License

Data Sharing Policy In the age of AI and deep learning, large amounts of data are the key to building high quality language models. While commercial entities have access to tremendous amounts of public and non-public data that they utilize for building models, the public initiatives 22 suffer from a paucity of data. Therefore, it is important for Bhashini's data sharing policy to level the playing field in the context of Indian Languages.

The Ministry of Science & Technology (MoST) has notified the National Data Sharing and Accessibility Policy (NDSAP) for open sharing of data created using public funds that are non-sensitive or non-personal (not explicitly classified as non-sharable). The policy shall apply to all data and information created, generated, collected and archived using public funds provided by the Government of India directly or through authorized agencies by various Ministries, Departments, Organizations, Agencies and Autonomous Bodies. The aim of the policy is to promote data sharing and enable access to government owned data for national planning and development.

In accordance with this, Bhashini will make available training and benchmarking datasets for development of AI models in Indian languages across domains and languages. Bhashini plans to use the **CC BY 4.0** license for its datasets to allow the free use of this data with attribution. Bhashini shall provide a hyperlink to the original data source for attribution, where possible. In certain cases, Bhashini may delay the release of the datasets outside the Bhashini ecosystem by no more than 6 months to provide a competitive advantage to startups and researchers within the Indian ecosystem.

To create datasets, pre-existing original language resources in Indian languages may be used. However, training datasets represent a significant value addition over the original data sources and will be licensed as **CC BY 4.0** by the creator of the dataset (Bhashini contributor). Bhashini will use original data sources only from the following licensing categories for the creation of datasets.

- Copyright free or public domain data
- Creative Commons or any open licensed data

- Government data: Based on the NDSAP policy Bhashini shall use language resources publicly available on the internet on government websites to develop training datasets. However, the Ministry/Department/Other Body shall have the option (on the Bhashini website) to request that a given data source not be used for the purpose of creating a training dataset. If “opt-out” option is exercised, data from the relevant data sources shall be excluded from the published datasets within 15 days (about 2 weeks) of receiving the notice.

- Publicly Available Licensed Data: Publicly available data refers to data that has been published for public consumption and is accessible online to the public. However, the copyright “owner” will have the option (on Bhashini website) to request that the data source not be used to create a training dataset. If “opt-out” option is exercised, data from the relevant data sources shall be excluded from the published datasets within 15 days of receiving the notice.

- Data with explicit permission from copyright holders: Bhashini will use privately held data only with explicit approval and according to the

terms. In the case of audio and video recordings privacy shall also be taken into account. If the recordings are available in the public domain, it shall be assumed that there is no additional impact to the individuals whose recordings have been used to develop datasets unless explicitly opted out. In the case of private recordings explicit consent to use the recordings shall be obtained at the time of data collection.

Link: https://creativecommons.org/licenses/by/4.0/