• XOBDO.ORG is a network of volunteers from across the globe developing and maintaining an online multilingual dictionary emphasizing the languages of North-East India, especially Assamese, Bodo, Dimasa, Khasi etc.

• More details here: https://en.wikipedia.org/wiki/Xobdo.org
HISTORY

• Born on 2006
  • It ran ASPs on Windows IIS and the primary database on RDBMS based on Microsoft SQL Server
• 2007, it migrated to open-source technologies like PHP, MYSQL
• 2012, migrated to GoDaddy platform
• The purpose of this document or presentation is to document the collective vision of XOBDO.ORG community.

• The document will be the basis of future path, way forwards for XOBDO, and guidelines for any projects to be taken up.
CURRENT FEATURES

• Dictionary
  • Multi-lingual, multi-directional dictionary with fuzzy search.
  • English-Assamese is the main emphasis, however, currently supports 22 languages.
  • Provisions for pronunciations, etymology, dialect, synonyms, antonyms, similar meanings, contexts, photos & sound files.

• Phrases / Idioms (ফক্ষা-যাজনা)

• Web-based spell-checker (Litikai)

• Assamese Crossword Solver

• Offline Phonetic Keyboard

• Offline Dictionary Software

• Apps : Database downloadable free. At least 3 (three) Android & 1 (one) iOS Apps on circulation
## CURRENT DATA GAPS & 5-YEAR ASPIRATIONS

<table>
<thead>
<tr>
<th>Feature</th>
<th>Current (Feb 2022)</th>
<th>Aspiration by 2027</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assamese Word Count (Root)¹</td>
<td>45000*</td>
<td>60000</td>
</tr>
<tr>
<td>Assamese Word Count (Total)²</td>
<td>95642</td>
<td>?</td>
</tr>
<tr>
<td>English Word Count (Root)³</td>
<td>20625</td>
<td>60000</td>
</tr>
<tr>
<td>Dimasa Word Count</td>
<td>3370</td>
<td>5000</td>
</tr>
<tr>
<td>Karbi Word Count</td>
<td>2619</td>
<td>5000</td>
</tr>
<tr>
<td>Bodo Word Count</td>
<td>1140</td>
<td>5000</td>
</tr>
<tr>
<td>Image Files</td>
<td>1386</td>
<td>5000</td>
</tr>
<tr>
<td>Sound Files</td>
<td>118</td>
<td>20000</td>
</tr>
<tr>
<td>Simple Sentences/Phrases in multiple languages</td>
<td>Limited</td>
<td>5000 nos x 5 languages</td>
</tr>
</tbody>
</table>

- Estimate
  1. Root Word or Stems or Dictionary Entries
  2. Root Words plus various inflected forms
  3. We should give attention to increase the English word count. This is one of the best way to make it more popular.
ASSAMESE DATA GAPS

1. Enhance terminologies/glossaries (পরিভাষা) in various domains, like
   a. Pure Sciences (Physics, Chemistry & Mathematics)
   b. Life Sciences (Zoology, Botany, Microbiology, Medical Science, Agriculture, Fishery, Veterinary Science etc)
   c. Applied Sciences (Engineering, Information Technology etc)
   d. Economics/Commerce/Finance/Business Management
   e. Judicial / Legal

2. Enhance words with multiple dialects.

3. Etymology of the words.

4. Collocation Words (e.g. what adjective should be used with what nouns)

5. Enhance words with example sentences for …
   a. Homograph (A homograph is a word that shares the same written form as another word but has a different meaning)
   b. Homophones (each of two or more words having the same pronunciation but different meanings, origins, or spelling, for example new and knew)
   c. Reduplication / Duplifix (?) – ভাত-ভাত, হাত-চাত, ভরি-চরি, ডা-ডাঙীয়া etc
## TECHNICAL DEVELOPMENTS: CURRENT VS 5-YEAR ASPIRATION

<table>
<thead>
<tr>
<th>Feature</th>
<th>Current Status (Feb 2022)</th>
<th>Aspiration by 2027</th>
</tr>
</thead>
</table>
| Dictionary Functions   | Limited search ability.   | • One stop window for detailed meanings and explanation for any Assamese word, (“stem” plus any inflected form)  
                          |                           | • Words begin with, exact word, words containing etc.                                   |
| Colocation Dictionary  | Concept Stage             | A working / functioning pilot                                                      |
| Website (Public)       | Functioning               | Make more interactive / dynamic                                                     |
| Website (Project)      | Functioning. Designed for Desktop. | Redesign for small window devices, especially mobile phones.                      |
| Spellchecker (Litikai) | Web-based                 | • Fully integrate with main dictionary and develop as 100% fool proof spell-checker.  
                          |                           | • Develop Apps.                                                                     |
| Machine Translation    | N/A                       | Something functioning                                                              |
| Eco-System             | Limited                   | A vibrant eco-system of developers and users.                                       |
LITIKAI – THE SPELL-CHECKER
(5-YEAR ASPIRATIONS)

• Fully integrate with main dictionary and develop as 100% full proof spell-checker.
  • Explore technology
  • Implement
  • Maybe conduct hackathon among enthusiast developers

• Develop Apps.

• Integrate with existing text editor such as Word, or web apps such as gmail, gdoc and chat app like Whatsapp.
  • It will be very helpful to enable inline spellcheck feedback to users while they input Assamese.
ADVANCED TECH

- XOBDO.ORG wants to participate and collaborate in projects related to:
  - Text-to-Speech System
  - Optical Character Recognition / Hand-writing recognition
  - Spellchecker
  - Grammar Checker
  - Machine Translation
  - Voice Recognition / Auto-transcription
  - Sentiment Analysis
COLLABORATION
(5-YEAR ASPIRATION)

• Engagements:
  • Engage More Independent Developers
  • Engage More Researchers
  • Engage More School-College Students (Conduct Webinars, Offline Workshop & Seminars)
  • Ensure XOBDO.ORG is duly credited

• Collaboration
  • Govt of Assam, Dept of Education, SCERT
  • Sahitya Sabhas of different languages (অসম সাহিত্য সভা (Assam Literary Society, ALS), বড়ো সাহিত্য সভা, দেউবী সাহিত্য সভা আদি)
  • Universities (IITG, Tezpur University, Assam Don Bosco University etc)
  • Private Companies involved in language technologies (Google, Microsoft etc)
NURTURING AN ECO-SYSTEM
(5-YEAR ASPIRATION)

• One of the key objectives of XOBDO.ORG is to nurture an ecosystem of independent users, developers & researchers that collectively develop various tools around XOBDO as the backbone.
• After declaring “copy-left” of our database, a number of developers have already downloaded the database and independently developing app:
  • https://play.google.com/store/apps/details?id=in.co.axomi
• Many researchers are also using the database, especially from Dept of Linguistics, IITG and already published technical papers. (To ensure XOBDO.ORG is duly credited)
• To support the ecosystem, many useful API have been developed and being actively maintained.
• Apart from the main database, the followings are also available or will be available for download:
  • Fonts (Engage Font Designers to develop XOBDO's own stylish fonts)
  • Corpus (Assamese Text, Parallel Corpus)
  • Apps
  • Keyboards
This website / sub-domain was created to share language learning materials, originally created to teach Assamese to children of non-resident Assamese people living in the Middle-East & the United States.

Many contributors developed these materials over time.

**Aspiration for 2027:**

- Develop more courses, targeting non-resident children, non-Assamese residing in Assam.
  - Modes: Passive Online / Live Online Coaching / Offline
- Develop certification program in association with ALS (Assam Literary Society/ অসম সাহিত্য সভা)
  - Working / Functional Assamese
  - Professional Translators
  - Voice Professionals (Voice-over Professionals, News Reader, Video Reporters)
FEEDBACKS
In order to make Xobdo more user friendly a in-depth usability audit to be carried out.

More emphasis will be provided on human computer interaction (HCI)

An End User survey to be carried out with the help of User Survey through a form prepared by Dr Keyur Sorathia, renowned HCI expert from IIT Guwahati

A phonetic Assamese keyboard to be developed for iOS system like Google Indic keyboard for Android system. For this Xobdo may tie-up with Assam Don Bosco University for Government Funded project

Information architecture design needs to be more systematic

The XOBDO web site to be enhanced for Cross Device Compatibility

Suggestion from Dr Vikramjit Kakati:
Anjal Borah

to me

সকলে ঠিকই আছে। মোব ফানর পরা এটা অন্তর্ভুক্ত কবির পাকিলে রেয়া নহয়।

Assamese corpus database

রেখেচ corpus database এটা তৈয়ার কবির পাকিলে গরেষণার বাবে আমি মুকলি কবি দিব পারে।

ঋষিক
Dear Bikram Da,

Thank you for initiating this. I only have 3 points for your review:

1. Language localization is going to grow more in the coming days. Can Xobdo (due to its authenticity and accuracy) can work as a service provider for the different localization projects?

2. Can the spell checker component be expanded more, with a viewpoint to accept different thoughts (at least for discussion purpose) on spelling in order to ease the language?

3. I believe many people use the online version to search for English words to Assamese words. Can we focus on that attribute as per our 2027 target?

Regards,

Manjit Nath
PANKAJ JYOTI MAHANTA

Dhanyabad. Utey dibole kibasake theke gachil.

Eyaab bhitpil kherent jagadit he'bole mon thekio alan kamar babe jagadit habor upata nohoara hoy. Sare ota kora julo, miito bho aagat monilie aahil. Saieta khibile hala h'er builit bhabo. Parasokalab anga bhoog sieeta loh pab--

Sahajk-Sahajkowor shreepita thaka chara-chatrik kelenkai jagadit khibi pabei seyta habor bhalo. Totoya natoon udambot kindo kama h'er. Toelokab samro olo mon habor, aroo kibi bhalo pabo. Toelokab pahlit natoonke pabgun alan Sahajk-Sahajkowor shreepita sakalak bichhib bore.

Pankaj Jyoti
ধন্যবাদ, স্লাইড কেইখন্ চালোঁ।ভাল হৈলে।

For spell checker (Litikai), effort to integrate into existing text editor such as Word, or web apps such as gmail, gdoc will be very helpful to enable inline spellcheck feedback to users while they input Assamese.

For machine translation, my suggestion is to create a English-Assamese corpus of parallel sentences (pair of sentences, one Assamese and the other English, where one is a translation of the other). Such a corpus is critical to doing high quality machine translation and is currently missing for Assamese.

Such an open-source corpus can then be offered to Google, Microsoft and others to improve the quality of their Assamese machine translation services, as well as to create one or more machine translation services directly

Regards,

Satyajit Nath
DR. ARUP NATH

• Future plans/areas to be improved from a User’s perspective
  • Can we track each and every searched lexical item and phrases by the users? If we can do that we would know what the users are searching and accordingly we can add them in the data base, rather than simple suggestions?
  • The dictionary lacks enough scientific terminologies/ glossaries, perhaps. While working on the ‘Engineering Glossary in Assamese’ the entire team at Tezpur University is referring to the Xobdo and most of the words were not found. This way, I am sure other technical terms from other subjects too might not be having much presence in the Xobdo database. Since academicians and language practioners and also the students would be our target users, I think we should chalk out a plan to include more and more ‘Axomiya Poribhaxa’ for each and every domain.
  • We can also formally give a proposal to the Govt of Assam, Dept of Education, SCERT etc. to include the dictionary in their reference book lists and if possible to explore the possibility of a formal collaboration with them by asking for fund allocation for it. Then it can go in a mission mode if we could appoint a few full time lexicographers to add lexicon in the database.
1) Simple sentence construction- 1st person+VERB, 2nd person+VERB, 3rd person +VERB in all tenses including singular and plural form

2) Homograph and Homophones in Assamese, CA may to include frequently used homograph and homophones with their examples-

   e.g a) বাঙলী আমার এজন্য গাইব নাম।

   b) গানটো গাই গাই চাইকেল বাবাই কমলা বজাবলৈ বুলি ৈৈল ৈল।

3) Uses of Reduplication in Assamese examples with sentence- e.g.

   a) গানটো গাই গাই চাইকেল বাবাই কমলা বজাবলৈ বুলি ৈৈল।

   b) সি বেছি চাহ-তাহ খাই ভাল নাপায।

   c) হঠাতে সকলো খেলি-মেলি লাগি গল।
Pranab Sarma: केवलीयाचा शंकर स्थितीक बनाने दिव नोबाबोले "predictive" बनाने सूझिल सूझिल हाव (instead of saying it is not in our database).

Satyajit Nath: उदाहरण एटाके "predictive" बनाने सूझ केब्ल्या बुजाई दिलचस्तान.

यकौन शब्दांची सुद्धा सुदृढ असता नाही.

Pranab Sarma: धनबाहे शंका एटा सूझिल, मानवीयता.

उत्तर आहेल Not in database.

Do you mean......

ती एटा मानवीयता, एटा मानवीयता......

एकत्रित मानवीयता व रूझाच्या विवि विषय दिवेळी येथे येथे वेळा उदाहरणार्थ एकत्रित "मानवीयता"चा उदा.

एकत्रित केब्वला बनाने शंकर अस्तु "मानवीयता"च्या गंभीर येथे अनेक वेळा येथे येथे हात सिद्धान्त "मानवीयता" येथे किंमत not found in the database रुपे उदा आहे?

*** मानवीयता येथे तसेच मानवीयता येथे तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता तसेच मानवीयता

Satyajit Nath: धनवान्याची, दस्ता.

The feature exists, as shown in the screenshot above (attached), but additional data is needed to cover those related words, which users can contribute.

Pranab Sarma: सेटूने तेर्डली गटिकलें सूझिले.

मिळेला तपत आणि तेलेवाळ हाने not in the database रूपी अटाच स्थिर हाना?

Satyajit Nath: That's a good point. As far as I know, the xobdo database has many different elements, and also a rules engine. So there may be some partial matches that triggers those suggestions.

Perhaps the feedback provided can be clearer on that front -- what is known, what users need to add.

I will pass that on.

धनवान्याची, दस्ता.

Pranab Sarma: धनवान्याची सद्धिचित. मानवीयता भागाते रुझान दस्ता नाही. केवलीयाची, विषयांच्या भागात, खडा चाल not found in the database दस्ता आपूर्तिक नाही.
Dear Bikram Da,

It will be good if we can do awareness webinars on XOBDO in Educational Institutes as much as possible. Then in different ways students and teachers may be involved with XOBDO.

Thank you,

Dr. Buljit Buragohain
Hello Bikram,

Thanks for sharing the document. I didn’t see anything about Text To Speech software in the presentation. Was this an oversight or a purposeful decision?

Regards,

Suresh

Suresh Bazaj is the creator of https://hear2read.org NGO involved in development of Indic Language Text-To-Speech system. XOBDO.ORG is sharing it’s database with them to create and improve the Assamese system. Specially Dr. Probodh Bora and Mr. Satyajit Nath are providing valuable feedback to them. You can access the system here: https://hear2read.org/prototype/app.php
THANKS

admin@xobdo.org