A STUDY ON EARTH OBSERVATION TRAINING DATA LANDSCAPE IN INDIA
A Study on Earth Observation Training Data Landscape in India

Author:
Dr. Sudhira H. S., Gubbi Labs LLP
www.gubbilabs.in

Editor:
Ms Susheela Srinivas, Gubbi Labs LLP

Design and Layout:
Ms Savitha Sekhar Nair, Gubbi Labs LLP

Photo Credits:
All images used in the report are openly accessible and made to use for reusability and modification. Some images are sourced from NASA and are as per their Image Usage Policy.

On behalf of the
German Federal Ministry for Economic Cooperation and Development (BMZ)

GIZ is responsible for the content of this publication
New Delhi, India
September 2021

Disclaimer:
The data in the publication has been collected, analysed and compiled with due care; and has been prepared in good faith based on information available at the date of publication without any independent verification. However, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH does not guarantee the accuracy, reliability, completeness or currency of the information in this publication. GIZ shall not be held liable for any loss, damage, cost or expense incurred or arising by reason of any person using or relying on information in this publication. GIZ makes no claims concerning the validity, accuracy or completeness of the maps nor assumes any liability resulting from the use of the information therein. The geographical map used is for informational purposes only and does not constitute recognition of boundaries or regions.
## CONTENTS

**PREFACE**

**EXECUTIVE SUMMARY**

1. **EARTH OBSERVATION IN INDIA**
   1.1. Geoportals And State Gis Initiatives
   1.2. EO Studies In India

2. **RESEARCH STUDY**
   2.1. Objectives
   2.2. Method

3. **EO TRAINING DATA LANDSCAPE**

4. **FINDINGS**
   4.1. Data Sources
   4.2. Willingness To Share Training Data
   4.3. Existing Training Datasets In India
   4.3.1. Derivable Training Data
   4.4. What Does It Take To Have A Data Sharing Ecosystem?
   4.5. Summary And Limitations

5. **CHALLENGES AND RECOMMENDATIONS**
   5.1. Challenges
   5.1.1. Portal For Data Sharing
   5.1.2. Ensuring Data Quality
   5.2. Recommendations
   5.2.1. Capacity Building
   5.2.2. Incentivising For Sharing Training Data
   5.2.3. Leveraging Citizen Science
   5.2.4. Drive through NDSAP and other Missions Under Pm-Stiac
   5.3. Summary

**REFERENCES**

**ANNEXURES**

Annexure 1: Draft of National Geospatial Policy and Guidelines
Annexure 2: National Data Sharing and Accessibility Policy (NDSAP)
Annexure 3: Open Government Data (OGD) License 26
Annexure 4: SpaceRS Policy-2020 and SpaceRS NGP-2020 26
Annexure 5: Study Method and Questionnaire 27
Annexure 6: List of Experts who Participated in Qualitative Discussions 30
Annexure 7: Tools and Libraries for Sharing EO Training Dataset 31
On behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ), the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH is implementing the project “FAIR Forward – Artificial Intelligence for All”. The project strives for a more open, inclusive, and sustainable approach to Artificial Intelligence (AI) on the international level.

The growing number of geospatial platform providers has led to a sharp increase in the supply of Earth observation (EO) data – especially satellite imagery. For example, the satellite data from NASA’s Landsat mission, the Sentinel satellites of the European Union (EU) or the available imagery from Indian satellites offer free and open data and services. The growing availability of EO data meets an expanding interest and potential real-world impact that Artificial Intelligence (AI) and Machine Learning (ML) offer.

However, often technical practitioners lack access to high-quality geo- and ground-referenced data to develop and train ML models. Existing EO training datasets do not sufficiently represent the Indian context but are skewed towards North America and Europe. The resulting models do not transfer accurately to different regions and produce biased or wrong results. Hence, the innovative potential of this technology is largely untapped. Open training datasets provide local innovators with the opportunity to test new applications and products by reducing the cost of developing a prototype. This minimises the main barriers to AI utilising EO technologies and creates a potential market for tech innovators and social entrepreneurs. Also, it enables the public, private and voluntary sectors to access the latest AI technology more freely.

With this in mind, one of the primary goals of the FAIR Forward: AI for All project is to provide open, non-discriminatory, inclusive training data and open-source AI applications for local innovation. This objective rolls into developing sustainable and scalable modes of data collection that produce easily accessible and locally relevant EO training datasets and models for Indian users in a consistent, unbiased, privacy-sensitive, and cost-efficient way. To do so, an evaluation of existing efforts to create EO datasets in India is needed, and hence this study is undertaken.

Based on this study, while the Indian policy frameworks support sharing datasets, the challenges and opportunities for enabling an ecosystem of training data sharing are identified and discussed. In addition, a few training data creation and sharing approaches are also recommended.

ML and EO are two complementary areas that can provide faster solutions to more complex problems: for example, calculating the expected yield of crops in a particular area, detecting pests on agricultural fields, monitoring deforestation, or predicting the likelihood and extent of a global disease outbreak. Therefore, open EO training data and models have great potential to make EO technology more inclusive and enable millions of people to access services they cannot use yet – be it in agriculture, education, health, or any other field.
EXECUTIVE SUMMARY

The present study attempts to capture the existing landscape of EO studies in India along with the numerous geportals and geographic information systems (GIS) initiatives by the respective state and central government agencies. One key aspect emerging from reviewing these geportals is that data quality attributes are not spelt out clearly. Besides, the data distribution policy is also not appropriately stated. Thus, although some portals share data, when data quality aspects are not stated, the reliability of using such data becomes questionable. Moreover, the absence of a data distribution policy leads to uncertainty in using them either for non-commercial, commercial, or other purposes.

Further, primary research on the EO training data was carried out by a combination of structured interviews, discussions and complemented by secondary research. In India, although there is the availability of homegrown EO data by ISRO, notably the LISS-III, LISS-IV, Cartosat-series, among others, their usage has been limited to the government agencies and a few in the academia, with very little being accessed by the private sector. Instead, academia and the private sector (including start-ups) have mainly used NASA’s Landsat or EU’s Sentinel data products for use cases and studies.

Some of the key efforts that gathers training datasets are captured, but it revealed that most of them are not shared, or if it has been shared, the distribution policy (license) is not spelt out. It appears there is only one effort in the country by the Space Applications Centre, ISRO, that has attempted to build an application to gather EO training data. The app has a web-based data visualisation, including downloading feature.

Interestingly, on the willingness to share EO training data, most within academia are open to sharing training data, while a few have reservations and look for incentives to share them. However, in the private sector (including start-ups), there is a clear requirement that unless they can recover the cost of data-gathering resources and instruments, it would not make sense to put them out in the open. On the other hand, the response from one source in the government indicated that there are certain concerns in sharing such training data, with one of them ascribing to strategic reasons. However, the majority were open to sharing EO training data; and, with the SpaceRS policy draft in the offing, the landscape of sharing EO training data is expected to change.

Although two-thirds of the respondents indicated a willingness to share training data, there seemed to emerge some key concerns on why practitioners do not share. Notable among them are:

- There is no appropriate platform or portal where one could post such training data
- There is a lack of quality standards for EO training data and how to share it efficiently
- Lack of incentives for those who share training data to justify financial and time resources of sharing

Given the outcomes from the study, some specific recommendations are:

- A targeted capacity building among geospatial professionals on data sharing
- A mechanism for incentivising the data shared
- Citizen science activities to enable a broader and scalable way of collecting training data

Clearly, with the increasing access to EO data and the availability of cloud-based computational infrastructure to analyse EO data using some ML algorithms, the outlook on EO data sharing is promising. However, it is imperative that the right amount of nudge by the government and complementary support by GIZ India towards enabling this can go a long way in achieving sustainable development goals.
India has been one of the few countries to foray into space and launch a host of Earth Observation (EO) satellites. For over four decades, various payloads on the satellites have captured voluminous data on earth systems and beyond. The rise in the number of geospatial data providers has led to a steep increase in EO data availability, particularly the satellite remote sensing imagery and numerous ground-based weather sensors. Starting with the data from NASA’s Landsat missions, the European Union’s Sentinel satellites, and our ISRO’s IRS missions have propelled the availability of medium resolution multispectral satellite remote sensing data to a great extent. In addition, there has been a growing interest in applying Artificial Intelligence (AI) and Machine Learning (ML) methods for automating several computation processes with the availability of cloud-based computing and enhancing the broader application of these in various domains like agriculture, conservation, planning, and many development-related projects.

In particular, the application of EO data has been attempted since the beginning of the Indian space program (Kasturirangan, 1985). Although Indian remote sensing data were available only from the 1980s, the hard copies of the images were being accessed and visually interpreted before that time. Moving forward, as the availability of computing resources complemented the availability of more EO data, these applications increased progressively (Roy et al., 2017; Townshend et al., 1991).

The geospatial domain in India has gained momentum and has been well adopted across many organisations – the government, private and academia (universities). Figure 1 gives a high-level snapshot of the state of the geospatial data landscape in India. The list of government, private, and academic organisations is not exhaustive, but it certainly captures the leading ones. A notable aspect is that while many organisations are engaged with the geospatial domain, several government organisations are making considerable efforts to bring out geoportals for different uses and applications. For example, Bhuvan by the National Remote Sensing Centre under the Department of Space is popular and often termed India’s response to Google Earth. However, despite hosting a wide range of thematic maps and access to specific raw satellite data, Bhuvan is often considered user un-friendly. The overall user experience it offers seems to limit its wider usage.

In a significant move, the Department of Science and Technology (DST), under the Ministry of Science and Technology, which oversees the Survey of India, came out with a new draft of the National Geospatial Policy and guidelines (See Annexure 1). Until recently, the earlier mapping policy restricted anyone apart from the Survey of India to produce geospatial data or maps within the country. A key highlight of the new draft policy is that it has paved the way for organisations to generate geospatial data and publish them. Earlier, DST released another forward-looking policy — the National Data Sharing and Accessibility Policy (NDSAP) — through a Gazette notification (See Annexure 2). This policy is significant because ample data collected by various government organisations, including academia, needs to be shared appropriately. The Ministry of Electronics & Information Technology (Meity) has been made the implementing agency, and accordingly, it has created the Open Government Data (OGD) Platform for India (https://data.gov.in/). It now hosts key data shared by the various government departments across the country. More significant is notifying a Government Open Data License (See Annexure 3) to share and distribute such data under the NDSAP. Notably, the training data collected by various publicly funded organisations are to be shared under this license on the OGD Platform.
### Department of Science and Technology, Govt of India
- Survey of India
- National Atlas and Thematic Mapping Organisation (NATMO)

### Department of Space, Govt of India
- National Remote Sensing Centre
- Space Applications Centre, Ahmedabad
- Regional Remote Sensing Service Centres

### Forest Survey of India
- Geological Survey of India
- National Bureau of Soil Survey Land use Planning; Soil and Land use Survey
- Central Water Commission
- Central Ground Water Board
- State GIS & Remote Sensing Application Centres

### OTHER RESOURCES
- Bhuvan: bhuvan.nrsc.gov.in/bhuvan_links.php
- Vedas by SAC, ISRO: vedas.sac.gov.in/en/
- e-Nakshe - Survey of India: soinakshe.uk.gov.in/Home.aspx
- BHARATH Maps: bharatmaps.gov.in/
- NSDI: nidiclearinghouse.gov.in/GeoportalNSDI/Home.aspx
- Bhoomi: bhoomigeoportal-nbolup.in/
- Krishi: krishi.icar.gov.in/
- Bhukosh: bhukosh.gsi.gov.in/Bhukosh/Public
- India WRIS: indiawris.gov.in/wris/#/
- NATMO: geoportal.natmo.gov.in/
- IIRS: iirs.gov.in/geomweb-services

###組織 & Geoportals

![India Geospatial Data](image)

**FIGURE 1: INDIA - GEOSPATIAL DATA LANDSCAPE.**
As a next step, the Department of Space has also come up with drafts — ‘Space-based Remote Sensing Policy of India (SpaceRS Policy 2020)’ and ‘Norms, Guidelines and Procedures (NGP) for implementation of SpaceRS Policy 2020’ (See Annexure 4). This is a huge step towards having more private players participate in space programs, to the extent of even launching private remote sensing satellites. Hopefully, this should pave the way for a greater impetus to the start-up ecosystem and private players to harness this. However, although the NGP of SpaceRS policy 2020 notes that up to 5 meters and coarser would be easily accessible on a ‘free and open’ basis, this is yet to be implemented. Importantly, the policy doesn’t mention anything on training data, and hence sharing them.
1.1. GEOPORTALS AND STATE GIS INITIATIVES

In India, there have been multiple efforts by different organisations to share geospatial data products. Most of these have been put out through the respective geoportals of some of the key organisations or by the state-level GIS initiatives. As shown in Figure 1, there are more than ten known efforts at the national scale. A common feature in all of them is that they render the political boundaries at state, district, and block/taluk levels, followed by major transportation networks, water bodies, and in some cases, the recent land cover map by NRSC. Most such data are, however, available through Bhuvan as well.

Interestingly, some states have attempted to have many aspects under one umbrella (Figure 2), particularly Tamil Nadu and Karnataka. In Tamil Nadu, the TN-GIS is anchored by the Tamil Nadu e-Governance Agency, ensuring that vital geospatial data required by multiple departments are served through TN-GIS alone, reducing significant duplicity of efforts. It boasts of hosting about 348 spatial layers and is used by over 450 departmental stakeholders.

Similarly, in Karnataka, the Karnataka State Remote Sensing Applications Centre (KSRSAC) works as a critical nodal agency on the state’s GIS and remote sensing aspects. On the lines of TN-GIS, Karnataka GIS has several geospatial layers curated and rendered for viewing and exploration. As a next step, K-GIS has also now shared some of the base layers, particularly the administrative boundaries in the public domain for downloading in KML/SHP file formats.

While it is noteworthy that some of these efforts combine a variety of geospatial data and some are open for distribution (apart from viewing, as in most geoportals), it is also a shortcoming that the distribution policy (licensing) and data quality are not spelt out or are left to the user’s interpretation. In either case, although there appears a lot of geospatial data (which indeed are), from the context of EO open training data, there is little or no such data available from these portals.

1.2. EO STUDIES IN INDIA

A critical aspect of classifying satellite remote sensing data using supervised classification methods is the availability of training data or signatures corresponding to the land cover features required to be classified. Furthermore, to expedite the classification process, spectral libraries are defined for some of the non-dynamic features, particularly soils and minerals (Bellinaso et al., 2010; Sanchez et al., 2009). These spectral libraries serve as a quick reference to classify EO data. This also minimises the need for collecting new training data for classifying.

In the recent past, there have been sporadic attempts to develop spectral libraries for soil and minerals (D. Gore et al., 2016; Shepherd & Walsh, 2002; Viscarra Rossel et al., 2016), urban areas (Herold et al., 2004; Kotthaus et al., 2014; Nasarudin & Shafri, 2011), agriculture crops (Awad et al., 2019; Jain & Bhatia, n.d.; Krishnayya, 2007; Nidamanuri & Zbell, 2011; Rao et al., 2007; Schmedtmann & Campagnolo, 2015), and even mangroves (Prasad et al., 2015; Selvaraj et al., 2020; Somdatta Chakravortty, 2013).

There have been attempts to characterise dynamic land cover features like biodiversity and vegetation changes at regional and national scales (Gillespie et al., 2008; Pettorelli et al., 2014; Reddy et al., 2013, 2016; Roy & Tomar, 2000; Sudhakar Reddy et al., 2016, 2017; Turner et al., 2003) and even fire incidences (Reddy et al., 2017) using remote sensing. Studies have also been carried out to characterise crops using remote sensing (Dadhwal et al., 2002; Mekonnen & Hoekstra, 2011; Rama Rao, 2008). In addition, developing spectral signatures for classifying crops using hyperspectral (Awad et al., 2019; Krishnayya, 2007; Nidamanuri & Zbell, 2012; Rama Rao, 2008; Rao et al., 2007) and machine learning-based approaches have been made (Zhang, He, et al., 2019).

There are already several research publications that have attempted to harness Google Earth Engine in the Indian context, on urban (S. Agarwal & Nagendra,
2019; Goldblatt et al., 2016), agriculture (Aneece & Thenkabail, 2018; Dong et al., 2016; Gumma et al., 2020; Srinet et al., 2020), and wetlands (Amani et al., 2019; Tiwari et al., 2020).

Added to the tremendous geo-computation capabilities is the availability of machine learning algorithms on Google Earth Engine (Cho et al., 2019; Gumma et al., 2020; Hird et al., 2017; Shetty, 2019; Srinet et al., 2020; Zhang, Okin, et al., 2019; Zhou et al., 2020). While all these developments have essentially paved the way for renewed focus in harnessing the earth observation data, a crucial aspect for applying a host of machine learning algorithms are training data corresponding to the intended land cover feature and time. The above has necessitated all players – the government, private and academia – to employ their means and methods to gather such training data. However, once these training data are generated and used for analysis, they often don’t make it to the public domain, resulting in others needing such data to redo or duplicate their efforts. In addition, some of these training data are used to create spectral libraries, particularly for non-dynamic land cover features like soil and minerals. Apart from that, there are no open spectral libraries for most dynamic land cover features, mainly vegetation, crop types, and urban areas. It is thus imperative that if training data is available under an appropriately licensed open data repository, several duplication efforts will reduce. Furthermore, they will pave the way for more such analysis using the cloud-based geo-computation facilities applying a several of machine learning algorithms.

Against this backdrop, an evaluation of existing efforts to create EO datasets in India is undertaken. In the following chapters, we discuss India’s EO training data landscape based on structured interviews and discussions; it is followed by a discussion on challenges and opportunities for enabling an ecosystem for EO training data sharing in India, including specific recommendations.
FIGURE 2: THE STATES AND UTS GIS PORTALS - HTTPS://STATEGISPORTAL.NIC.IN/STATEGISPORTAL/
With India offering a vast canvas of applications for EO datasets, the need for relevant training data for using various ML methods for analysis has gained increased importance. Furthermore, with the larger goal of achieving the Sustainable Development Goals (SDGs) by 2030, there is a greater impetus for using and applying EO data. This can support and bring positive change in achieving them.

FAIR Forward is committed to improving the conditions for Indian developers and EO experts to use geospatial data and ML to promote sustainable development. The long-term goal of GIZ and its partners is to develop sustainable and scalable modes of data collection that produce easily accessible and locally relevant EO training datasets and models for Indian users in a consistent, unbiased, privacy-sensitive and cost-efficient way.

2.1. OBJECTIVES

The primary objective of this research is to identify and evaluate existing (open) EO datasets available for India. The key activities include:

- Mapping and describing existing EO training datasets in including ongoing efforts to collect and share ground reference data in India
- Identifying and evaluating approaches for creating EO training datasets in India
- Mapping and analysing the significant challenges of EO practitioners in India in the area of ML for sustainable development
- Outlining recommendations
  - for improving the conditions for Indian ML developers and EO experts to use geospatial data and
  - for promoting future sustainable EO data collections in India

2.2. METHOD

Primary research was carried out using structured interviews and discussions and complemented by secondary research to assess India’s prevalent EO training data landscape.

The study method and the set of questions asked during the interviews and discussions are provided in Annexure 5. The list of experts who participated in qualitative discussions are mentioned in Annexure 6.

The interviews were planned by short-listing and contacting about 25 experts who have been either in academia, practitioners or in the government. The researcher interviewed the respondents at a time suggested by the latter and engaged in the discussion. The context, background and questions were shared with the respondents prior to the interview. Primarily, the survey aimed to gather any ongoing efforts on collecting training data, their quality, willingness to share and other issues concerning them. Seventeen experts responded to the interview and engaged in qualitative discussions. Table 1 indicates the break-up of the domains represented by the respondents.

<table>
<thead>
<tr>
<th>Domain</th>
<th>No of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working in Academia</td>
<td>5</td>
</tr>
<tr>
<td>Working in the Private sector</td>
<td>3</td>
</tr>
<tr>
<td>Worked with the Government</td>
<td>3</td>
</tr>
<tr>
<td>Working with Non-Governmental Organisation</td>
<td>3</td>
</tr>
<tr>
<td>Advocacy</td>
<td>1</td>
</tr>
</tbody>
</table>
EO TRAINING DATA LANDSCAPE

Analysis (Gorelick et al., 2017; Moore et al., 2011) has become a gamechanger. In addition, there have been several cloud-based computing and geospatial data infrastructures like Open Data Cube (https://www.opendatacube.org/), Open EO (https://openeo.org/), SEPAL (https://sepal.io/), and Sentinel Hub (https://www.sentinel-hub.com/) to name a few leading ones (Gomes et al., 2020). This has enabled several large-scale and rapid analyses of satellite remote sensing data accessed on the cloud. Besides, Google’s cloud-based infrastructure Earth on Amazon AWS is also available for cloud computation. In the past few years, this has also accelerated outputs and enabled researchers to take up more ambitious analyses that were otherwise computationally intensive.

With many such platforms now available, the need for EO training data to train and develop ML models is more than ever. Availability of adequate training data is imperative for applying any supervised classification methods. Although there are no specific criteria on the amount of training data required for analysis, a rule of thumb is to have at least ten times the number of variables (classes) (Maxwell et al., 2018). However, it is also a function of the classification algorithm used, the number of input variables, and the spatial characteristic of the area to be mapped.

Another crucial aspect that can affect the classification of EO data is the quality of training data. If the training data are gathered from field visits, they are mostly accurate. However, those generated through thematic maps or composites using sources like Google Earth are prone to have inaccuracies as the composites are primarily created based on the best available cloud-free data, and they need not necessarily correspond to the time EO data is analysed. Such variations can affect dynamic land cover features like croplands, whether they were cultivated or not during that season, for instance. Yet, for large scale studies and in cases where field visits are not possible, training data is derived from available composites or thematic maps.
METHOD OF TRAINING DATA COLLECTION

Training data collection for classifying satellite remote sensing data is done either through extensive field visits or from secondary sources (Bakker et al., 2001), based on available maps/satellite data without a field visit. In the former case, field visits are carried out in the region of interest using a handheld location device that facilitates geocoding points of interest or landscapes based on the land cover or land use. Ideally, every distinguishable land cover or land use feature is marked as a point (in some cases, polygons are also created depending on the extent) and geocoded through the handheld location devices. For instance, if there are paddy fields and plantations, a point/polygon feature is marked, and relevant attribute details are noted. Likewise, if there are inaccessible landscapes (like rocky outcrop or a distant waterbody) but can be ascertained from the handheld device, point/polygon features are marked appropriately. Typically, about 100 such points for each distinguishable land cover type are gathered, to qualify as training data. These are primarily in .GPX or .CSV file formats.

When field visits are not possible and analysis is being carried on historical data, training data can be generated using a combination of maps and available satellite data (using true/false colour composites, as appropriate). For example, if thematic maps on geology or land cover maps indicating cropping areas or water spread areas are present, random points can be generated from such maps corresponding to the respective land cover feature. These random points can be then used as training data as well. Since the training data is derived from a past classified or thematic map, their reliability will also be as good as the classification accuracy of respective maps.

There are also instances of creating training data set by image interpretation using popular satellite true-colour composites (like Google Earth, Google Maps – Satellite, Bing Maps – Aerial, Mapbox – Satellite, to name a few). However, in these cases, since the composites are generated over time, their interpretation may be time-sensitive, resulting in data quality issues.
The following findings emerged from the EO landscape and training datasets based on the interviews and the qualitative discussions with experts. At the outset, it also revealed some of the challenges and opportunities that can enable creating a sustainable ecosystem for sharing EO training datasets in India.

A highlight of the study was meeting with Prof. K. VijayRaghavan, the Principal Scientific Adviser (PSA) to the Government of India, who leads many initiatives and oversees key futuristic policies on science, technology and innovation for the country.

The Government of India is committed to open up data as appropriate and the new SpaceRS policy once notified, sets the right guidelines and allows liberalised use of EO data infrastructure. This promises unlimited potential for industry, academia and start-up ecosystems to make use of EO as well as open data to enable many innovative applications that can contribute to larger societal benefits.

- Prof. K. VijayRaghavan, Principal Scientific Adviser (PSA) to the Government of India.
4.1. DATA SOURCES

As can be seen from the summary of responses (Table 2), almost all of them have been using NASA’s Landsat, followed by European Union’s Sentinel along with Indian Remote-sensing Satellite’s dataset. In addition, some have used Spot, Aster, and other hyperspectral datasets. Only those in the private sector gather additional EO data through ground-based sensors, which they have deployed privately. Barring that, academia is mainly relying on EO data made available by NASA and the EU. In most cases, it was also revealed that, unlike in the past, there is now little or no money budgeted for acquiring satellite remote sensing data since the relevant ones are now made available by the respective agencies. In India, although there is the availability of homegrown EO data by ISRO, notably the LISS-III, LISS-IV, Cartosat-series, among others, their usage has been at best only by the government agencies and a few in the academia, and very little by the private sector. The academia and the private sector, including start-ups, have mainly used NASA’s Landsat or EU’s Sentinel data products for various use cases and studies. Even with popular open-source geospatial applications like the QGIS, accessing Landsat, Modis or Sentinel data is accessible through plugins, but the same is not valid for IRS data products. This should alert the policymakers to recognise and address specific issues concerning the ease of access, usage, and distribution policies.

4.2. WILLINGNESS TO SHARE TRAINING DATA

Interestingly, most of them are open to sharing training data within academia, while a few have reservations and look for incentives to share them. Specifically, those in academia who are willing to share would be forthcoming
to share once they have published the results based on the data collected. One of them did suggest that, like how academicians publish their results based on data, they should be encouraged to publish data as a publication itself. Such data papers could have unique DOI and be treated on par with publishing in a journal. In a progressive move, one of the researchers shared that they are open to sharing such training data on request.

However, the private sector (including start-ups) has a clear view that unless they can recover the investment cost in resources and instruments for gathering the data, it would not make sense for them to put out the data in the open. While this applies to training data per se, some private players are also putting out some of their derived data products at no charge.

The response from one of those in the government indicated specific concerns in sharing such training data, one of them ascribing to strategic reasons. They also note that some of the policies have been shaped in the right direction enabling sharing of such data, particularly the NDSAP and the latest National Geospatial and SpaceRS Policy. However, even among these, there seemed to be a larger consensus on data sharing than having them held up internally, particularly when this is public-funded. Interestingly, there is less hope on its delivery on the state of the National Spatial Data Infrastructure (NSDI), which was initially envisioned as a clearinghouse for such data repository. Evidently, the accomplishments of NSDI have not been visible, and their contribution to the larger geospatial domain remains to be seen.

4.3. EXISTING TRAINING DATASETS IN INDIA

The creation or generation of training data for different land cover and land use is mainly gathered by field visits coinciding with the time when the satellite data was captured. As most of the instances of training data generation involve field visits, this involves more time, effort, and resources. Moreover, gathering and generating training data would be impossible in situations like extreme weather events or the ongoing pandemic. The training data essentially capture the point or polygon of the intended land cover feature requiring one to go around to all such locations. Typically, in academia, depending on the scale of the study, training data are gathered that are in the range of hundreds. However, since most training data also augment with verification from independent sources, their numbers are in thousands in the private sector.

The research was carried out through data gathered from secondary sources and publications to ascertain the availability of EO training data in India. Although this is not exhaustive, the key ones are tabulated in Table 3. Table 3 lists down the current available EO trained datasets about India in various sectors. Some of the key fields that attempt to capture the dimensions of the existing training data are:

- Whether training data shared? (Yes / No)
- Nature of training data (The thematic / topic on which the training data is collected)
As shown in Table 3, barring a few of them, most of the training data are not shared or not in the public domain. Some of the sources have considerable datapoints (> 1000) that can very useful for applying in ML models. Each datapoint is one signature and depending on the extent of area being classified hundreds of such datapoints would be required for training. This study has only found one effort in the country that has attempted to build an application to gather EO training data through an app and has a web-based data visualisation, including downloading features. This has been developed under VEDAS by the Space Applications Centre, ISRO (A. Agarwal, 2019). The website also states under its copyright policy that the “material featured on this site may be reproduced free of charge in any format or media without requiring specific permission”. However, the only aspect it lacks is an assessment of data quality.

In recent efforts, researchers have attempted to map and create the spatial data of the Indian grasslands/savannas, calling them open natural ecosystems (ONEs). The researchers have derived the training points from publicly available data from the National Remote Sensing Centre’s 2018-19 Land Use Land Cover (LULC) map to generate 181,812 points of ONEs and 116,447 points, not ONEs (Madhusudan & Vanak, 2021). Although the researchers have not yet shared the training data, the derived spatial layer of ONEs is shared through Google Earth Engine. There have also been large-scale efforts like biodiversity characterisation at the landscape level, where more than 16,000 data points have been used. However, when finishing this report, one of the authors informed that they plan to share the training data next year.

Apart from these, some of the prominent research labs and centres in universities have been generating training data and continue to do so. However, they are not shared. Since many of these have been used in several publications by these institutions, it would be worthwhile to share them to enable wider usage, particularly in the context of applying machine learning algorithms on earth observation data.

Further, the private sector has made considerable efforts in creating training datasets. However, they have been primarily catering to some commercial applications, notably on weather and crop information for insurance and other sectors.

### 4.3.1. Derivable Training Data

Perhaps, one of the ways to have a stop-gap mechanism such as a common repository that presents an open access and availability of training datasets and ensures standardised, clean datasets to be stored. This shall ensure the availability of training data for different land cover features from the most recent classified outputs or thematic maps by random sampling. A fraction of this can be used for validation and then ascertain the training data quality.

Table 4 lists some popular sources for deriving and sharing training data at scale considering the available country-wide data. Indeed, Dr. P. G. Diwakar who was heading the Earth Observation Systems at ISRO earlier suggest that since there are many derived and thematic data products on Bhuvan, one could use them and extract training data (signatures). For non-dynamic land cover like geomorphology (soil and other geological layers), one could access the Bhukosh – GSI portal data and derive the training data. However, what is often required are for dynamic land cover types like crop type and their growth stage, for instance.
### TABLE 2: SUMMARY OF INTERVIEW RESPONSES ON EO TRAINING DATA SHARING IN INDIA

<table>
<thead>
<tr>
<th>Expert</th>
<th>Domain</th>
<th>Type of organisation</th>
<th>Area/region covered</th>
<th>Method of training data collection</th>
<th>Availability of training data &amp; license</th>
<th>Willingness to share training data</th>
<th>Training Data Quality</th>
<th>Training Data Quality</th>
<th>EO Data Sources &amp; Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Parth Sarathi Roy</td>
<td>Biodiversity characterisation, Land cover change</td>
<td>Past Govt and University, now with NGO</td>
<td>All India, North-east in particular</td>
<td>In-person field-level data collection</td>
<td>On request, not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, Spot, ASTER</td>
<td></td>
</tr>
<tr>
<td>Dr T V Ramachandra</td>
<td>Land use land cover change, Urban, Water resources, Climate Change</td>
<td>Academic</td>
<td>Karnataka</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Some</td>
<td>Self-assessed</td>
<td>Landsat, IRS, Modis, ASTER, SRTM</td>
<td></td>
</tr>
<tr>
<td>Dr C. Jeganathan</td>
<td>Forestry, Climate Change, Agriculture</td>
<td>Academic</td>
<td>Central, North and North-east India</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, Spot, ASTER</td>
<td></td>
</tr>
<tr>
<td>Dr. Jagdish Krishnaswamy</td>
<td>Ecohydrology</td>
<td>NGO</td>
<td>Karnataka</td>
<td>In-person field-level data collection</td>
<td>Only one data paper, otherwise not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS</td>
<td></td>
</tr>
<tr>
<td>Dr Harini Nagendra</td>
<td>Urban, Land use land cover, Agriculture, Social</td>
<td>Academic</td>
<td>All India</td>
<td>In-person field-level data collection and through Google Earth</td>
<td>Not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, MODIS</td>
<td></td>
</tr>
<tr>
<td>Dr S. Pazhanivelan</td>
<td>Land use land cover, Agriculture</td>
<td>Academic</td>
<td>Tamil Nadu</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Based on incentives</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS</td>
<td></td>
</tr>
<tr>
<td>Mr Sarvesh Kurane</td>
<td>Agriculture, Climate, Land use land cover</td>
<td>Private</td>
<td>All India</td>
<td>In-person field-level data collection</td>
<td>Through Satsure Sparta</td>
<td>Yes</td>
<td>Self-assessed and independently validated through third-party sources</td>
<td>Landsat, Sentinel, IRS and ground-based sensors</td>
<td></td>
</tr>
</tbody>
</table>
### Findings

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Type of Organization</th>
<th>Area/Region Covered</th>
<th>Method of Training Data Collection</th>
<th>Availability of Training Data &amp; License</th>
<th>Willingness to Share Training Data</th>
<th>Training Data Quality</th>
<th>EO Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Parth Sarathi Roy</td>
<td>Biodiversity characterisation, Land cover change, Land cover change</td>
<td>Govt and University, now with NGO</td>
<td>All India, North-east in particular</td>
<td>In-person field-level data collection</td>
<td>On request, not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, Spot</td>
</tr>
<tr>
<td>Dr T V Ramachandra</td>
<td>Land use land cover change, Urban, Water resources, Climate Change</td>
<td>Academic Karnataka</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Some data through OSM</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, IRS, ASTER, SRTM</td>
</tr>
<tr>
<td>Dr C. Jeganathan</td>
<td>Forestry, Climate Change, Agriculture</td>
<td>Academic Central, North and North-east India</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Some data through India Biodiversity Portal, but training data is not in the public domain.</td>
<td>Yes</td>
<td>Relied on user-contributed data</td>
<td>Landsat, Sentinel, IRS, Spot</td>
</tr>
<tr>
<td>Dr Jagdish Krishnaswamy</td>
<td>Ecohydrology</td>
<td>NGO Karnataka</td>
<td>In-person field-level data collection</td>
<td>Only one data paper, otherwise not in public domain</td>
<td>Available on request, not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS</td>
</tr>
<tr>
<td>Dr Harini Nagendra</td>
<td>Urban, Land use land cover, Agriculture, Social</td>
<td>Academic All India</td>
<td>In-person field-level data collection and through Google Earth</td>
<td>Not in public domain</td>
<td>Some data through OSM</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS</td>
</tr>
<tr>
<td>Dr S. Pazhanivelan</td>
<td>Land use land cover, Agriculture</td>
<td>Academic Tamil Nadu</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Based on incentives</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, Spot</td>
</tr>
<tr>
<td>Mr Sarvesh Kurane</td>
<td>Agriculture, Climate, Land use land cover</td>
<td>Private All India</td>
<td>In-person field-level data collection</td>
<td>Through Satsure Sparta</td>
<td>Yes</td>
<td>Self-assessed and independently validated through third-party sources</td>
<td>Landsat, Sentinel, IRS, Spot</td>
<td></td>
</tr>
<tr>
<td>Mr Yogesh Patil</td>
<td>Weather &amp; Climate, Agriculture, Land use land cover</td>
<td>Private All India, but more for central Indian states</td>
<td>In-person field-level data collection and through Google Earth and a host of sensors for weather/climate data</td>
<td>Not in public domain</td>
<td>Based on incentives</td>
<td>Self-assessed and independently validated through third-party sources</td>
<td>Landsat, Sentinel, IRS, Spot</td>
<td></td>
</tr>
<tr>
<td>Mr Sajjad Anwar</td>
<td>Land cover</td>
<td>NGO US</td>
<td>In-person field-level data collection</td>
<td>Some through OSM</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel</td>
<td></td>
</tr>
<tr>
<td>Mr Thejesh GN</td>
<td>Open Data</td>
<td>Advocacy All India</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Dr R Prabhakar</td>
<td>Ecology and Conservation</td>
<td>NGO All India</td>
<td>User-generated data</td>
<td>Some through OSM</td>
<td>Yes</td>
<td>Relied on user-contributed data</td>
<td>Landsat, Sentinel, IRS, Spot</td>
<td></td>
</tr>
<tr>
<td>Dr Rajani MB</td>
<td>Archaeology</td>
<td>Academic Karnataka, Bihar, Chhattisgarh</td>
<td>In-person field-level data collection and through Google Earth</td>
<td>Not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS</td>
<td></td>
</tr>
<tr>
<td>Dr P G Diwakar</td>
<td>Land cover, Agriculture, Urban</td>
<td>Past Govt, now Academic</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>No, only for internal use</td>
<td>Self-assessed</td>
<td>IRS, Landsat, MODIS</td>
<td></td>
</tr>
<tr>
<td>Dr Vinay Kumar Dadhwal</td>
<td>Land use land cover, Agriculture</td>
<td>Past Govt, now Academic</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Yes</td>
<td>Self-assessed</td>
<td>Landsat, Sentinel, IRS, Spot</td>
<td></td>
</tr>
<tr>
<td>Mr Devdatta Tengie</td>
<td>Land use land cover, Agriculture, Urban</td>
<td>Private All India</td>
<td>In-person field-level data collection</td>
<td>Not in public domain</td>
<td>Based on incentives</td>
<td>Self-assessed and independently validated through third-party sources</td>
<td>Landsat, Sentinel, IRS</td>
<td></td>
</tr>
</tbody>
</table>

Note: # Sources listed in the order of data accessed.
<table>
<thead>
<tr>
<th>No</th>
<th>Source</th>
<th>Training data shared</th>
<th>Nature of training data</th>
<th>Area/region covered</th>
<th>Description of Labels</th>
<th>Method of training data collection</th>
<th>Any quality assessment</th>
<th>No of data points</th>
<th>Data sharing policy, if indicated</th>
<th>Link Source</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vedas by Space Applications Centre (SAC), ISRO</td>
<td>Yes</td>
<td>On Fodder and crop-related</td>
<td>Gujarat</td>
<td>Fodder crop type, Field Size, Crop Growth Stage, Crop Stress, Adjacent Crop</td>
<td>Field visits</td>
<td>Data accuracies are mentioned</td>
<td>&gt; 500</td>
<td>Distributed free of charge as per their Copyright policy <a href="https://vedas.sac.gov.in/en/copyright.html">https://vedas.sac.gov.in/en/copyright.html</a></td>
<td>Link</td>
<td>Only application from India that allows collection of EO training data.</td>
</tr>
<tr>
<td>2</td>
<td>Open Natural Ecosystems (ONEs)</td>
<td>No</td>
<td>Points as ONEs and not-ONEs</td>
<td>All India</td>
<td>Scrub Land, Degraded Forest, Barren Rocky Area, and Gullied Land as ONEs; Lands under cultivation that varied from horticultural crops and irrigated farmlands to marginal rained agriculture as not-ONEs</td>
<td>Derived from LULC map and high-resolution base maps in Google Earth</td>
<td>NA</td>
<td>181,812 points of ONEs and 116,447 points not-ONEs</td>
<td>Paper and ONE data under CC-By-NC 4.0</td>
<td>Link</td>
<td>A recent paper that has mapped open natural ecosystems in India.</td>
</tr>
<tr>
<td>3</td>
<td>Indian Council for Agricultural Research (ICAR)-Survey Data</td>
<td>No</td>
<td>Crop related information</td>
<td>All India</td>
<td>Location, identification details, subject level details, unit-level data</td>
<td>Surveys and field visits</td>
<td>NA</td>
<td>NA</td>
<td>Restricted under ICAR data sharing policy</td>
<td>Link</td>
<td>ICAR claims to have data but is restricted for general users.</td>
</tr>
<tr>
<td>4</td>
<td>Biodiversity Information System (BIS)</td>
<td>No</td>
<td>Vegetation Type map, spatial locations of road &amp; village, Fire occurrence</td>
<td>All India</td>
<td>Species-abundance values, measured environmental variables at plot level</td>
<td>Field visits</td>
<td>NA</td>
<td>16,000+ sample plots for entire India</td>
<td>Not mentioned</td>
<td>Link</td>
<td>Has geospatially referenced field sample plots.</td>
</tr>
<tr>
<td>5</td>
<td>Urban dataset</td>
<td>No</td>
<td>Points as urbanised or non-urbanised</td>
<td>Saharanpur, UP</td>
<td>Built-up and non-built-up</td>
<td>Derived</td>
<td>NA</td>
<td>900 built-up and 900 non-built-up (total 1800 points)</td>
<td>NA</td>
<td>Link</td>
<td>Training data created for 1995-2001 can serve long-term monitoring if shared.</td>
</tr>
<tr>
<td>6</td>
<td>A global reference database of crowd-sourced cropland data collected using the Geo-Wiki platform</td>
<td>Yes</td>
<td>Croplands as per GEOGLAM/JECAM definition</td>
<td>Global, including India</td>
<td>Cropland, whether used Google background imagery or viewed in Google Earth</td>
<td>Crowd-sourced through Citizen science</td>
<td>32,287 of 35,866 points were validated 4 to 7 times</td>
<td>35,866 sample units</td>
<td>Creative Commons Attribution 4.0 International License</td>
<td>Link</td>
<td>A crowd-sourcing campaign using the Geo-Wiki crowd-sourcing tool.</td>
</tr>
<tr>
<td>No.</td>
<td>Source</td>
<td>Data sharing</td>
<td>Method of aggregation</td>
<td>Nature of data</td>
<td>Quality indication</td>
<td>Link</td>
<td>Remains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>------------------------------------------------------------------------</td>
<td>--------------</td>
<td>-----------------------</td>
<td>----------------</td>
<td>-------------------</td>
<td>------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>A global dataset of crowd-sourced land cover and land use reference data</td>
<td>Yes</td>
<td>Land use, land cover</td>
<td>GEOGLAM/JECAM</td>
<td>66-80%</td>
<td><a href="https://doi.org/10.1038/sdata.2017.75">https://doi.org/10.1038/sdata.2017.75</a></td>
<td>A crowd-sourcing campaign using the Geo-Wiki tool.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Landuse Land Cover at EWRG, Centre for Ecological Sciences (CES), Indian Institute of Science (IISc)</td>
<td>No</td>
<td>Points of various land use and land cover types</td>
<td>Extensive for the Western Ghats</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Eco-informatics Lab at Ashoka Trust for Research in Ecology and the Environment (ATREE)</td>
<td>No</td>
<td>Points of various land use and land cover types</td>
<td>Extensive for the Western Ghats</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 5,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Tamil Nadu Agriculture University (TNAU)</td>
<td>No</td>
<td>Points on crop type and growth</td>
<td>Tamil Nadu</td>
<td>Crop type, growth stage, irrigation, etc.</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 1,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SatSure Ltd</td>
<td>No</td>
<td>Land use land cover, Agriculture</td>
<td>All India</td>
<td>Land cover and land use with greater emphasis on crop information</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>GeoSpoc Geospatial Services Pvt. Ltd.</td>
<td>No</td>
<td>Land use land cover, Agriculture, Urban</td>
<td>All India</td>
<td>Land cover land use type ranging from agriculture, forest, to urban</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Azim Premji University (APU)</td>
<td>No</td>
<td>Urban</td>
<td>Top 100 cities in India</td>
<td>Mostly on urban land cover and land use</td>
<td>Field visits and derived</td>
<td>NA</td>
<td>&gt; 500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>National Institute of Advanced Studies (NIAS)</td>
<td>NO</td>
<td>Archaeology</td>
<td>Select sites in India</td>
<td>On archaeological sites</td>
<td>Field visits</td>
<td>NA</td>
<td>&gt; 500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 4: TRAINING DATA – OPPORTUNITY FOR DERIVING AND SHARING.

<table>
<thead>
<tr>
<th>No</th>
<th>Source</th>
<th>Area/region covered</th>
<th>Description of Labels</th>
<th>Data sharing policy, if indicated</th>
<th>Source</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bhuvan</td>
<td>All India</td>
<td>Needs to be derived</td>
<td>NA</td>
<td>Link 1 and Link 2</td>
<td>Has access to a lot of thematic data and IRS satellite data products. It can be used to derive training data.</td>
</tr>
<tr>
<td>2</td>
<td>Bhukosh – Geological Survey of India (GSI)</td>
<td>All India</td>
<td>Needs to be derived</td>
<td>Not mentioned</td>
<td>Link</td>
<td>Has a host of geological and geomorphological data available for download for registered users. It can be used to derive training data.</td>
</tr>
<tr>
<td>3</td>
<td>Indian Council for Agricultural Research (ICAR)-Krishi Geoportal</td>
<td>All India</td>
<td>Crop residue burning, event date, satellite, instrument. Needs to be derived</td>
<td>NA</td>
<td>Link</td>
<td>Geoportal has maps on crop residue burning incidents derived from EO satellites, among other crop statistics. It can be used to derive training data.</td>
</tr>
<tr>
<td>4</td>
<td>Open Street Map</td>
<td>All India</td>
<td>Labelled data under landuse-forest, natural-wood, for vegetation/tree cover</td>
<td>Open data, licensed under the Open Data Commons Open Database License (ODbL)</td>
<td>Link</td>
<td>User-contributed data on OSM can be used as a proxy for some land cover, and training data can be derived.</td>
</tr>
</tbody>
</table>

#### 4.4. WHAT DOES IT TAKE TO HAVE A DATA SHARING ECOSYSTEM?

With two-thirds of the respondents indicating a willingness to share training data, some key concerns on this seemed to emerge. Notable among them are listed below:

- There is no appropriate platform or portal where one could post such training data.
- There is a lack of quality standards for EO training data and how to share it efficiently
- Incentives for those who share training data to justify financial and time resources of sharing.

#### 4.5. SUMMARY AND LIMITATIONS

The present study revealed some of the key perceptions on willingness to share training data. In addition, the study showed at least three top-level aspects for enabling an ecosystem for data sharing. A fundamental shortcoming of this study is that it is based on a small sample of academicians and practitioners (government and private). In the absence of a larger sample and limited responses, the inference drawn is bound to this set. However, efforts have also been made to gather and draw out from select secondary research on some aspects relevant to this study.
The current study reveals that the EO training data sharing landscape has some key challenges and opportunities. Adequately addressing the challenges and harnessing the opportunities would pave the way for achieving the intended sustainable development goals. However, specific concerns have emerged across academia, industry, and the government, as elucidated by the respective stakeholders.

5.1. CHALLENGES

5.1.1. Portal For Data Sharing

Most of those who wanted to share data expressed the need for a portal for data sharing. NSDI should have facilitated it; instead, the Open Government Data (OGD) portal serves this need for those emanating from the government sector. However, for those in academia, including NGOs and the private sector, the OGD portal does not encourage them. Instead of creating a portal from scratch, specific existing tools and libraries aid in using training data for ML applications, notably Development Seed’s Label Maker and ML Hub by Radiant Earth (See Annexure 7). There is also Collect Earth from Open Foris that can be considered.

Nevertheless, ML Hub by Radiant Earth is indeed promising, and it has already managed to curate some data for Africa. Perhaps, some push through specific capacity building workshops can enable the adoption of Radiant ML Hub. While there are domain-specific data repositories and some generalist repositories (Recommended Data Repositories | Scientific Data, n.d.), there exists only one such application for EO training data developed by VEDAS at Space Applications Centre, ISRO. However, without adequate push on the VEDAS data collection application, this may not see enough traction. It is thus crucial to ensure all EO training data generators share them in a common portal. Perhaps, the funding calls can spell such portals (like ML Hub or Vedas) and encourage grantees to share such data.

5.1.2. Ensuring Data Quality

As noted earlier, while there are multiple efforts to put out geospatial data through geoportals or state-backed GIS portals, there is very little or no information on their data quality. Even for those where some EO training data is shared, the assertion on data quality is lacking. It should therefore become mandatory to disclose the data quality appropriately should any such geospatial data be shared. There can be community-based guidelines and imposing self-policing of shared data to ensure data quality. The community-based approaches and policing work best when a large community has a shared interest, like the OpenStreetMap movement. However, in the absence of such a community, the standard data quality assurance
A data QA protocol can follow standard methods gathering statistics on – inconsistency, incompleteness, accuracy, precision, and missing/unknown values in the data.

In addition to the data quality, for any training data to be shared, their data and metadata standards information should be appropriately populated/documented, which often slips out in many instances. An unintended consequence of ensuring sharing through specific established EO data sharing portals like ML Hub would be implicitly enforcing adhering to particular data and metadata standards. Thus, there is no need to reinvent, but following the Open Geospatial Consortium (OGC) standards is most appropriate. Members develop the OGC standards to make location information and services FAIR – Findable, Accessible, Interoperable and Reusable (OGC Standards and Resources | OGC, n.d.). In this context, and the larger context of this initiative, it is appropriate to adhere to the Catalogue Services standard and specifications published by OGC (Catalogue Service | OGC, n.d.).

5.2. RECOMMENDATIONS

5.2.1. Capacity Building

It is a misnomer to think of universities as centres of research alone. Instead, they need to have a shared sense of purpose to cater to changing times and set goals for transforming societies and pushing the frontiers of knowledge. Besides these, they ought to create the suitable capacity - grooming individuals to become doers and leaders akin to nation-building.

The academia, while largely open to sharing training data, is unsure of where to share them. Of course, this is true for all those who want to share training data. However, there also emerged that the more considerable nuances of data sharing (standards, quality, or licensing) might also have to be addressed.

As most of them are keen to share the data, it may be appropriate to build capacity to provide directions on some of the best practices and a few existing methods to share training data. For instance, an orientation workshop on ML Hub of Radiant Earth can pave the way for such a movement. However, instead of a specific orientation workshop on ML Hub, it can be combined with a training workshop on Google Earth Engine or Python-based tutorials, which can also throw light on how to use such training data and use them in ML algorithms. A series of such workshops can be planned over a year spread across different parts of the country.

5.2.2. Incentivising For Sharing Training Data

The industry segment comprises seasoned geospatial players and relatively newer entrants, like the start-ups, plays a crucial role. Many of them are already generating a host of training data for various use cases, like agriculture insurance, flood inundated areas, assessing crop types and their yield, weather, and climatic aspects. Specifically, the start-ups are a new breeze and taking on the geospatial industry with innovative ideas and solutions that enable their intended customers. However, across the industry segment, they have specific concerns about EO training data sharing. Essentially, the industry would have invested in generating this data on resources and/or equipment they feel the need to monetise or break even at worse. Therefore, it is fair for the private players to seek incentives, in the absence of which their sustenance can be a challenge.

There could be a commercial version of ML Hub equivalent, where the users can form a consortium and have participating members contribute data for a price. This can allow data exchange and save costs within the private sector by avoiding duplicating such data collection efforts. Another alternative is that if an external fund is available to incentivise the private sector, it could be explored only for those sharing the training data. However, it may have to be deliberated whether the private players forming a consortium and charging a fee for sharing training data is viable.

For those in academia, if sharing data conforming to specific quality and standards can assure them an equivalent of a publication, they could be satisfied. However, those in the private sector and any other independent or individual may expect monetary incentives. In such a case, a creative way of incentivising such data sharing may have to be thought of.
5.2.3. Leveraging Citizen Science

India has 742 districts (in 2021), up from 640 districts as per the 2011 Census. In most districts, there have been District Science Centres. For various reasons, they have been notional and with limited activities, mainly engaging with schools. However, there has been little public engagement at large. In addition, there exists a host of government-run educational institutions – primary, higher primary and high schools – in all these districts. High school students could be mobilised as part of larger citizen science activities. They can be oriented to gather data on different land cover features, including crops or water bodies which would become an excellent training data resource. An orientation workshop across each state can be initiated detailing the activity. The citizen scientists can be asked to submit or share such data through a portal like ML Hub. A vital aspect of this approach is that it is scalable and repeatable over the years. Also, it can save high costs for the government and the private sector, who would otherwise have to spend considerable resources in gathering such data. In such a case, they only pay their resources validating such crowd-sourced data. This would be practical outdoor training for the students, enabling them to appreciate what is around them and document. The latter may not have direct tangible benefits, but certainly, they would have intangible benefits.

5.2.4. Drive through NDSAP and other Missions Under Pm-STIAC

The Government of India made the correct moves to frame the right policies and even the licensing. The NDSAP, NGP and the OGD License are essential aspects of an ecosystem of open data sharing, particularly among the public sector and academia. Specifically, those in academia funded by public money are held accountable, and it is fair for them to share the data as appropriate. Although the NDSAP has been for a while, including the OGD portal, they don’t yet host training data. This would, however, require a nudge from the senior leadership levels in the government, either political or bureaucratic leadership.

It is also essential to note that while there are concerns due to neighbours keeping the larger goal on SDG and other development needs, the government should seriously consider nudging all concerned agencies to share data under NDSAP under the OGD license. One of the critical aspects of why Landsat and Sentinel are used widely is that they have addressed these concerns well, prompting the Indian geoportals and EO data providers to take a leaf or two from these.

In addition, there are several national-level scientific missions (like the Artificial Intelligence Mission) under the Prime Minister’s Science, Technology, and Innovation Advisory Council (PM-STIAC), anchored by the Office of the Principal Scientific Adviser (PSA) to the Government of India. Therefore, it would be prudent for the Office of PSA to anchor the collation of training data by all government and academic institutions through some of its missions, including those gathered through citizen science. Since it has overarching responsibility across all departments and ministries of the Government of India dealing with research/any aspect of science and technology, the Office of PSA is ideally positioned to anchor such an effort at the national scale.

5.3. SUMMARY

Clearly, with the increasing access to EO data and the availability of cloud-based computational infrastructure to analyse EO data using some ML algorithms, the outlook on EO data sharing is promising. However, it is imperative that the right amount of nudge by the government and complementary support by GIZ towards enabling this can go a long way in achieving sustainable development goals.
REFERENCES


ANNEXURES

ANNEXURE 1: DRAFT OF NATIONAL GEOSPATIAL POLICY AND GUIDELINES


ANNEXURE 2: NATIONAL DATA SHARING AND ACCESSIBILITY POLICY (NDSAP)

https://dst.gov.in/sites/default/files/gazetteNotificationNDSAP.pdf

ANNEXURE 3: OPEN GOVERNMENT DATA (OGD) LICENSE


ANNEXURE 4: SPACERS POLICY-2020 AND SPACERS NGP-2020

ANNEXURE 5: STUDY METHOD AND QUESTIONNAIRE

Study method

The study was carried out through structured interviews and discussions with a select list of professionals working with the government, academia, and practitioners working in the industry and start-up ecosystems. In addition, efforts were made to gather responses through DataMeet, a leading group of open-data enthusiasts in India.

Questionnaire

1. Which earth observation data are you using/access?
   [Please list]

2. How do you access the earth observation data?
   [Describe]

3. Are you aware of the data sharing and access policy of the data you are accessing?

4. What are you using/applying earth observation data for?
   • Research / Academic purposes
   • Training
   • Practice / Generating reports
   • Policy Advisory
   • No, not using/applying at the moment
   • Interested in using it in the future
   • Other: __________

5. What areas/domains are you using/applying earth observation data?
   • Ecology and Conservation
   • Land-use Land cover change studies
   • Water resources
   • Urban planning
   • Social systems – education, healthcare, etc.
   • Other: __________

6. Do you publish the results of your analysis for public (open) access?
   Yes / No

7. Do you also publish/share the data used for analysis for public (open) access?
8. If yes, under what license are you publishing the data?

[List]

9. If yes, please share the details of the website/source from where they can be accessed.

[List]

10. What is the quality of the data set?

On training data

11. Are you / your organisation involved in the collection of earth observation training data?

Yes/No

12. Are you making use of the earth observation training data for any machine learning / AI-based model?

13. If yes, please describe:

14. How often are you collecting the earth observation training data?

15. If yes (for 11), are you making the datasets openly and freely accessible?

Yes/No

16. If yes, where is it available?

[List]

17. If yes (for 15), what is the quality of the training data?

18. If no (for 15), do you have plans or are you willing to open/share the datasets?

19. Are there issues/challenges in making the training data open (accessible)?

Yes/No

20. If yes, please chose which of these challenges/issues are applicable:

- Technical
  - Lack of server space/resources (hardware)
  - Data format (interoperability)
21. What are the major challenges for sharing earth observation training data?

[Describe]

22. In your opinion, what measures can enable an ecosystem of earth observation and training data more shareable and accessible?

[Describe]

23. In your opinion, what is the future outlook for earth observation training data when made open/accessible?

[Describe]


[Describe]
# ANNEXURE 6: LIST OF EXPERTS WHO PARTICIPATED IN QUALITATIVE DISCUSSIONS

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Designation</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Prof K VijayRaghavan</td>
<td>Principal Scientific Adviser to the Government of India</td>
<td>Office of Principal Scientific Adviser to the Government of India</td>
</tr>
<tr>
<td>2</td>
<td>Dr K R Murali Mohan</td>
<td>Head, Frontier and Futuristic Technologies (FFT) Division</td>
<td>Department of Science &amp; Technology (DST), Government of India</td>
</tr>
<tr>
<td>3</td>
<td>Dr P G Diwakar</td>
<td>ISRO Chair Professor</td>
<td>National Institute of Advanced Studies and Former Director, Earth Observation &amp; Disaster Management, ISRO</td>
</tr>
<tr>
<td>4</td>
<td>Dr Parth Sarathi Roy</td>
<td>Senior Fellow</td>
<td>Sustainable Landscapes and Restoration, WRI India and Former Deputy Director, National Remote Sensing Centre, ISRO</td>
</tr>
<tr>
<td>5</td>
<td>Dr Vinay Kumar Dadiwal</td>
<td>Indira Gandhi Chair Professor</td>
<td>National Institute of Advanced Studies Formerly, Director, Indian Institute of Remote Sensing</td>
</tr>
<tr>
<td>6</td>
<td>Dr T V Ramachandra</td>
<td>Coordinator</td>
<td>Energy and Wetlands Research Group,</td>
</tr>
<tr>
<td>7</td>
<td>Dr C. Jeganathan</td>
<td>Professor and Dean (Research, Innovation &amp; Entrepreneurship)</td>
<td>Department of Remote Sensing</td>
</tr>
<tr>
<td>8</td>
<td>Dr Jagdish Krishnaswamy</td>
<td>Senior Fellow, Suri Sehgal Centre for Biodiversity and Conservation</td>
<td>Ashoka Trust for Research in Ecology and the Environment (ATREE)</td>
</tr>
<tr>
<td>9</td>
<td>Dr Harini Nagendra</td>
<td>Director, Research Center, and Professor of Sustainability</td>
<td>Azim Premji University</td>
</tr>
<tr>
<td>10</td>
<td>Dr S. Pazhanivelan</td>
<td>Professor and Head, Remote Sensing and GIS</td>
<td>Department of Remote Sensing and GIS</td>
</tr>
<tr>
<td>11</td>
<td>Dr Rajani MB</td>
<td>Associate Professor</td>
<td>National Institute of Advanced Studies</td>
</tr>
<tr>
<td>12</td>
<td>Dr R Prabhakar</td>
<td>Director</td>
<td>Strand Life Sciences and India Biodiversity Portal</td>
</tr>
<tr>
<td>13</td>
<td>Mr Sarvesh Kurane</td>
<td>Vice President</td>
<td>SatSure Ltd</td>
</tr>
<tr>
<td>14</td>
<td>Mr Yogesh Patil</td>
<td>CEO</td>
<td>Skymet Weather Services Pvt. Ltd.</td>
</tr>
<tr>
<td>15</td>
<td>Mr Sajjad Anwar</td>
<td>Data and Strategy</td>
<td>Development Seed, Formerly with Mapbox</td>
</tr>
<tr>
<td>16</td>
<td>Mr Thejesh GN</td>
<td>Co-Founder</td>
<td>DataMeet</td>
</tr>
<tr>
<td>17</td>
<td>Mr Devdatta Tengshe</td>
<td>Solutions Architect - GIS</td>
<td>GeoSpoc Geospatial Services Pvt. Ltd.</td>
</tr>
</tbody>
</table>
ANNEXURE 7: TOOLS AND LIBRARIES FOR SHARING EO TRAINING DATASET

Noting the challenges for creating the EO training dataset, there are some tools and applications that can be used to create the EO training dataset. Though they are specific to certain domains, mostly in ecology/natural history, some are listed below.

Vedas

At the Space Applications Centre, ISRO in Ahmedabad, an Android app and web interface application has been developed to gather EO training data (A. Agarwal, 2019). Perhaps, this is the only application in India dedicated to such EO training data collection and a web interface for the users to view and download. However, it is not apparent from the website and the publication on any sharing policy, although the portal allows for data sharing.

Collect Earth from Open Foris

A custom-built, open-source satellite image viewing and interpretation system called Collect Earth Online was developed by SERVIR. It is a joint NASA and USAID program in partnership with regional technical organisations worldwide and the FAO as a tool for use in projects that require land cover and/or land use reference data (Collect Earth Online, n.d.).
**Label Maker by Development Seed**

Label Maker generates training data for ML algorithms focused on overhead imagery (e.g., from satellites or drones). It downloads OpenStreetMap QA Tile information and overhead imagery tiles and saves them as a Numpy .npz file for easy use in ML pipelines (Label Maker Documentation — Label-Maker 0.9.0 Documentation, n.d.).

![Label Maker Examples](https://github.com/developmentseed/label-maker/tree/master/examples)

**Radiant Earth – ML Hub**

Radiant Earth ML Hub is the world’s first cloud-based open library dedicated to EO training data for use with machine learning algorithms. Designed to encourage widespread data collaboration, Radiant ML Hub allows anyone to access, store, register, and share open training datasets for high-quality Earth observations (ML Hub – Radiant Earth Foundation, n.d.).

![ML Hub – Radiant Earth Foundation](https://www.radiant.earth/mlhub/)