DESIDOC Journal of Library & Information Technology, Vol. 40, No. 5, Sept 2020, pp. 277-285, DOI : 10.14429/djlit.40.5.16034

Language Analysis in Library OPAC: Designing an Open Source Software based Framework for Bibliographic Records in Mainstream and Tribal Languages

Parthasarathi Mukhopadhyay* and Anirban Dutta

Department of Library and Information Science, University of Kalyani, Nadia - 741 235, India

**E-mail: psmukhopadhyay@gmail.com*

ABSTRACT

It reports the development of an enhanced library OPAC prototype through integration of language analysis tool and book reader in the retrieval interface. Language analysis or text analytics is considered as one of the components of language documentation and when integrated with library OPAC can extend supports to analyse corpus of the retrieved document in terms of word/phrase frequency, term circus, term links, term context etc through visual representation in a single-window along with the other datasets generally expected in a typical library OPAC. The open source software based integration mechanism is tested with English and Bengali as mainstream languages and a Unicode-compliant Indian official tribal language Santali (Ol Chiki script) as minority language.

Keywords: Language analysis; Text analytics; Enhanced OPAC, Book reader; Ol Chiki; Santali.

1. INTRODUCTION

As opined by Boerger, B.H., *et al.*[1], the general work-flow of language documentation includes recording, metadata maintenance, transcribing, annotation & analysis, translation, archiving and dissemination. In simple words, language documentation refers to documenting and describing verbal, written, and gesture (for sign language) forms of a language systematically in their suitable sociocultural context with emphasis on long-term archiving and access[2,3,4]. Library professionals can prove themselves helpful in many of these aspects like annotation, metadata encoding & standardisation, long-term archiving and effective dissemination. One of the major purposes of language documentation or documentary linguistics, apart from the activities as given, is to provide scope for linguistic analysis in any given language from the corpora of textual materials available in that language[5,6]. The advancement of the NLP tools and techniques, availability of the open source language analysis tools/services and increasing use of the open source software in library management are the factors that provide an opportunity to enhance library OPAC by rendering not only bibliographic metadata in retrieval interface but also to extend supports to obtain results of language analysis such as word/phrase frequency, term circus, term links, term context and many more from the corpus of retrieved document through a managed visual interface.

2. REVIEW OF LITERATURE AND OVERVIEW OF SYSTEMS

Language documentation is a multidisciplinary research domain and consists of facets like specific stories, samples of speeches, writing codes etc[7]. Furbee[8] said that the theories of this emerging research area are still premature but Hill[9] suggested that language ideologies are integral part of the domain. Varela & Lee[10] reported that inclusion of language analysis ensures better utilities in language documentation and Austin[11] pointed out four primary constraints related to this: i) ambiguous nature of content (makes it difficult to determine quality of the resources); ii) multidisciplinary approaches (includes applied linguistics, anthropology, information science, etc.); iii) non-availability of comprehensive metadata schema; and iv) sustainability (trainings and capacity building programmes are required for achieving combined skills). Many models of language analysis have been developed by researchers to address the issues raised by domain experts like cognitive folk model for schema analysis[12,13], schema analysis model[14] and content analysis model[15]. A few researchers came up with frameworks and algorithms to solve problems of syntactic and semantic structures of languages[16], morphological alterations and morphotactics[17], and language analysis techniques compatible with Natural Language Processing ('meaning extraction' and 'named entity extraction')[18]. There are some ground-breaking software models in the domain of language documentation specifically meant for language analysis and results visualisation like 'Computer Aided Text Analysis (CATA)'[19], 'Corpus Linguistics (CL)'[20] and text analyses of indigenous literature[21]. The primary targets of these software solutions were schema analysis and identifications of narratives, metaphors, proverbs and other linguistic and paralinguistic features of languages. Many text analysis tools and guidelines have also been developed by Illinois Library[22]. Only a few studies are so far carried out on Indian tribal

Received : 15 June 2020, Revised : 25 August 2020
Accepted : 23 September 2020, Online published : 04 November 2020

languages documentation till date. The Santali language is one of the constitutionally approved Indian tribal languages. It is a sub-group of Austroasiatic-Mundari language and the only tribal language which has extensive web presence (Santali version of Wikipedia was launched in 2018 in Unicode-compliant Ol Chiki script)[23]. The 'Ol Chiki' script, invented by Pandit Raghunath Murmu in the 1920s is now used as an official script of Santali language. It consists of six vowels, twenty four consonants, and five sound symbols[24].

3. OBJECTIVE

Language documentation means creating or collecting, processing, preserving and disseminating linguistic data (such as, texts, audios, videos)[25]. Hundsberger[26] said way back in 2009 that "Libraries and language labs should be two closely cooperating entities, possibly even under the same roof, to make it easier for students to follow up links". This research work is aimed to support views of the domain experts that a library should support both retrieval and analysis from the same search interface and the main objective of this research is to design and development of an enhanced OPAC in an open source ILS (Integrated Library System) where end-users can access retrieved bibliographic resource in full-text systematically through a book reader and can visualise different results of language analysis in the same retrieval interface. It requires seamless integration of ILS, book reader software and language analysis tool to provide such enhanced information services in OPAC. The mechanism as proposed can support

Table 1. Tasks and steps related to framework design

Task	Steps	Result
Identification and selection of front-end ILS	Major considerations 1. Open source, Web-enabled, Unicode-compliant and REST/API support; 2. Scope for API development without hampering future upgradation; and 3. Ability to handle selective occurrences of a repeatable MARC tag/subfield and conditional display control	Possible candidate software: Koha, NewGenLib, Evergreen Selection Koha (keeping in view its flexibility in API handling and user base)
Identification and selection of back-end language analysis tool/service	Major considerations 1. Open source, Web-enabled, Unicode-compliant and data visualization support; 2. Scope for selective export of result display URL for linking in bibliographic framework selectively; and 3. Ability to generate a series of analysis services, separate URL for each analysis and control at the user end to navigate from one analysis to another.	Possible candidate software: Voyant, Lexos, Textalyser, Word & Phrase and WordSeer Selection Voyant (keeping in view its support for wide array of filetypes, different analysis services, improved visualization facilities and local installation options)
Linking mechanism	Major attributes 1. The architecture of the mechanism must support URL based linking between Koha catalogue dataset and Voyant corpus; 2. At Voyant end the corpus (Unicode-compliant full-text document associated with a MARC record in Koha) must produce unique URLs for available analysis results; 3. A repeatable MARC tag/subfield needs to be selected which can record corpus data URL (analysis service URL) along with other links; and 4. Linking script at Koha end needs to create a "Language analysis" tab conditionally in OPAC, if MARC tag/subfield provides analysis URL, and then can fetch data visualization from corresponding corpus available at Voyant end in the given OPAC tab with facility to navigate other analysis URLs.	Features: 1. As it is targeting OPAC, the logical selection of API end in Koha are - OPACUserCSS, and OPACUserJS (these two system preferences may be populated with the script that will link Koha and Voyant); 2. MARC tag 856 sub-field u is a logical choice to include locally available URL for language analysis services; as $u is repeatable, any number of analysis URLs and link to book reader URL can be included here; and 3. The script has two logical section – a) linking algorithm to connect book reader, if first occurrence of 856 $u is populated in MARC record; and b) linking algorithm to connect language analysis, if second occurrence of 856 $u is available in MARC record.

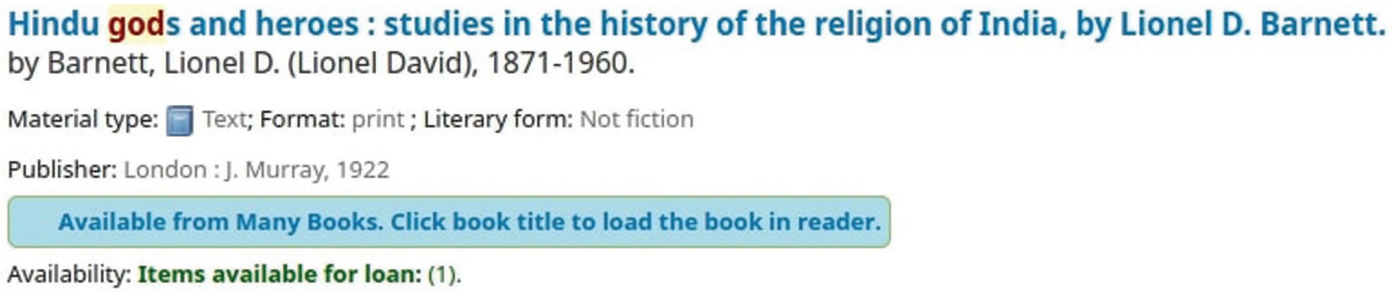

Figure 1. Indication of corpus availability in OPAC.

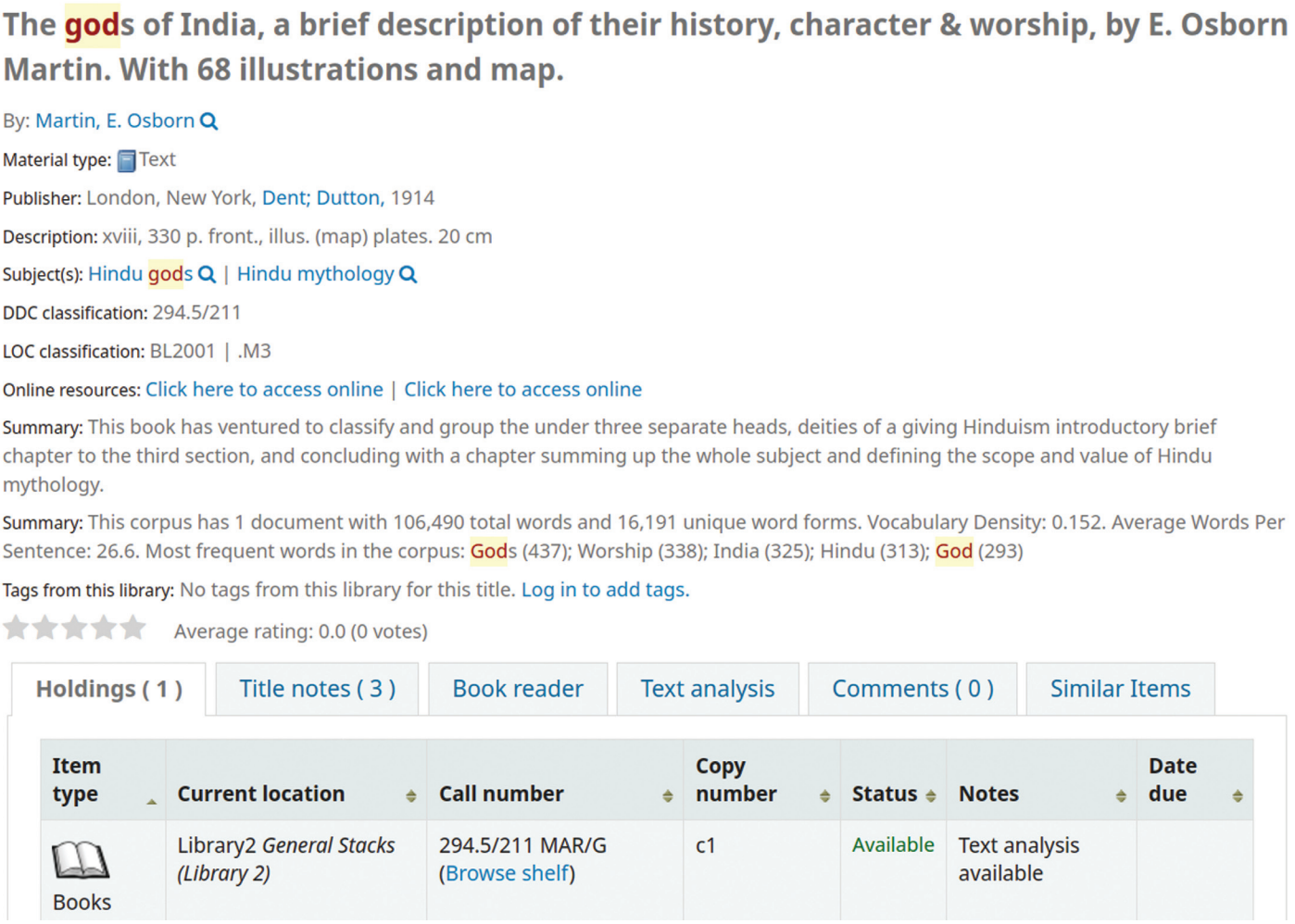

Figure 2. Integration of book reader and text analysis tabs in OPAC.

library materials in mainstream languages as well as for Unicode-compliant minority languages (like tribal languages of India).

4. STATEMENT OF PROBLEM

The above-stated objective related to design and development of an enhanced OPAC to support language analysis as well as visualisation of the analysis results is based on the following three interlinked research questions:

RQ 1: How and to what extent is it possible to customise OPAC of a selected web-enabled open source ILS so that it can display results of language analysis (provided through a back-end text analytics tool) for text corpus of a retrieved bibliographic record in the same interface along with other OPAC data (like holdings, availability, notes etc)?

RQ 2: What should be the parameters for selection of open source ILS and open source language analysis tool? How can these two software be linked to generate language analysis services in OPAC for documents having corpus and availability of the corpus is indicated at the metadata

Figure 3. Term circus in Voyant for the corpus of "The story of my experiments with truth".

minority languages (such as tribal languages in India) alongside mainstream languages?

5. METHODOLOGY

The research problems enumerated above need to be addressed through a systematic approach involving following major steps: 1) identification of selection parameters for front-end ILS and back-end language/text analysis service keeping in view the objective as stated; 2) selection of open source ILS and open source language analysis software for the framework on the basis of the parameters as identified in step 1; 3) development of linking mechanism so that selected ILS can be linked seamlessly with

Table 2. Voyant server – categories, services and facilities

Main Categories	Major services	Facilities
Corpus tools	Term circus, Term links, Term berry, Term radio, Term bubble, Mandala etc.	These are mainly visualization tools for displaying word cloud, network graph of high frequency terms, relationships between the terms etc.
Document tools	Summary, Reading, Term context, Phrases, Trends, Collocates, Correlation etc.	These services are meant for generic data related with the corpus like document information, summary report, frequency of terms and phrases, keywords in context, variation in term frequency, term proximity, line graph of the relative frequencies across the corpus etc.
Visualization tools	In addition to corpus tools, a set of advanced visualization tools are available like Loom, Knots, ScatterPlot, StreamGraph, TextualArc etc.	These are meant for generating real-time dynamic graphics related to change of the frequency of the words in a corpus, weighted centroid of terms in an arc, statistical analysis of word's correspondence etc.

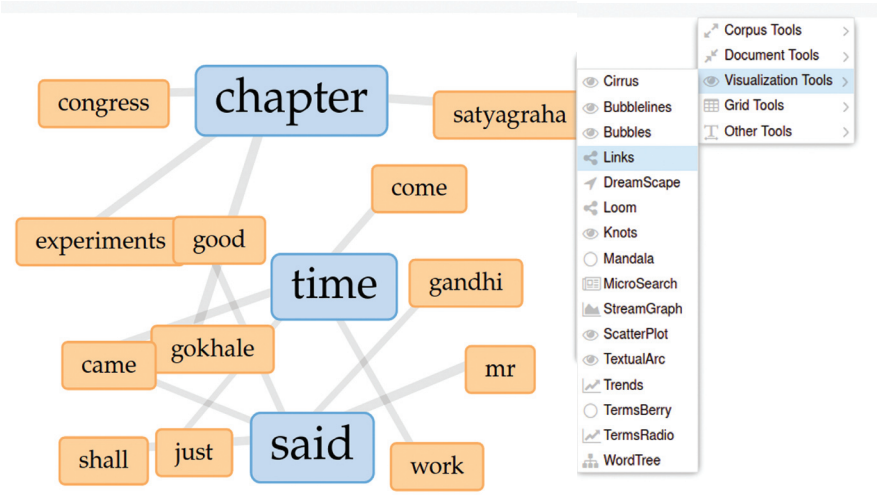

Figure 4. Language analysis tab in OPAC with navigation facility.

level? What MARC tag/subfield is the logical choice for linking corpus data in OPAC from the back-end language analysis service?

RQ 3: What will be the method to interlink the selected open source ILS with the selected open source language analysis tool in real-time and 'on-the-fly'? Is it possible to develop a linking mechanism that would not create any issues/problems for future upgradation of the selected ILS and the selected language analysis tool? Is it possible to extend the mechanism to support Unicode-compliant

the back-end language analysis service and can display analysis results automatically for the retrieved documents having corpus uploaded in language analysis tool (it is quite possible that all documents in catalogue database won't be having corpus in the back-end and thereby selective display control in linking mechanism is a must); 4) preparation of bibliographic data entry rules to indicate the presence of corpus in a selected MARC tag/subfield (the script must be able to handle different occurrences of a repeatable tag/subfield for accurate extraction of data required for link creation); and 5) testing and debugging with documents in two mainstream languages (Unicode-compliant full-text in English and Bengali) and a minority language (Santali - a tribal language). The steps as mentioned can further be explained under two groups – a) framework design; and b) script development to support linking mechanism in the framework.

5.1 Framework Design

The tasks and steps related to framework design may be represented in tabular format (Table 1) for better explanation. It includes a total of three task areas supported by ten parameters and attributes to provide a theoretical base for developing the proposed enhanced OPAC.

The testing and debugging are done with MARC formatted bibliographic records in three languages as mentioned above (step 5 in methodology). The bibliographic metadata and full-text resources for documents in all three languages need to be Unicode-compliant.

5.2 Linking Mechanism

The linking mechanism forms the core of the framework as described in section 5.1. The entire linking mechanism may be divided into two groups and can be represented by using Fig. 1 through Fig. 3.

5.2.1 Koha Side Workflow

The ILS side workflow includes development of the following two scripts to trigger the linking mechanism of the framework.

- Use of OPACUserCSS to create a class which can be called (in OPACUserJS) to indicate corpus availability and to control OPAC display (Fig. 1).
- Use of OPACUserJS to create – (a) link to book reader if first occurrence of 856 $u is present; (b) link to corresponding corpus in Voyant if second occurrence of 856 $u is present; and c) link to call opac-detail.tt to create tab(s) depending on the presence of (a) and/or (b). The perl script opac-detail.tt creates tab(s) on the basis of call from OPACUserJS (Fig. 2).

5.2.2 Voyant Side Workflow

The language analysis works in the Voyant side includes creation of corpus in local Voyant server. It is followed by unique corpus id creation in the form of URL for easy identification of any one language analysis service *(http://localhost:8888/?corpus=lib2bid36&view=Cirrus)* and subsequent encoding of the service URL in the second occurrence of 856 $u (Fig. 3).

6. ENHANCED OPAC THROUGH LANGUAGE ANALYSIS

Van Esch, D., *et al.*[27] opined that language documentation deals with huge amount of raw textual data and it is difficult to be annotated and analysed in timely fashion. In this context, Acharya, S., *et al.*[28] also said that text analytics is a low-hanging fruit in the domain of language documentation and an application of visualisation techniques may generate effective information services in addition to the typical library services.

The bits and pieces of the framework as described in section 5 are connected seamlessly to provide users an opportunity to access full-text in book reader tab and language analysis in another tab along with other typical dataset generally expected from a library OPAC. Thereby, this open source software based framework is completely amenable to the advices as provided by the experts in the domain of language documentation (Fig. 4). Voyant Tools is an open-source toolkit for developing web-enabled text analysis environment. The source code is available from Github (http://github.com/sgsinclair/Voyant) under a GPL3 license. Voyant server is a Java application and therefore system must be preloaded with Java for local installation (tested with openjdk version 1.8 in this framework). It is a global initiative under the project leaderships of Stéfan Sinclair, McGill University and Geoffrey Rockwell, University of Alberta[29]. The major text analysis services (Table 2), generated by backend Voyant server (Fig. 4) are available in the enhanced OPAC with control in the hand of end users to navigate from one analysis to another without difficulties.

The availability of Unicode-compliant corpus of a document in Indic-script can also produce similar language analysis results. For example, the text corpus of Gitanjali (a collection of poems by Rabindranath Tagore, originally published in Bengali in 1910, subsequently translated in English in 1912 and won the Nobel Prize for Literature in

Figure 5. Language analysis tab in OPAC for Indic-script (Bengali).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+1C5x	᱐	᱑	᱒	᱓	᱔	᱕	᱖	᱗	᱘	᱙	ᱚ	ᱛ	ᱜ	ᱝ	ᱞ	ᱟ
U+1C6x	ᱠ	ᱡ	ᱢ	ᱣ	ᱤ	ᱥ	ᱦ	ᱧ	ᱨ	ᱩ	ᱪ	ᱫ	ᱬ	ᱭ	ᱮ	ᱯ
U+1C7x	ᱰ	ᱱ	ᱲ	ᱳ	ᱴ	ᱵ	ᱶ	ᱷ	ᱸ	ᱹ	ᱺ	ᱻ	ᱼ	ᱽ	᱾	᱿

Figure 6. Unicode block for Ol Chiki, Unicode, ver 13.0, March 2020.

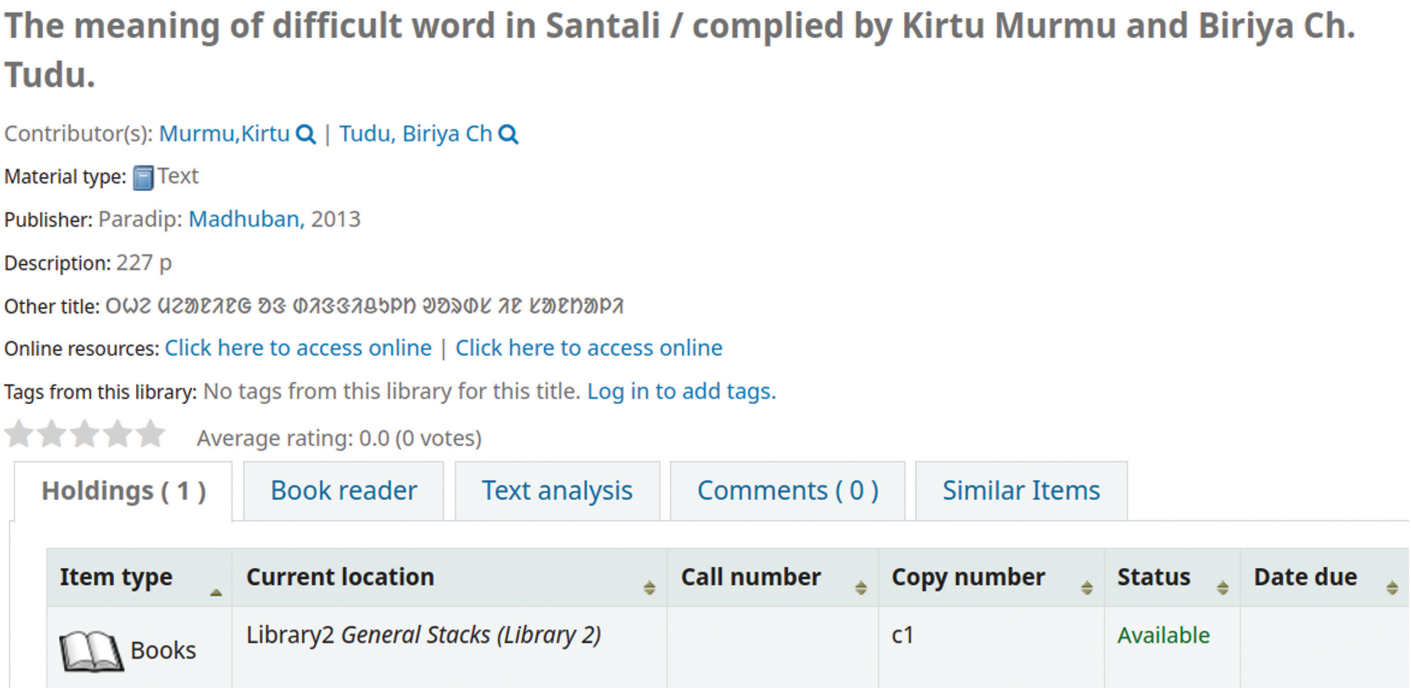

Figure 7. Display of a book in OL Chiki script in book reader tab of OPAC.

Figure 8. Language analysis tab in OPAC for Indic-script (Ol Chiki).

1913) if available in the library catalogue can lead to Indic-script based language analysis services in OPAC (Fig. 5).

7. EXTENSION OF MECHANISM TO A TRIBAL LANGUAGE

The definition of minority language varies but in most of the cases a minority language is a language which is spoken by less than 50 per cent of a population in a given region, state or country[30]. Obviously the size of the speaker population in a conferred geographic area is the key characteristic of a minority language and interestingly a minority language in one area can be a majority language in another region[30-31]. The report on language (Table C-6, published in 2018) by Census of India includes two parts – a) 22 languages that are included in the 18th Schedule of the Indian Constitution; and b) those not included in it (total 99). The group 'a' includes two major tribal languages (listed in the Eighth Schedule) namely Bodo and Santali. Many of the tribal languages in India have shown negative growth in compare with 2001 census including two major tribal languages Bodo and Santali (included in the Eighth Schedule).

This paper has selected Santali to extend the language analysis mechanism as discussed in the foregoing sections (from two tribal languages included in the Eighth Schedule of the Constitution namely Bodo and Santali) as a minority tribal language of India. The reasons are :

- Santali is spoken by 69,73,345 people whereas Bodo is spoken by 14,54,547 people[32,33];
- Santali has now its own script Ol Chiki (included in Unicode standard 5.1.0, released in April, 2008 and continues - http://unicode.org/versions/Unicode 5.1.0 – Fig. 6) whereas Bodo is written in Devnagri script;
- Ol Chiki script has now many Open Type Fonts (OTF) supports like Sakal Bharati (from Technology Development for Indian Languages, Ministry of Electronics & Information Technology, Govt. of India), Noto Sans Ol Chiki (developed by Google) etc. All the major web browsers now support display in Ol Chiki and thereby full-text display in OPAC is an easy task now with the help of a local flip-book reader tool (Fig. 7);
- Santali has many literary and cultural heritage objects and recently made significant stride towards mainstream Indian languages;
- Santali language Wikipedia in Ol Chiki script (sat.wikipedia.org) was launched in August 2018 after obtaining necessary approval from Wikimedia Language Committee.

The technical factors as mentioned in point 2,3 and 4 are key ingredients in extending the language analysis mechanisms as applied in mainstream languages (Fig. 1 through Fig. 5) to Santali language resources in Unicode-complaint Ol Chiki script without much trouble (Fig. 8).

8. CONCLUSIONS

Language documentation helps to stop the endangerment of languages and holds its cultural diversity. UNESCO atlas of the world's languages in danger[34] reported that India has most endangered languages in the world (197 endangered languages) and the Endangered Languages Project[35] identified 201 endangered Indian languages. Krauss[36] predicted that up to 90 per cent of all languages will be extinct by the end of this century. Library can play a vital role for the activities related to integration of information retrieval with language analysis and other facets of language documentation. All the tools and techniques applied in this research work are available as open source and can be implemented in libraries of any type or size in developing enhanced OPAC.

REFERENCES

1. Boerger, B.H.; Self, S.N.; Moeller, S.R. & Reiman, D.W. Language and culture documentation manual, 2020. https://leanpub.com/languageandculturedocumentationmanual (accessed on 13 June 2020).
2. Gippert, J.; Himmelmann, N.P. & Mosel, U. Essentials of language documentation. Mouton de Gruyter, Berlin, 2006. doi: 10.1515/9783110197730.
3. Austin, P.K. Data and language documentation. In Essentials of language documentation, edited by Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel. Mouton de Gruyter, Berlin, 2006, 87–112. doi: 10.1515/9783110197730.
4. Austin, P.K. Training for language documentation: The SOAS experience. *In* The International Conference on Endangered Austronesian Languages, 5–7 June, 2007, Providence University, Taiwan. 2007. pp. 223-34. https://www.researchgate.net/publication/238506014_Training_for_Language_Documentation_the_SOAS_experience (accessed on 03 June 2020).
5. Himmelmann, N.P. Documentary and descriptive linguistics. *Linguistics,* 1998, **36**(1), 161–95. doi: 10.1515/ling.1998.36.1.161.
6. Woodbury, A.C. Defining documentary linguistics. In Language documentation and description (volume 1), edited by Peter K. Austin. Hans Rausing Endangered Languages Project, SOAS, London, 2003, 35–51. http://www.ddl.cnrs.fr/Colloques/3L_2008/3LCourseMaterial/Woodbury_2003_Documentation.pdf (accessed on 03 June 2020)
7. Woodbury, A.C. Language documentation. In the cambridge handbook of endangered languages, edited by Peter K. Austin & Julia Sallabank. Cambridge University Press, Cambridge, 2011, 159-86. doi: 10.1017/CBO9780511975981.009.
8. Furbee, N.L. Language documentation: Theory and practice. In Language documentation: Practice and values, edited by Lenore A. Grenoble & N. Louanna Furbee. John Benjamins Publishing Company, Amsterdam, 2010, 3-24. doi: 10.1075/z.158.04fur.
9. Hill, J.H. The ethnography of language and language documentation. In Essentials of language documentation, edited by Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel. Mouton de Gruyter, Berlin, 2006, 113–28. doi: 10.1515/9783110197730.113.

10. Varela, M.E. & Lee, N.H. Language documentation: A reference point for theatre and performance archives? *Int. J. Perform. Arts Digital Media*, 2018, **14**(2), 17-33. doi: 10.1080/14794713.2018.1453242.
11. Austin, P.K. Current issues in language documentation. In Language documentation and description, vol. 7, edited by Peter K. Austin. SOAS, London, 2010, 12-33. http://www.elpublishing.org/PID/080 (accessed on 04 June 2020).
12. D'Andrade, R. A folk model of the mind. In Cultural models in language and thought, edited by D. Holland & N. Quinn. Cambridge University Press, 1987, 112-48. doi: 10.1017/CBO9780511607660.006.
13. D'Andrade, R. The identification of schemas in naturalistic data. In Person schemas and maladaptive interpersonal patterns, edited by Mardi J. Horowitz. University of Chicago Press, Chicago, 1991, 279-301
14. Mckee, A. A beginner's guide to textual analysis. Metro magazine: Media & education Magazine, 127/128, 2001, 138-49. https://eprints.qut.edu.au/41993/ (accessed on 11 June 2020).
15. Frey, L.; Botan, C. & Kreps, G. Investigating communication: An introduction to research methods (2nd ed.). Allyn & Bacon, Boston, 1999. http://mason.gmu.edu/~afinn/html/teaching/courses/f03_comm250/fbk_chapters/09.pdf (accessed on 11 June 2020)
16. Zaenen, A. & Uszkoreit, H. Overview of language analysis and understanding. In Survey of the state of the art in human language technology, edited by Ron Cole *et al.* Cambridge University Press, 1997, 112-13. http://www.dfki.de/~hansu/HLT-Survey.pdf (accessed on 04 June 2020).
17. Karlsson, F. & Karttunen, L. Sub-sentential processing of language analysis and understanding. In Survey of the state of the art in human language technology, edited by Ron Cole et al. Cambridge University Press, 1997, 113-14. http://www.dfki.de/~hansu/HLT-Survey.pdf (accessed on 04 June 2020).
18. Goldenstein, J.; Poschmann, P. & Händschke, S.G.M. Linguistic analysis: The study of textual data in management and organisation studies with NLP. *Acad. Manage*. 2007. doi: 10.5465/ambpp.2015.10882abstract.
19. Krippendorff, K. Content analysis: An introduction to its methodology. Thousand Oaks, SAGE, 2004, **2**, doi: 10.1177/1094428108324513.
20. Pollach, I. Taming textual data: The contribution of corpus linguistics to computer aided text analysis. *Organ. Res. Methods*, 2012, **15**(2), 263-87. doi: 10.1177/1094428111417451.
21. Bernard, H.R. & Ryan, G. Text analysis: Qualitative and quantitative methods. In Handbook of methods in cultural anthropology. 1998. http://www.analytictech.com/mb870/readings/bernard_ryan_text_analysis.pdf (accessed on 11 June 2020).
22. Text mining tools and methods. Illinois Library, https://guides.library.illinois.edu/textmining (accessed on 11 June 2020).
23. Murmu, C. Ol Seched (in Al Chiki). Pandit Raghunath Murmu Human Welfare Trust, Rairungpur, 2004.
24. Mohanta, B.K. Invention in Al Chiki script for Santali language: An attempt to preserve an endangered tribal dialect. In Endangered cultures and languages in India: Empirical observations, edited by G. K. Bera, K. Jose, & North Eastern Institute of Culture & Religion (Gauhati, India), 2015, 211-29.
25. Klessa, K. Language documentation. In Book of knowledge of languages in danger, edited by Michael Hornsby. 2014, 106-120. http://languagesindanger.eu/book-of-knowledge/ (accessed on 03 June 2020)
26. Hundsberger, S. Foreign language learning in second life and the implication for resources provision in academic libraries. ARCADIA, 2009. http://www.arcadiaproject.lib.cam.ac.uk (accessed on 03 June 2020)
27. Van Esch, D.; Foley, B. & San, N. Future directions in technological support for language documentation. *In* Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, 26-27 February 2019, Honolulu, USA, 2019, 14-22. https://journals.colorado.edu/index.php/computel/article/view/341 (accessed on 13 June 2020)
28. Acharya, S.; Sanyal, D.K.; Mazumdar, J. & Das, P.P. Archiving endangered mundā languages in a digital library. https://ndlproject.iitkgp.ac.in/sites/default/files/website_images/publications/Paper_Archiving%20Endangered%20Mund%C4%81%20Languages%20in%20a%20Digital%20Library.pdf (accessed on 13 June 2020).
29 Sinclair, S. & Rockwell, G. Voyant Tools (v. 2.4). http://voyant-tools.org/ (accessed on 13 June 2020).
30. Grenoble, L.A. & Singerman, A.R. Minority languages. Oxford University Press, 2014. doi: 10.1093/OBO/9780199772810-0176.
31. Pandharipande, R. Minority matters: Issues in minority languages in India. *Int. J. Multicultural Soc.*, 2002, **4**(2), 213-34. www.unesco.org/shs/ijms/vol4/issue2/art4 (accessed on 13 June 2020).
32. Down to Earth. Language census: Many tribal tongues now have fewer takers. 2018 https://www.downtoearth.org.in/news/environment/language-census-many-tribal-tongues-now-have-fewer-takers-61044 (accessed on 14 June 2020)
33. Office of the Registrar General, India. Census of India 2011: Paper 1 of 2018, Language India, States and Union territories (Table C-16). https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf (accessed on 14 June 2020).
34. UNESCO atlas of the world's languages in danger: http://www.unesco.org/languages-atlas/ (accessed on 03 June 2020).
35. Ethnologue: Languages of the world https://www.ethnologue.com/ (accessed on 03 June 2020).
36. Krauss, Michael E. The World's languages in Crisis. *Language,* 1992, **68**(1), 4-10. doi: 10.1353/lan.1992.0052.

CONTRIBUTORS

Parthasarathi Mukhopadhyay is a Professor in the Department of Library Science, University of Kalyani. His contributions to this paper are - developing the technical architecture including the scripts required for integration of language analysis services in Koha OPAC, building mechanism for text analytics for Ol Chiki script and finalizing the objectives, the research questions and the methodology for this research oriented publication.

Anirban Dutta is a Junior Research Fellow (JRF) of the Department of Library Science, University of Kalyani. He has completed his BLIS and MLIS from the same university. He has performed the following activities for this research work – review of literature, overview of the related initiatives, metadata encoding for software framework, and testing & debugging of the language analysis services.