Response to RFI on the Existence and Use of Large Datasets To Address Education Research Questions

Issued by Institute of Education Sciences, US Department of Education

Provided by the DXtera Institute
May 31, 2022

This submission is provided by the DXtera Institute (DXtera), a non-profit consortium that is dedicated to collaboratively develop shared technology solutions and practices that remove barriers to create, exchange, utilize and optimize data from disparate systems or applications. DXtera is a membership organization based in Boston, MA that was formed in 2015 out of collective efforts by institutions of higher education in the U.S. and Europe. DXtera's membership includes all varieties of institutions of K-12 and higher education, non-profit organizations, private-sector entities, funders and other non-academic organizations.

The DXtera team, and our communities, have been actively engaging organizations in the U.S. and internationally to build a membership network of trust that is focused on developing solutions related to making more fit-for-purpose data available for research and development. We connect with thought leaders, public entities, and education facing organizations to share knowledge, provide leadership, and increase visibility and understanding of the broad societal benefits of access to open source technology solutions to the too common barriers in education delivery.

This response was developed and builds upon the growing interest of the membership of DXtera to address this significant failure in the current market and will be utilized as use cases for testing of concepts and deployments. The specific community at DXtera involved in this response is the **Open Authentic Data (OAD)** in Ed community.

The participants in the OAD community engage with organizations throughout the education and employment continuum, participate in and attend workshops and conferences, and connect with organizations providing resources and information about advances in educational and employment technology, and assist states and employers who want to benefit from the availability of OAD. The engaged participants in the DXtera OAD community include the following:

- John N. Gardner Institute
- University of Michigan
- University of Hawai'i System
- Federation of American Scientists
- Bay Path University
- Credential Engine
- Dallas College
- Community College System of New Hampshire
- Open University of Catalonia
- Riiid Labs
- Ascend Learning
- Indiana Commission for Higher Education
- Smart Republic
- Open Geospatial Consortium
- Ad Astra
- Strada Education Network
- Brighthive

The OAD Community was initiated by organizations involved across the educational data ecosystem with a clear goal to establish an open source environment that will enable growth in needed and trusted solutions. The purpose of this effort is to bring together through an open invitation of collaboration of the multiple parties engaged in all aspects of the ecosystem engaged with creating OAD. This includes existing networks, frameworks, education organizations, research and standards bodies to learn from, and work with each other, and provide global leadership in the creation of a safe, more trusted ecosystem of OAD innovation within the education landscape.

The direct work of the OAD is not done in isolation. This community directly supports the work of two international policy and practice efforts that are seeking to increase safety in the market for technology use in education. Our team is serving as experts and facilitators on the Broadband Commission's Data for Learning Working Group. In addition, the OAD community is also embedded with the work of the EdSAFE AI Alliance (ESAA) which is dedicated to establishing trust in the marketplace for the use of AI tools in education. To create this trust we need collectively established benchmarks and frameworks that can inform policy. These must be supported by access to large open data to enable the tools to be trained and tested against data that would address bias and efficacy issues. The OAD is serving as the working group of the ESAA to lead this effort internationally.

Responses to RFI Questions:

(Question 1) What public or restricted use education-related datasets are available for training students in data mining/machine learning methods? What training needs are not being met by the datasets that are currently available?

We will leave this question to others more experienced in AI / ML education.

(Question 2) What open or restricted use education-related datasets are available to train new artificial intelligence models or to test hypotheses using data mining/machine learning methods? What research needs are not being met by the datasets that are currently available?

Authentic data that tracks student academic trajectory and completion information is generally unavailable to the data scientists who could create the tools to identify and address inequitable conditions and optimize student completion. This is simply the reality of the need to protect individual privacy (FERPA, GDPR, local regulations, etc), but it means that countless millions of records exist that can not be openly or easily mined to help develop, train and test advanced analytics, AI algorithms, or machine learning that could make a difference.

In contrast, robust applications of machine learning require vast amounts of authentic data to train and test models, and today this is not readily available. Currently, new products can only be created by established data science companies with a customer base that generates sufficient data. Developing authentic, openly available datasets will help to level the playing field and allow data science to be used broadly by researchers and educational technology providers to increase student success.

There are a number of good data sets available covering student success and test scores at the aggregate level. However, the biggest issue is the lack of data at the individual student level (PII) (student demographics, courses taken, program enrollment history, individual homework, grades, test scores, etc.).

To further explore this, the DXtera Institute launched, in spring of 2020, its Open Authentic Data (OAD) community, as described above.

At around the same time we also launched our initial exploration into data synthesis and simulated populations based on the vast amounts of higher education data that our members who have deployed the DXtera Data Management stack now have easier access to. DXtera's DM solutions help an organization align data across numerous operational areas, and is predominantly focused on institutions of higher education. The data sources we are currently dealing with include Student Information Systems, Learning Management Systems, Admissions Systems, and Financial Systems. Data areas include, but are not limited to:

Applicant Education History
Application Status
Placement Test Results
Student Personal Info
Instructors and Assignments
Course Catalog Info
Degrees Awarded
Student Demographics

Program Catalog
Course Registrations
Institutional Org Structures
Student Billing
Program Enrollments
User Engagement
Learning Outcomes
Assignments and Tests

Assignment Submissions
Test Results
Account Structures
Financial Postings
Financial Commitments
Budgets
Financial Aid Catalog Info
Student Aid Awards

(Question 3) What work do researchers need to do to access, and then explore the quality of, an existing dataset before conducting research with it? What aspects of this work could be reduced or conducted just once so that future researchers can reduce the time needed to complete a research project?

In the higher education context, it is very difficult to access data that can be correlated across systems, at a useful grain that ties students and their academic progress to the educational events that they are experiencing, such as interactions with content, assessment results, and other activities related to their world as students, and future workers. The current state of the art relies on third parties, like Institutional Researchers, to try to prepare data for research or product consumption. This can introduce certain bias into the data as information is filtered, massaged, combined, or corrected along the way. (See DXtera blog Farm to Table Information Processing)

Concerning the time to access institutional operational data, DXtera's experience is that once a decision has been made by an educational institution to enter into an agreement with an external party, the average time to complete a formal data access or data sharing agreement is about 6 months.

It should be noted that in our experience most institutions of higher education, particularly non-R1 colleges and universities, do not have documented processes in place, nor data access or sharing agreements to start with. DXtera has developed its own boilerplate agreement that we offer to those institutions that have none of their own (the majority it seems). We believe there is a pressing need for guidance for educational organizations on how to enter into such agreements with third party researchers or companies.

(Question 4) How do researchers determine the validity of data elements within previously collected datasets? What challenges are frequently encountered related to how those data align to constructs of interest?

We can't speak to previously collected data sets. Our focus is on moving forward to get the right data, at the right grain, with sufficient links between learning experience, educational content and overall student success.

(Question 5) What are promising approaches to testing and improving the validity of metrics within large datasets, especially those datasets that are developed through interactions with education technology?

By generating Open Authentic Data directly from an institution's Operational data, and being open and transparent in the techniques used to generate that data, DXtera and its members hope to address some of the challenges of data validity.

We believe that to create open data sets that are fit-for-purpose for training, testing and demonstrating new advanced analytics, AI and machine learning tools it must have certain qualities:

- Operational - Useful data must reflect what is occurring at the most granular level within actual operational systems. It can't be based on analytics, reports or other aggregate transformations of that data that can introduce certain biases introduced by individuals or algorithms that are performing the transforms.

- Current - Useful data must be contemporary to its purposeful use. We can't effectively train, test or demonstrate tools in question based on data that is years or even months out of date. Global, local and individual dynamics such as shifts in economy, politics, culture, the influence of pandemics, climate change and the like are all important components of realizing authentic data. An open data infrastructure that allows organizations to regularly publish anonymized or synthetic data based on current data would be an ideal goal.

- Aligned - Useful data must be aligned across multiple operational sources, such as student information systems, learning management systems, content management systems, human resource systems, financial etc, across K12, higher education, further education and workforce, to allow researchers to have the largest possible cross section of useful data. To reduce bias, these alignments must occur as close to the operational systems as possible, before any data anonymization or simulation transformations are applied.

It will be critical, in an open data infrastructure, to allow researchers to inspect the algorithms and the nature of the noise being applied to create open data sets, and that those data sets are generated directly from granular operational student and institutional data rather than from data the has already been aggregated by other data scientists or reporting algorithms along the way.

(Question 6) How likely is it that existing datasets, especially those that come out of education technology, contain data that are valuable for researchers and of sufficient quality that research could be conducted with a high amount of rigor?

The situation is that there are terabytes to petabytes of information out there, but it is generally locked in vendor specific operational systems. Quality remains an issue in the systems themselves, but is much less of an issue at the atomic grain and cross-system alignment required for training, testing and demonstrating advanced analytics, AI and ML tools.

At DXtera we are successfully able to integrate with these operational systems, via a growing software repository of integration connectors, to gain access to and align SIS and LMS data (and data from other sources) that goes back years and even decades.

---

(7) To what extent do existing datasets capture enough information to address research questions related to diversity, equity, inclusion, and accessibility? What additional data should be collected to address these questions?

For higher education institutions, data relevant to the real-world issues that students are dealing with, like family care, needing to fit education around work schedules, transportation issues, access to technology, and other stresses are also generally unavailable. Data like this might exist in advising systems, but typically not in a way that can be aligned and correlated with other operational data.

Data bias is also an issue. DXtera's strategy of working directly with operational data essentially eliminates bias that might result from data aggregation or reporting techniques that require handling of data prior to release at the cost of generating a potentially sensitive data set. However, the selection of automated analytic or reporting schemas to apply to operational data also has the potential to introduce bias, as does the practice of using additive noise to improve data security.

It should also be noted that no matter how closely authentic synthesized data is aligned with the entities, attributes and relationships of real operational data, the operational data itself is still not free of bias. Real operational data can do no more than accurately capture the experiences of the actors (students, teachers, advisors, etc) in the educational process. Unfortunately, these experiences are rooted in institutionalized educational models that have evolved, or not, from centuries of predominantly white, male, western perspectives on how best to educate.

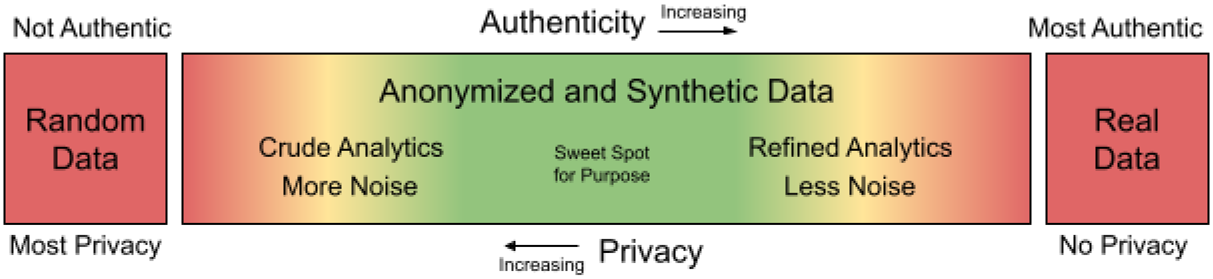
---

(8) What are the best practices for creating new datasets or linking existing datasets and sharing them with researchers (open or restricted use) while prioritizing the privacy of individuals and adhering to local, State, and Federal laws? What barriers and limitations exist?

To refine the usefulness vs privacy considerations of synthetic and anonymized data and in creating simulated populations based on real data we use two techniques:

1) Increasing refinement of analytics of real data increases the authenticity and fit for purpose of the simulated populations, but also increases the likelihood that a third party can potentially identify an individual. The most refined synthesis techniques amount to simply anonymization of PII.

2) Increasing the amount of “noise” or randomness applied to anonymized or analyzed data decreases authenticity and fit for purpose of simulated populations while decreasing the likelihood that a third party can potentially identify an individual.



When it comes to authenticity, we need to better understand what, if any, research or software development purposes can be served at various points along the scale from totally random to totally real. Combine this with the privacy concerns associated with data as it becomes more and more authentic, and we would expect to begin identifying “sweet spots” related to particular purposes. This, we believe, is an area of enquiry that requires more research and development.

Inauthentic/Totally Safe	Somewhere in the Middle	Highly Authentic/Not as Safe
This data is fit for very limited purposes. Typically testing systems at load or demonstrating basic user experience (think “lorem ipsum” for educational data).	Data that can be used for training advanced analytics, AI and ML algorithms to work on synthesized data. This is valuable for testing potential efficacy and demonstrating functionality.	This data (whether anonymized or real) can potentially be used to train advanced analytics, AI and ML algorithms to actually work on real data.

When it comes to privacy, we must also consider institutional perception. This is new territory, and we don’t yet have evidence of where along this authenticity/privacy spectrum serious PII concerns begin to arise. To that end, convincing the stewards of institutional data, who see themselves as protectors of student privacy, and driven by FERPA and GDPR concerns, to participate in open authentic data processes will be one of the most difficult challenges of developing significant simulated populations from real data. To make matters worse, institutions and instructors often consider educational **content** as proprietary (“special sauce”) and therefore requiring its own privacy practices. This is problematic when the purpose of the analytics or AI interventions relate to better understanding of the effectiveness of certain kinds of content towards educational success.

To date DXtera’s data simulations are focused primarily on student record data, and sit on the not-authentic but highly private/safe end of this spectrum. We have been relying exclusively on simple statistical analysis of key data, like the percentage of students who take certain courses or complete certain degree programs as members of a particular cohort. These statistics are then randomly distributed across the simulated population (which is currently not mapped against population demographics), with no regard for whether things like course trajectories make sense for meeting program requirements. This gives us data that is only fit for limited purposes, like load testing or customer demonstrations of certain functionalities.

(9) What role can IES play in developing infrastructure that supports the use of large-scale datasets for education research?

IES can be a catalyst for the required research and development and investments in open infrastructure that would spur this opportunity. Large open datasets must be supported by frameworks, practices and policies that are developed by participants from the affected market. Investing in an open community to establish a process for engagement and developing common practices and open infrastructure will be critical in this process. Providing a long-term commitment to funding the community and the needed technology solutions will be critical in bringing this important national issue to life.