Position paper of the roundtable meeting on

Common Education Data Models for Schools

Organised on 6th July, 2022

The Raj Reddy Center for Technology and Society (RCTS), IIIT Hyderabad, organised a closed door brainstorming roundtable meeting on *'Common Education Data Models for Schools'*. The objective of this meeting was to have a brainstorming session with the grassroot NGOs, pedagogy experts, technologists, researchers and innovators on how to build common data models for existing demographic and education assessment data, to help with school management at every level of administration. The models developed can allow in the quantitative categorization of problems in the education space and aid the development and application of AI/emerging technologies-based solutions to problems in the education space.

This roundtable had tabled the different types of school data currently being collected, prevailing challenges, and explored possibilities to build new models. It also explored technology solutions that could be implemented using the school data. The three major areas discussed were:

a) What information is currently being collected in relevance to education in schools?
b) What are the hurdles/requirements for implementing a solution developed in schools?
c) What data-based insights or technological solutions could help in either school administration/education?

A. Existing Data Collected in Relevance to Education in Schools

The two largest datasets available publicly in the context of education in India are the ones collected by the Annual Status of Education Report (ASER) and National Achievement Survey (NAS). However, these two are datasets collected at a cluster or district/national level and not at an individual level. Few comparatively smaller scale data collection initiatives are the State Achievement Survey, School Application Data, School Scheme data, etc. which all also collect information at an individual level. To better complete our understanding of the students' physical learning environment, programs such as *"Mana Ooru Mana Badi",* also collect information regarding the infrastructural facilities that the students have access to.

The data collected through all the above systems can be segregated into three different levels based on what information is being recorded:

1. *Student Level*
    a. Demographic information about the students including their parents' socio-economic background, education backgrounds, etc.
    b. Student and teacher attendance.

2. *Administration Level*
    a. Nutritional Tracking: Information regarding support programs such as the 'National Programme of Nutritional Support to Primary Education" (NP-NSPE), which provide free food to the students during school days.
    b. Information regarding available school infrastructural facilities like classrooms, washrooms, labs, teaching tools, tables, chairs, etc.
3. *Outcome Level*
    a. Continuous Formative Assessments
    b. Monthly/Yearly Assessments

B. Problems and Challenges in the Existing Data Collection Systems

The main problems that any data collection effort faces can be divided into three distinct phases:

1. *Conceptualization phase*: The communication of the need for certain data collection is not communicated to all the entities in the pipeline. The data systems have to be integrated into the day-to-day workflow, along with systemized mandates and vision communication.
2. *Generation and data entry phase*: The data collection and storage procedures should be standardized.
3. *Utilization phase*: Should be in real-time and the appropriate analytics and indicators should be made available to the entities at all levels. The possible outcomes of changes in different metrics should also be communicated in a clear and easy to understand fashion.

Expanding on specific issues in the above three phases, the panel raised the following issues that require attention when designing new data collection efforts:

1. Addressing Missing Links in Data Collection Arising Due to Only Implementing "Need-Based" Systems:

Current top down solutions collect data only for known problems that we want to address. However student education is a highly complex problem where not all issues and their interdependencies are completely known. In order to more holistically approach the problem in the education space, we require comprehensive data collection efforts. These comprehensive datasets should also be analysed and monitored at a longitudinal level, to study the efficacy of different intervention efforts, to allow for better allocation of resources to solving the known issues while trying to detect any unknown issues or repercussions of the intervention efforts.

The comprehensive data collection effort should cover not just Government schools but also private schools to allow us to get a better picture of how different levels of resources are contributing to student education and well-being. This would require policy level efforts, to address administrative discrepancies in the current scenario where private schools operate in

a separate bubble where, the spectrum of data collected/considered of importance is quite different from that at government schools for example, health data which is collected in government schools under programs such as the "School Health Program" launched as part of Ayushman Bharat, does not officially apply to private schools and hence given very little importance in many private schools.

2. Lack of Necessary Importance Towards "Mental health":

Current data collection efforts do not account for information regarding the mental health of the students. This also extends to students who have visual, auditory, speech, or other physical impairments which for the most part are not recorded or catered to during solution design. This has led to some of the most vulnerable in our society being left behind by the broadly well received digital content programs. This is a significant issue both in the government as well as the private sector and requires particular focus in terms of solutions designers as well as policy makers to properly address.

3. Lack of Stress on Longitudinal Studies of Student Development:

A student's performance, measured in terms of marks/grades, is currently annually refreshed in a cyclic process, with the only carry over from one academic year to next being has the student passed all their subjects. The treatment of passing students with different performance levels to new topics of the same difficulty, acts to diminish the capacity of the good performers while leaving behind the poor performers. A longitudinal approach to student development and learning needs to be performed as standard practice, to ensure that every student learns to their full potential while filling in learning gaps created due to the current system. This would require the standardized recording, maintenance, and accessibility of student performance data and corresponding analytics over multiple years and across schools.

4. Importance of a Micro & Macro Approach:

The currently existing systems mostly record data for the purpose of sanity checking, missing out many features that would capture actual learning progress of students that the current NEP policy is interested in. This issue is not limited to just government schools but also plagues most private institutes too.

Thus, it is of utmost importance that the data collection processes be standardized at a micro level (from individual student/teacher), and the data aggregation and analytics dissemination processes be streamlined at a macro level (at the state/central). There must also be trusted third party validation & auditing of the data at multiple steps of the collection process, to ensure any information extracted from the data can be trusted to genuinely reflect the scenario of education in the country.

5. ***Lack of Standardised Data Linking Protocols Among the Different Data Collection Agencies:***

As mentioned earlier, multiple agencies are capturing different datasets in relation to student education such as the ASER, National Achievement Data, State Achievement Data, etc. However these datasets have not been designed to be linked in a standardised fashion, reducing the usability of the various datasets. Any standards developed, should also take into account the linking of student demographic information, parental involvement information, school management and environmental information with the learner achievement information, to allow for complete understanding of the current educational status in the country as well as allow for extracting maximum information form the various datasets collected.

C. Hurdles for Successful Deployment of Any Data Collection Solutions

With multiple data collection systems already in play, the current problem is one of plenty rather than scarcity. These different systems as previously specified have been designed to address different data collection problems with their own sets of goals and constraints. They have also been collected by a wide variety of personnel with varying levels of training, tools, access, and incentive to collect the data. These different factors lead to a multitude of issues even when comparing the common features in the different datasets. Below are the main hurdles towards deploying any data collection solution, that the panel raised at the roundtable:

1. ***Reliability of the Data:***

With most collected data passing through multiple agents with different values and incentives, authenticity of the dataset is a very important issue to consider. Many of the data collection systems in place, lack even simple validity and logic checks, leading to little safety checks to prevent deliberate data manipulation at the time of collection. The other main and non-deliberate cause of unreliable data is that many of the data collection systems currently designed have not taken into account the existing workloads and work-flows of the personnel collecting the data (especially when the party collecting the data is the teacher). This leads to data entry and collection being performed in an inconsistent manner, leading to unintentional data loss or mistakes. There is also a lack in standardized practices for collecting, storing, and maintaining data.

To address these issues, data collection and storage standards need to be communicated more clearly to the personnel on the ground and they need to be strictly enforced through automated validity checks and periodic trusted third-party audits to detect non-deliberate

mistakes and prevent deliberate manipulations. Employing dedicated personnel for the data collection process when possible and designing systems to better integrate into the day-to-day workflow is necessary.

2. ***Unqualified Human Resources:***

As explained above, overburdening the data collection personnel with solutions designed without taking them into account is a major source of non-deliberate mistakes in the data. However even when data collection solutions are designed to properly integrate with collection personnel work-flows, lack in training in usage of the collection tools, improper communication of the importance of the task at hand, lack of incentives, and deliberate malpractice could still affect data quality and reliability.

To address these aspects, dedicated resources and infrastructure need to be made available for data collection and maintenance. The system design should also account for ways to reduce the colouring of data due to the experience of the data collector. This can be resolved to some extent by making all data collection modules highly informative in both aspects of "*What is being collected?*" and "*Why it is being collected?*" The complete data collection pipeline should be designed to be robust to personnel changes through usage of both system design and policy making.

3. Data Access and Privacy:

Given the need to collect data at a micro level to better study the individual performances and the sharing of data across organizations, introduces the issue of data access and privacy. As such before even designing the collection process, the purpose of who, why and how the data will be handled must be envisioned clearly. The lack in such a process is also the main reason behind systems being designed with little though given to interoperability with existing/future systems and failure of unified systems being developed over the years.

The only way to address this issue is through adherence to strict access control and anonymization tools being built into any data collections systems being developed from the conceptualization phase, and these measures and their usage must be standardized at a policy level, to ensure widespread compliance.

4. Misinterpretation Due to Language Discrepancies:

The last issue raised was one specific to a nation such as India which has a very diverse range of languages. To accommodate this diversity, data collection systems are generally designed to work in multiple languages, introducing an avenue to misinterpret the same question during translation due to local dialect and language structures.

The language in which the datasets are collected also may vary from the one in which the final analysis is performed or in many cases (especially when looking at a state/national level), data collected in multiple languages need to be integrated into one language, introducing large avenues for misinterpretation of the data due to different interpretation of the data during translation. Therefore any multilanguage data collection system designed must take into account possible misunderstandings that may arise during data collection and aggregation and work on reducing such cases as much as possible. And in cases where such confusions cannot be eliminated, possible confusion must be clearly marked and communicated to the data collectors and analysers. Addition of supportive clarification questions could also be considered.

D. Possible Data-driven Insights or Tech. Solutions

The final item discussed by the panel was the aspects any possible solution must satisfy for widespread use, these are as follows: The data entry must be a daily life process, associating all student activity to individuals with anonymity and gamification in the telemetry and learning process to improve and lower the friction of student participation. NDEA (National Digital Education Architecture) certification requirement for the operation of the systems. To allow for data portability. (e.g.: like HIPAA). The content and systems should be regionalized (accessible in regional language and regional context) to encourage usage of the system.

1. ***Digitization and Automation of the Data Collection Processes:***

All large percentages of the data collection in government schools still take place as a manual process with physical records. Hence, the first task before considering building any new tech solution would be how to digitize the current data collection and storage processes with minimal friction for implementation. The developed systems must be able to satisfy three major factors affecting data collection, namely consistency, frequency and dynamicity. The digitized solutions need to also taken into account access to technology and internet connectivity and have alternative avenues to handle the lack of resources and seamlessly onboard these alternate data streams into the main data stream in order for widespread adoption of the digitization solutions. Automation of the data collection processes should also be considered to the highest extent possible to avoid manual errors (deliberate or non-deliberate).

2. ***Solutions Should be Seamless, Simple, User friendly, and Politically Agnostic:***

The lowest skill level of the end user should be taken into consideration when designing technology solutions for the problem at hand. The solutions should be self-evident in their

usage even towards minimally tech savvy individuals as much as possible with any required training minimized to the maximum extent possible to allow for easy onboarding of new users (be it data collectors, analysers, or management). The solutions should be designed to balance being politically agnostic to avoid being thrown out due to change in political favours while still being able to utilize political will as a tailwind to expand adoption of the system.

3. Cross Pollination of Data Collection Efforts:

Cross pollination can be considered in multiple aspects; firstly in terms of best practices of data collections can be adopted from other successful large data collection efforts. Secondly the data being collected as part of other efforts can also be onboarded as part of the new system to allow for quick access to larger populations. And finally look beyond education for possible sources for cross pollination for example, consider re-using the Ayushman Bharat Health Account (ABHA) ID, the unique health ID created under the National Health Mission (NHM) for the purpose of unique student and teacher IDs to avoid having to create multiple IDs for the same individual and also allow for future integration of the health and education information of individuals.

An important aspect to keep in mind for any cross pollination efforts is that any solutions developed must be able to operate as an individual system (catering to the current user base) until any multi-system integration becomes possible.

4. Personalized Learning Solutions:

Finally, as previously stated a key issue with the current education system is its "One size fits all" nature. Student assessment information should be expanded to allow for tracking personal learning beyond simple scores. Automated learning recommendation engines could then be used to allow students to learn by the learning levels rather than age group. Solutions designed must also keep track of engagement levels by collecting relevant data to see how each student is engaged in a class, to allow for better content design and suggestions. Such metrics may not only be used for personalised learning recommendation but also as teaching assistance tools, to better monitor and address student needs, such as with which topics and why (lack of understanding or effort).

Summary and Next Steps

The center has collaborated with one school-complex cluster to pilot technology interventions and explore possible tech solutions for teaching tools, designing holistic progress reports, formative assessments and such. As next steps the center will make an effort to understand all the current models, data sources, and build models that can ingest data from different sources. The center will involve end users from the start for repeated review during development.

We may start the model building process by drawing up requirements and challenges from the points this round-table has tabled. Look into existing data collection systems and attempt to develop a system that can utilize existing systems, fetch data, and has a backend model flexibility to allow for addition of features as they are required. Further algorithm development for projection and prediction in the proposed information management system with real time dashboards catering to the different parties in the data pipeline would be considered, to aid in generating a positive loop for encouraging individual acceptance and usage of any systems developed and deployed.

Participants:

We greatly appreciate the time and insights from our panelists:

Tanmay Mahapatra (Care India), Rohit Agarwal (Talent Sprint), SreeLakshmi Reddy (Key Stone Foundation), Venkatesh Datla (Creya Learning), Kiran Ng (Infosys Springboard), Sandeep Kavety (ConveGenius), Mahadevan NV (TCS-ION), Raj Janagam (AIC-IIITH) and few others.

Moderator:

Dr. Rohit Kandakatla (KG Reddy College of Engineering and Technology).

About

Raj Reddy Center for Technology and Society (RCTS) is an initiative of IIITH to enable research and emerging technology-led solutions for grassroot education and public health, with specific emphasis on rural areas. It will identify a few socially relevant themes and take up projects in them to create a sequentially linked amplification of impact. It will take a wholesome view of the problem addressing all aspects, technology and otherwise, through to realising value on ground. Sustainability and Scalability are the key concerns in all engagements at the Center. It aspires to be a global hub to bring the latest research to solve societal problems.