The *shabd* portal – searchable lexical resources for Indian languages by Government of India

Mercy Lalrohluo Hmar[1], Girish Nath Jha[2], Dhananjay Singh[3]

Commission for Scientific and Technical Terminology[1]; Jawaharlal Nehru University[2]; Chairman, Commission for Scientific and Technical Terminology[3]
New Delhi, India[1], New Delhi, India[2], New Delhi, India[3]
mercylhmar.cstt@gmail.com[1] , girishjha@jnu.ac.in[2], dhananjaysinghcstt@gmail.com[3]

Abstract

The shabd portal (https://shabd.education.gov.in) of Commission for Scientific and Technical Terminology (CSTT), a subordinate office under the Ministry of Education, Department of Higher Education, Government of India (GOI) is a data server designed and developed by Prof. Girish Nath Jha, former Chairman CSTT, featuring all the standardized scientific and technical glossaries of CSTT in digital searchable mode. The aim is to launch a central repository for the terminologies prepared in Indian Languages, thus enriching the language bank of India enabling user friendly and free access to standardized terminology. This website is available in 22 Indian languages. The data covers several domains of science, humanities, engineering, medical science and agriculture subjects. The data is dynamic with regular updates in various domains. Users can search the equivalents of terms in Indian languages and submit their feedback for those equivalents prepared by CSTT. The unique feature of the search platform is that users have various options for search, based on languages, subjects, dictionary type and language pairs. The user can also choose to search in a specific glossary or the entire collection which includes about 471 glossaries having about (29,56,125 headwords).

Keywords: standardized terminology, Indian languages, domains, searchable mode, scientific and technical glossaries

About CSTT -

The Commission for Scientific and Technical Terminology of GOI

The Commission for Scientific and Technical Terminology was established with the objective to evolve technical terminology in all Indian Languages, on 1st October 1961 by the Presidential Order dated April 27, 1960, through a resolution of the Government of India (Ministry of Education), as per the recommendations of the Committee constituted under the provisions of the Clause (4) of the Article 344 of the Constitution of India. The Commission was established with the objective to evolve standardized technical terminology in all Indian Languages. Prof. Dhananjay Singh is the current Chairman of CSTT.

The main objectives of the Commission are to evolve standard terminology, propagate its use, obtaining useful feedback and distribute it widely. In the process of evolution of scientific and technical terminology and reference material in Indian Languages, the Commission collaborates with State Governments/ Universities/ Regional Text-Book Boards and State Granth Academies to ensure uniformity of terminology in Indian languages.

The Commission publish glossaries, definitional dictionaries, journals, monographs, encyclopaedia etc; to see that the evolved terms and their definitions reach the students, teachers, scholars, scientists, officers etc., to ensure proper usage necessary updating/ correction improvement on the work done (through workshops/ seminars/ orientation programmes) by obtaining useful feedback, to coordinate with all States to ensure uniformity of terminology in Hindi and other Indian languages.

Some of the Schemes of the Commission are as follows:

1. Preparation of English-Hindi and Hindi-English Technical Glossaries/Dictionaries
2. Preparation of Bilingual Technical Glossaries/Dictionaries
3. Preparation of Trilingual Technical Glossaries/Dictionaries
4. Preparation of National Technical Terminology
5. Preparation of Definitional Dictionaries
6. Preparation of Technical Encyclopedias
7. Preparation of School-Level Terminology
8. Preparation and/or Approval of Departmental Glossaries

Proceedings of WILDRE-8, pages 61–66

9. Revision and Updating of Glossaries
10. Identification and Publication of Pan-Indian Terms
11. Propagation, Expansion and Critical Review of Terms Coined and Defined via Seminars/ Conference/ Workshops.
12. Scheme of Production of University Level Books in Hindi and Regional Languages
13. Preparation and Publication of Monographs
14. Preparation and Publication of Journals (Gyan Garima Sindhu (for Humanities subjects) and Vigyan Garima Sindhu (for science subjects).
15. Sales of Publications
16. Free Distribution of Publications
17. Organizing Exhibitions

About the *shabd* portal

One of the new initiatives taken by the CSTT in the field of propagation of scientific and technical terminology is the *shabd* portal hosted at https://shabd.education.gov.in .

The *shabd* is a data server which features all the glossaries of CSTT in digital searchable mode. Other institutions/ agencies preparing dictionaries will also be able to host their work in digital form on this platform. The aim is to showcase a central repository for all the terminologies prepared in/for Indian Languages.

Through this platform the users will not only be able to search the equivalent terms of scientific and technical terminology in Indian languages but will also be able to register their feedback for the equivalents already prepared by CSTT. This aims at providing a user-friendly search environment and also in involving the users via the feedback mechanism. The unique feature of the search platform 'https://shabd.education.gov.in' is that the user will have various options for search, whether based on languages, subjects, dictionary type, or language pairs. Not only this, the user can choose to search in a specific glossary or through the entire collection which currently includes about **471** glossaries having about (**29,56,125** headwords).

One website – 22 languages- wide user base

The 'Shabd' website is accessible in 22 languages and contains the data in the form of searchable glossaries in 22 Scheduled Indian languages. The users can easily search the glossary and use the plethora of Indian language equivalents available for a wide range of subject domains.

Concept design execution, website development and programming was done by Prof. Girish Nath Jha, former Chairman, CSTT and was launched in March 2024.

The following partners of CSTT have greatly contributed to make the content available for the website:

- Udaan Project team, IIT Bombay (All Languages) and International Centre for Free and Open Source Software (ICFOSS), Kerala (Malayalam) for OCR of printed glossaries

- GIST group, CDAC Pune for UNICODE Consultation and support

- Central Institute of Indian Languages (CIIL) Mysore for Data Typing support

- Bhartiya Bhasha Samiti (BBS), Overall Guidance

- Language Division, Ministry of Education for Administrative Support

History of *shabd*

The inception

CSTT prepares the terminologies through the Expert Advisory Committees, consisting of subject and language experts along with linguists and Sanskritist who are focused on finding out the equivalent terms in the specific subject areas and language. The terminology prepared by CSTT has not only been used by Granth academies, textbook boards and publication cells to prepare textbooks but is also being used by institutions such as NCERT, NTM, AICTE and so on.

CSTT glossaries available in printed form or pdf form on the official website were not searchable and this made it less popular among the user group. The need to have easily accessible and searchable glossaries resulted in the inception of the vision for a website wherein the glossaries prepared by CSTT are available in searchable form.

Thus began the humongous task of building such a website and getting the data of the glossaries in digital form available for the task to take shape. Prof. Girish Nath Jha, designed and programmed

a server which could host the data locally. After a long and strenuous brainstorming, the initial layout of the local server took shape. Based on several trials, hits and errors, the server slowly evolved to take a stable form. The demo of the server was shown on several occasions to collect valuable inputs and suggestions from the persons of interest and such constructive suggestions were implemented to make the server more user friendly and useful.

For the data, CSTT began with the glossaries available in digital soft copy which were initially digitized as part of the agreement between CSTT and CIIL. Then it was found that many legacy publications of CSTT were not available in such digital form. Thus began the search for partners who could do this task in record time. This led to our MoU with UDAAN, IITB team led by Prof. Ganesh Ramakrishnan. The team has helped CSTT in digitizing most of the legacy documents and are continuing to do so.

Once the server was tested and checked locally, the website team helped to carry out the necessary task required to acquire the domain and host the *shabd* website in a public domain for use by the users. The site first went active in March, 2024, since then, it has had 20,66,090 hits from across the country and the world as of now.

The journey till date

Initially the website was built having an English and Hindi layout. Both the layouts were not interconnected. Changes had to be made on both pages regularly when any suggestions or bugs were identified. Thereafter the layouts were merged and programmed by Prof Jha to switch, without having to make corrections in every page and content.

The search was simple and user had to search in a specific glossary initially. Next the search was extended to all glossaries and the feedback mechanism was also introduced. The next big jump was the introduction of all 22 Indian languages. All the code strings were translated to the 22 target languages by a panel of linguists who were a part of CSTT's meetings to update the principles of terminology preparation. The linguists timely helped CSTT whenever there was any issue to give language equivalents for the various contents available on the website. The panel of linguists headed by Prof. K. L. Verma, Former Chairman, CSTT are:

1. Dr Bipasha Patgiri, Assistant Professor, Department of Linguistics and Language Technology, Tezpur University, Assam
2. Prof. Niladri Sekhar Dash, Head, Linguistic Research Unit, Indian Statistical Institute, Kolkata
3. Prof. Swarna Prabha Chainary, Professor, Dept. of Bodo, Gauhati University
4. Prof. Lalit Mangotra, President, Dogri Sanstha, Jammu
5. Dr. Baldevaanand Sagar,National President, World Sanskrit Media Council
6. Prof. Chandan Kumar,Professor, Faculty Member,Dept. Of Hindi, University of Delhi
7. Dr. Girisha Bhat A, Professor and Principal, Govt First Grade College, Siddakatte, Karnataka
8. Prof Aadil Amin Kak, Dean, School of Arts, Languages and Literatures University of Kashmir
9. Dr. Kiran Budkuley,Author, Critic& Translator, President, Aksharpath, Goa.
10. Prof. Awadesh Kumar Mishra, Professor, Chief Coordinator, Bharatiya Bhasha Samiti, Vishwakarma Bhawan, IIT Delhi, New Delhi
11. Dr. Shobha L, Member Research Staff, AU-KBC Research Centre , Madras Institute of Technology, Anna University, Chennai
12. Dr Hanjabam Surmangol Sharma, Assoc. Professor, Department of Linguistics, Manipur University, Imphal
13. Dr. Shakuntala Gawde, Head and Assistant Professor, Department of Sanskrit, University of Mumbai
14. Shri. Vishnu Bahadur Gurung, In-Charge (Rtd.), Nepali Service, External Services Division, All India Radio, New Delhi
15. Prof. Panchanan Mohanty, Dean, School of Languages and Literature/Humanities, Nalanda University
16. Dr Suman Preet, Professor , Department of Linguistics and Punjabi Lexicography, Punjabi University Patiala
17. Prof. Satyapal Singh, Professor, Department of Sanskrit, University of Delhi

18. Dr. Thakur Prasad Murmu, Assistant Professor, Department of Santali, Sidho Kanho Birsha University
19. Smt. Shalini Sagar, Senior Broadcaster All India Radio, Sindhi Language Expert, NCERT
20. Dr. S. Arulmozi, Professor & Head, Centre for Applied Linguistics and Translation Studies, University of Hyderabad
21. Dr. MC Kesava Murty, Professor, Dept. of Dravidian and Computational Linguistics, Dravidian University, Kuppam
22. Prof. Syed Imtiaz Hasnain, Retd. Prof of Sociolinguistics, Chair-Professor, Maulana Azad National Urdu University (MANUU)

Figure 1: Shabd website localization

Main features of the Shabd Website:

The various features which were added slowly as the 'shabd' website began to become popular among the users are:

- **Localization of website content**

Figure 1: Main content of the website is available in 22 Indian languages which can be manually selected

- **Website link to language information page/ website of organization working for that language**

Figure 2: Link to the major organization working in the specific language provided on the main page when language preference is selected

- **Addition of the footer with features such as -popular website, credits, contact us etc**

Figure 3: Footer has all the popular and websites along with the credit and contact details

- **Search options – language wise, subject wise, language pair wise, glossary type, etc.**

Figure 4: many options to customize the search is available for the user

- **Transliteration feature**

Figure 5: Transliteration feature for all non-Devnagari glossaries

- **Glossary details**

Figure 6: Glossary title, status of the glossary, headword count and Officer in-charge details and contact details are available for the glossaries

- **List of Expert involved in the project**

Figure 7: Expert panel involved in terminology preparation is mentioned in this section

- **Log of the user stat & feedback**

Figure 8: Feedback of user is recorded and necessary action is taken from time to time

- **User feedback mechanism**

Figure 9: User can provide feedback for every equivalent and contribute. Feedbacks are considered and placed before the Expert Advisory Committee for review

- **Popular glossaries on main page**

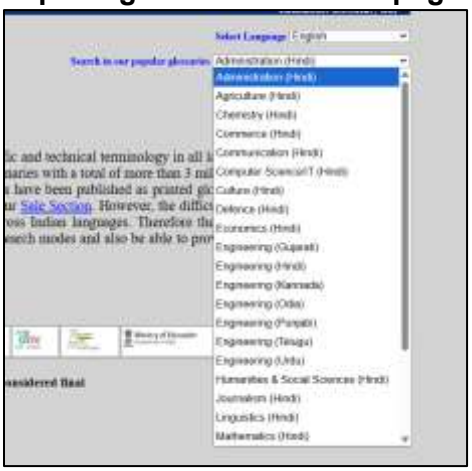

Figure 10: All popular glossaries are enlisted in this as per the frequency of user hits

- **List of glossaries in alphabetical order on main page**

Figure 11: The alphabetical list of all glossaries is available as drop-down menu on main page

The entire collection which as of now includes about **471** glossaries having about **29,56,125** headwords. This covers disciplines from Humanities, Social Science, Medical Science, Engineering, Agricultural Sciences and Science which include **more than 60 subjects** such as Agriculture, Public Administration, Chemistry, Botany, Zoology, Psychology, Physics, Economics, Ayurveda, Mathematics, Civil and Electrical Engineering, Computer Science, Political Science, Culture, Transport, Geology,

Capital Market, Cell Biology, Broadcasting, Journalism, Music and Fine arts, CSIT, AIML, Linguistics, Forestry, Entomology, Plant Pathology, Soil Science, Sports, Nematology, Sericulture, LIS, etc. Many more domains are being added as and when the terminology preparation meetings are being held across the country.

Bibliographical reference:

1. Resource available from CSTT main website (https://cstt.education.gov.in)
2. Resource available from CSTT shabd website (https://cstt.education.gov.in)