have two properties: language and publication date or year
- The author should be an instance of a human (Q5)
- An occupation (P106) must be associated with the author, which can be a poet, novelist, playwright, or writer

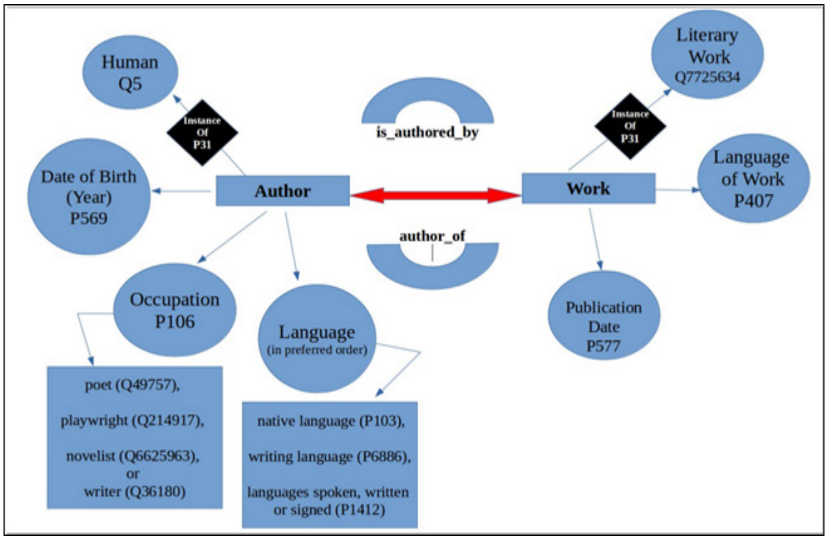
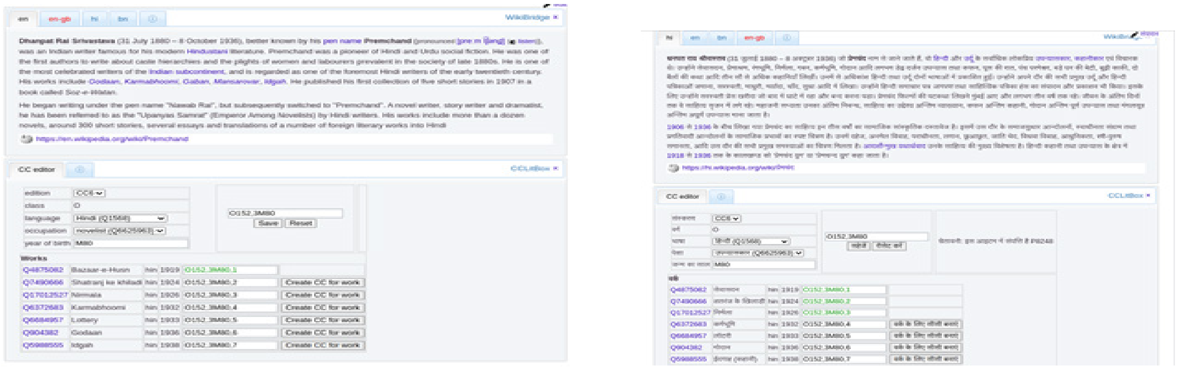


Figure 1. Semantic mapping of facets of literature class and literary works in Wikidata.



en | en-gb | hi | bn | ⓘ — WikiBridge ×

Dhanpat Rai Srivastava (31 July 1880 – 8 October 1936), better known by his pen name **Premchand** (pronounced [pre:m tʃənd̪] listen), was an Indian writer famous for his modern Hindustani literature. Premchand was a pioneer of Hindi and Urdu social fiction. He was one of the first authors to write about caste hierarchies and the plights of women and labourers prevalent in the society of late 1880s. He is one of the most celebrated writers of the Indian subcontinent, and is regarded as one of the foremost Hindi writers of the early twentieth century. His works include Godaan, Karmabhoomi, Gaban, Mansarovar, Idgah. He published his first collection of five short stories in 1907 in a book called Soz-e-Watan.

He began writing under the pen name "Nawab Rai", but subsequently switched to "Premchand". A novel writer, story writer and dramatist, he has been referred to as the "Upanyas Samrat" (Emperor Among Novelists) by Hindi writers. His works include more than a dozen novels, around 300 short stories, several essays and translations of a number of foreign literary works into Hindi.

https://en.wikipedia.org/wiki/Premchand

CC editor | ⓘ — CCLitBox ×

edition	CC6
class	O
language	Hindi (Q1568)
occupation	novelist (Q6625963)
year of birth	M80

O152,3M80 — Save | Reset

Works

Q4875082	Bazaar-e-Husn	hin	1919	O152,3M80,1	
Q7490666	Shatranj ke khiladi	hin	1924	O152,3M80,2	Create CC for work
Q17012527	Nirmala	hin	1926	O152,3M80,3	Create CC for work
Q6372683	Karmabhoomi	hin	1932	O152,3M80,4	Create CC for work
Q6684957	Lottery	hin	1933	O152,3M80,5	Create CC for work
Q904382	Godaan	hin	1936	O152,3M80,6	Create CC for work
Q5988555	Idgah	hin	1938	O152,3M80,7	Create CC for work

hi | en | bn | en-gb | ⓘ — WikiBridge ×

[illegible]

https://hi.wikipedia.org/wiki/प्रेमचंद

CC editor | ⓘ — CCLitBox ×

संस्करण	CC6
वर्ग	O
भाषा	हिन्दी (Q1568)
पेशा	उपन्यासकार (Q6625963)
जन्म का साल	M80

O152,3M80 — सहेजें | रीसेट करें

[illegible]

कार्य

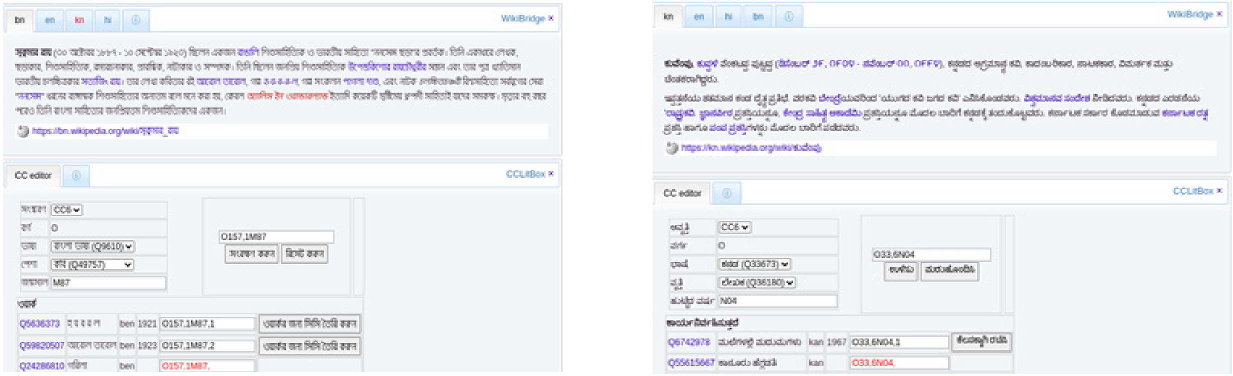
Q4875082	[illegible]	hin	1919	O152,3M80,1	
Q7490666	[illegible]	hin	1924	O152,3M80,2	
Q17012527	निर्मला	hin	1926	O152,3M80,3	
Q6372683	कर्मभूमि	hin	1932	O152,3M80,4	[illegible]
Q6684957	[illegible]	hin	1933	O152,3M80,5	[illegible]
Q904382	गोदान	hin	1936	O152,3M80,6	[illegible]
Q5988555	[illegible]	hin	1938	O152,3M80,7	[illegible]

Figure 2. CCLitBox interface in default mode (left) and in multilingual mode (right).

- A language must be associated with the author, which can be either the native language, the writing language, or languages spoken
- The year of birth (P569) must be assigned to that person

An illustration (Figure 1) shows the semantic mapping between facets of the literature class of the Colon Classification 6th edition and that of literary works in Wikidata.

Several steps were performed to ascertain the utility of Colon Classification in automated classification, and representation and formation of call numbers. Classification for literary works in CC can be reduced to an algorithm with the help of its classic device and facet formula (O [P_1], [P2] [P3], [P4]) of "Literature Class O". Unlike Dewey Decimal Classification (DDC) or Universal Decimal Classification (UDC), literature class in CC comprises of not more than 10 isolates, making it unique. Information and data collected from Ranganathan's Colon classification provided the basis to conduct this study. The value added to the application of CCLitBox as presented by us, however, focuses on how this gadget is working in the Indian multilingual environment. Our first step was the identification of Indian languages having ISO 639-3 codes, and if those language numbers were not found in Colon Classification they were constructed following the rules of Colon Classification 6th edition. As in the case of Maithili, Konkani, Dogri, and a few, we had to construct the numbers. Then there was the conversion of the CCLitBox gadget to support Indian languages by doing some changes in the script on the "*common.js*" *page* in the registered user's Wikidata platform. The next step was the translation of the gadget into 22 constitutionally recognized Indian languages using a modified JAVA script. After the gadget was made ready, then we performed testing and debugging the gadget with a sample set of literary works in those 22 constitutionally recognized languages. Although while performing these tests, the gadget was working perfectly, for some languages problems were faced.



bn | en | kn | hi | ⓘ — WikiBridge ×

[illegible]

https://bn.wikipedia.org/wiki/সুকুমার_রায়

CC editor | ⓘ — CCLitBox ×

সংস্করণ	CC6
বর্গ	O
ভাষা	বাংলা ভাষা (Q9610)
পেশা	কবি (Q49757)
জন্মসাল	M87

O157,1M87 — সংরক্ষণ করুন | রিসেট করুন

কার্য

Q5636373	[illegible]	ben	1921	O157,1M87,1	[illegible]
Q59820507	[illegible]	ben	1923	O157,1M87,2	[illegible]
Q24286810	[illegible]	ben		O157,1M87,	

kn | en | hi | bn | ⓘ — WikiBridge ×

[illegible]

https://kn.wikipedia.org/wiki/[illegible]

CC editor | ⓘ — CCLitBox ×

[illegible]	CC6
[illegible]	O
[illegible]	[illegible] (Q33673)
[illegible]	[illegible] (Q36180)
[illegible]	N04

O33,6N04 — [illegible] | [illegible]

[illegible]

Q6742978	[illegible]	kan	1967	O33,6N04,1	[illegible]
Q55615667	[illegible]	kan		O33,6N04,	

Figure 3. Automatic generation of CC class numbers in Bengali language (left) and in Kannada language (right).

5. Conclusion

With the advancement of knowledge and the amount of new information being created, these systems have rendered themselves not only invaluable but indispensable. Automated classification schemes have a major role to play in aiding information retrieval in the Wikidata environment, especially for providing item structures for information gateways on the Internet. Advantages of using automated classification schemes include improved subject classification facilities, potential multi-lingual access, and improved interoperability with other services.

In summay the following conclusions have emanated from the above study. In the present, Wikidata environment, CC is the most suitable scheme of classification because of its hospitality and the ability to include emerging new subjects from the ever-growing universe of subjects. The inclusion of new concepts can be made possible at any point in the scheme without disturbing the present order of concepts. We conclude that the CCLitbox gadget works, but it needs to improve consistency and efficiency. Natural language processing, along with the automatic generation of class numbers, is difficult, and without the intervention of any human indexer, it becomes even more difficult and complicated. Keeping aside the benefits of Colon classification, we faced some drawbacks too. In the scheme, the facet formula of literature class is not a unique identifier for authors and their works because if we follow the rules strictly, then poems and novels by the same author would be located in two different sections, thus creating confusion. Also, different authors having the same year of birth may result in the same class numbers, which again creates ambiguity. CC language schedules are mainly devoted to Indo-European, Semitic, and Dravidian languages (currently, 54 ready-made class numbers are available); therefore, to make the gadget compatible with other different languages of the world, ISO 639-3 codes may be brought into use. The major reasons for the limited use of the electronic version of the CC schemes are budgetary constraints, a lack of requisite infrastructure, and awareness about online classification systems. Automated classification schemes have a bright future in the Wikidata environment and therefore cannot be replaced, although with the current online classification tools available, these can be supported in a much mor e convenient and better manner.

6. Future Prospects

The CCLitBox gadget presently supports a single instance-based automatic building of Colon classification numbers. The large-scale utilization of the gadget requires a REST/API approach, based on which Wikidata item IDs can automatically create CC numbers in data-wrangling software like OpenRefine. Another major possibility may be the integration of the CCLitBox gadget with ISO 3166 (place codes) and ISO 639-3 (language codes). This would eventually lead to the fact that if at any time in the near future the 8th edition of CC is published, there will be no need for Language and Place Isolates, and the volume will eventually be smaller than the 6th and the 7th editions. Presently, there are tools available through which the entire catalog of a given library in MARC format can be added to Wikidata automatically. There is a future possibility that if bibliographic data sets of all Indian libraries, including the Indian National Bibliography, are made available on the Wikidata platform, then reconciliation of CC numbers would be possible in any local bibliographic dataset. This would immensely help libraries in India and elsewhere that use or wish to use Colon classification as their scheme of classification.

7. References

Bianchini, C. and Bargioni, S. (2021). Automated classification using linked open data: A case study on faceted classification and Wikidata. Cataloging and Classification Quarterly, *59*(8), 835-852. https://doi.org/10.1080/01639374.2021.1977447

Chudamani, K. S. (2004). Classification model for libraries in the digital environment. In 2nd International CALIBER; INFLIBNET; Ahmedabad. p. 487-490.

Markey, K. (2009). Forty years of classification online: Final chapter or future unlimited? Cataloguing and Classification Quarterly, *42*, 3, 1-63.

Panigrahi, P, Prasad, A. R. D. and Basu, A. (2003). NLP based automatic classification system for analytico-synthetic scheme. SRELS Journal of Information Management, *40*, 4, 289-312.

Ranganathan, S. R. (1963). Colon classification. Basic classification. reprinted with amendments. Bombay: Asia Publishing House. p. 450.

Rao, K. S. and Gupta, U. (2011). Applied classification in web knowledge management. SRELS Journal of Information Management, *48*, 1, 23-32.

Stefano Bargioni is the author of the scripts of the gadget, and they are available, in Wiki-data (https://www.wikidata.org/wiki/User:Bargioni/CC_Lit_Box.js and https://www.wikidata.org/wiki/User:Bargioni/CC_Lit_Box_conf.js) and are openly published in Zenodo (https://zenodo.org/record/5090640)

Wikibase. Welcome to the Wikibase project. Available at: https://wikiba.se/

Wikidata. Colon Classification (P8248). Available at: https://www.wikidata.org/wiki/Property:P8248