Preliminary Findings of the Working Group on Data for Learning

The Transformative Potential of Data for Learning

Interim Report
September 2022

BROADBAND COMMISSION FOR SUSTAINABLE DEVELOPMENT

ITU

unesco

Preliminary Findings of the Working Group on Data for Learning

The Transformative Potential of Data for Learning

Interim Report
September 2022

Disclaimer

This interim report presents the preliminary findings from the first year of the two-year cycle of the Broadband Commission Working Group on Data for Learning, which will be further developed into a final report in September 2023. It has not been endorsed by the Broadband Commission for Sustainable Development and as such does not commit the organization to the findings or interim recommendations presented.

Acknowledgments

This interim report was written in a collaborative manner, with members of the Broadband Commission for Sustainable Development's Working Group on Digital Learning contributing their knowledge and expertise. The working group was chaired by Audrey Azoulay (Director-General, UNESCO) and coordinated by Borhene Chakroun (Director of the Division of Policies and Lifelong Learning Systems, UNESCO) who guided the production of this interim report.

The UNESCO writing team included Juliette Norrmén-Smith, Patrick John Devaney and Saurabh Roy. Other contributors to the writing include Wayne Holmes and Carmela Salzano. The interim report would not have been possible without the working group contributions of Valtencir Mendes, the guidance of the Broadband Commission Secretariat team including Anna Polomska, Leah Mann and Julia Gorlovetskaya, and the support of Kelsey Knutson. Editing, design and infographics in the interim report are thanks to the collaborative efforts of the Strategic Agenda team, including Kieran Colvin, Darcy Amber, Lucy Baines, Julia Spaeth, Gabriel Polley and Karen Wooddissee.

The views expressed here are not attributed to any one organization or individual, except in the case studies contributed by respective working group members, nor do the views necessarily reflect the position of the Broadband Commission members or their affiliated organizations. Not all working group members participated in every meeting that led to the content of this interim report. We wish to thank the working group's Broadband Commissioners, Commissioners' focal point members and external experts for their invaluable contributions, kind reviews and helpful comments.

Working Group members

Commissioners

1. Professor Mercedes Aráoz, Peru
2. Mr Bocar Ba, SAMENA Telecommunications Council
3. Ms Sima Sami Bahous, UN Women (represented by Diana Rusu)
4. Dr Lee Yee Cheong, International Science Technology and Innovation Centre for South-South Cooperation, UNESCO
5. Dr Rumman Chowdhury, Parity Responsible Innovation Fund
6. Dr Mr Mutlu Cukurova, Institute of Education, University College London
7. Mr Piotr Dmochowski-Lipski, EUTELSAT IGO (represented by Mr Sajn Petra)
8. Dr Qu Dongyu, FAO (represented by Mr Maximo Torero, Mr José Rosero Moncayo and Mr Henry Burgsteden)
9. Mr Amir Dossal, Global Partnerships Forum
10. Mr Erik Ekudden, Ericsson (represented by Ms Zohra Yermeche)
11. Dr Tedros Ghebreyesus, WHO (represented by Mr Derrick Muneene)
12. Baroness Beeban Kidron, 5Rights Foundation (represented by Ms Sonia Livingstone)
13. Mr Lacina Koné, Smart Africa (represented by Mr Osman Issah, Ms Aretha Mare, Mr Calvain Nangue and Ms Thelma Efua Quaye
14. Mr Mauricio Ramos, Millicom (represented by Ms Giuseppina Curreli)
15. Professor Jeffrey Sachs, Columbia University (represented by Ms Radhika Iyengar)
16. Mr Achim Steiner, UNDP (represented by Mr Gayan Peiris)

Experts

1. Mr Dale Allen, DXtera Institute/EdSAFE AI Alliance
2. Mr Alexandre Barbosa, Cetic.br, Regional Centre for Studies on the Development of the Information Society, UNESCO
3. Ms Tracey Burns, National Center on Education and the Economy
4. Mr Doug Casey, Connecticut Commission for Educational Technology
5. Mr Ashok Goel, National AI Institute for Adult Learning and Online Education
6. Professor Ms Ellen J. Helsper, Department of Media and Communications, London School of Economics
7. Mr Wayne Holmes, Institute of Education, University College London
8. Mr Shiloh Naiken, Department of Basic Education, South Africa
9. Ms Carol O'Donnell, Smithsonian Science Education Center
10. Mr Amit Prakash, Centre for IT and Public Policy, International Institute of Information Technology Bangalore
11. Ms Maria Florencia Ripani, Fundación Ceibal
12. Ms Joanna Rubinstein, World Childhood Foundation (former commissioner)
13. Ms Carmela Salzano, Global Partnership for Education
14. Mr Fabio Senne, Cetic.br, UNESCO Regional Centre for Studies on the Development of the Information Society
15. Ms Karine Sonigo, ILO
16. Ms Kati Tiainen, Microsoft
17. Mr Quentin Vidal, Mr Stéphan Vincent-Lancrin and Mr Jun Yu, OECD
18. Mr Werner Westermann Juárez, Programa de Formación Cívica
19. Mr Niall Winters, Department of Education, University of Oxford

Table of contents

List of tables and figures

Acronyms and abbreviations

5G	Fifth generation of wireless mobile telecommunications technology
AI	Artificial intelligence
API	Application programming interface
AR	Augmented reality

Cetic.br	Regional Centre for Studies on the Development of the Information Society
CTE	Cybersecurity career and technical education
DBE	Department of Basic Education
EdTech	Educational technology
EMIS	Education management information system
GDPR	General data protection regulation
ICT	Information and communication technology
IGO	Intergovernmental organization
IIEP	UNESCO International Institute for Education Planning
IPR	Intellectual property rights
IT	Information technology
ITU	International Telecommunication Union
LMS	Learning management system
MIS	Management information system
MIT	Massachusetts Institute of Technology
MOOC	Massive open online course
NDEAR	National Digital Education Architecture
NGO	Non-governmental organization
OECD	Organisation for Economic Co-operation and Development
OER	Open educational resource
PSET	post-school education and training
RUSA	Rashtriya Uchchtar Shiksha Abhiyan
SA-SAMS	South African School Administration and Management System
SDGs	Sustainable Development Goals
SIS	School information system
TVET	Technical and vocational education and training
UIS	UNESCO Institute for Statistics
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations Children's Fund
VC	Verifiable credentials
VICT	Computational Thinking for Persons with Visual Impairment
VR	Virtual reality
WGDL	Working Group on Data for Learning

Executive summary

The demand for data has grown dramatically in recent years, driven by a deepening dependence on digital technologies that require large amounts of data to operate. Although data use in education is nothing new, the COVID-19 pandemic's fast-paced propulsion of technology into teaching and learning processes has added both complexity and urgency to conversations about data in education.

The current international landscape of learning data lacks clarity, coherence and consensus on the true meaning of Data for Learning, including definitions of key aspects of Data for Learning ecosystems or a common vision for data governance in educational environments. A **multistakeholder consensus** is necessary to unleash the potential of Data for Learning and ensure the safe and efficient use of data as a tool to drive the transformation of education.

In the current landscape, **data are a double-edged sword**, offering both transformational potential to improve education systems, as well as numerous risks that need to be taken seriously. Data-fuelled technologies offer exciting possibilities for empowering educators, learners, schools and education policy-makers. Amid this promise, risks relating to data profiling, deterministic algorithmic interference, and learners' privacy and security, simply cannot be ignored. Both edges of the sword must drive the development of education data ecosystems to ensure the successful, safe realization of Data for Learning's transformational potential.

Responding to the need for a greater understanding of the data ecosystem in education, the Broadband Commission convened a working group of experts and Commissioners in January 2022 to discuss this double-edged nature of Data for Learning and define strategies to build bolder partnerships, actions and investments needed to advance the aspiration of transforming education. This interim report explores the potential assets and risks of Data for Learning, as well as the key challenges and tensions that must be considered.

The Broadband Commission Working Group on Data for Learning (WGDL) considered three keys to unlocking the power of Data for Learning and to making data a pillar of inclusive and quality education systems: (1) **infrastructure:** focusing on data infrastructure and learning ecosystems, (2) **capacities:** emphasizing data skills and competence framework for life and work; and (3) **governance:** considering ethics, governance, data flows and sovereignty. These keys structured the work of the group and were adopted as strands of work and reflection.

Within these three keys, the WGDL concluded that the following concepts are crucial to the development of **inclusive, diverse,** and **equitable** Data for Learning ecosystems.

- Regarding **infrastructure**, technical, semantic, organizational and legal **interoperability** between multisectoral stakeholders will be a key driver of success. Data used for teaching and learning should be **accurate, reliable** and of sufficient quality to be deemed **trustworthy**. The high **financial and environmental costs** attached to more technically advanced data systems should also be factored into the design of data-informed learning models.

- Regarding **capacities**, progress will be impossible without sufficient **digital and data literacy and skills**. Furthermore, meaningful **transparency, explainability** and **accountability** will play important roles, not only in safeguarding data subjects such as learners and educators, but also in paving the way for the **meaningful consent** required to ensure **privacy** and **security** in learning data ecosystems.

- Finally, regarding **governance**, it is of paramount importance that stakeholders build a shared understanding of data **ownership** that both supports innovation and the **common good** without infringing upon personal privacy rights or creating an undue concentration of power.

The WGDL concluded its work with five interim recommendations to policy-makers and stakeholders engaged in the education data landscape.

Firstly, there is a need to define a clear multistakeholder consensus on the need to develop a whole-of-government approach to implementing and unlocking the true value of Data for Learning. **Secondly**, a sustainable financing strategy grounded in multilateral partnerships is required. **Thirdly**, effective data literacy and skills development is required across all levels of the education ecosystem, from the youngest learner to the most senior education policy-maker. **Fourthly**, to ensure the realization of the transformational potential Data for Learning offers, efforts must target the enduring obstacles and challenges of education. **Finally**, a multilateral approach built on international cooperation and solidarity is needed to help bridge the digital divide that exists both within and across countries, to ensure the benefits of Data for Learning reach all learners, everywhere.

Background

Since its establishment in 2010 by the International Telecommunication Union (ITU) and the United Nations Educational, Scientific and Cultural Organization (UNESCO), the Broadband Commission has expanded both the breadth and depth of international dialogue on sustainable development, leading advisory work and advocacy for the transformational impact of broadband technologies on human lives. Working groups are at the heart of the Commission's work. With more than 30 groups to date, the Broadband Commission's working groups bring together stakeholders from all sectors to advocate for meaningful, universal connectivity and achieve its seven Broadband Advocacy Targets (Broadband Commission, n.d.). All working groups leverage the expertise and perspectives of a unique composition of membership, comprising some of the key players in the technology industry, civil society, intergovernmental organizations (IGOs), non-governmental organizations (NGOs), academia and government.

Education is a core focus of this work, and to date the Broadband Commission has convened seven working groups on the theme of education: Data for Learning (Broadband Commission, 2022*b*), AI Capacity Building (Broadband Commission, 2022*a*), Digital Learning (Broadband Commission, 2021*b*), School Connectivity (Broadband Commission, 2020), Child Online Safety (Broadband Commission, 2019), the four-year Working Group on Education (Broadband Commission, 2017) and Multilingualism (Broadband Commission, 2011). Together with the 2017 Working Group on Digital Skills for Life and Work, these working groups have convened industry leaders, government officials and civil society to address prominent issues specifically dedicated to the intersection of education and technology. Building on the work and research of these groups, the active WGDL is positioned as a key consultation group for international dialogue on data for education and training recovery, resilience and future development, with specific focus on Data for Learning.

Globally today, 2.7 billion people are still offline, 90 per cent of whom live in developing countries. In addition, there are significant differences in Internet affordability worldwide such that the poorest people often have the most expensive mobile data fees. Due largely to access and cost differentials, people in wealthy countries use, on average, 35 times more digital data than people in poorer countries. This digital divide within and across countries narrows opportunities for far too many young people and adults to engage in lifelong learning, fulfil their potential and contribute to sustainable development of their communities.

Inspired by the belief of the Broadband Commission that the digital divide is more than simply technological, this dialogue operates within the understanding that present gaps in access to broadband networks and new technologies, including data-driven technologies, are significant contributors to persistent and widening disparities across economies and societies. To explore the many dimensions and possible implications of digital divides on data use in education, the WGDL has focused on **three strands** related to Data for Learning: (1) data infrastructure and learning ecosystems, (2) data skills and competence framework for life and work, and (3) ethics, governance, national sovereignty and cross-border data flow regulation.

In monthly meetings since January 2022, the group has shared experiences and case studies on subjects related to education data ecosystems, such as the development of data-fuelled learning systems, interoperability frameworks, and ethics and inclusion in AI-driven technologies used in education and training. The group will continue its monthly meetings for the duration of its unique two-year cycle, releasing this interim report in September 2022 and culminating in a final report in September 2023.

Objectives and structure

The key objectives of the WGDL are to promote all learners' data protection, advocate for the democratization of data delivery through open data initiatives in education, propose an approach to financing models of investment in Data for Learning, explore linkages with other related initiatives from Broadband Commissioners and/or members of the WGDL, develop scenarios for future development of data-driven learning ecosystems, and connect to the Global Education Foresights and other foresight works that give evidence to policy-makers to tackle cross-cutting issues and build resilient policies for the future.

A primary objective of the final report, as outlined in this interim report, is to **map the evolving data landscape within a lifelong learning perspective**. To accomplish this, the present report integrates the three main strands of the WGDL into its synthesis of group discussions, seeks to stabilize a consensus on key definitions relating to data for education including terms and concepts, sets out a vision for the WGDL's final year, and provides preliminary guiding principles for data use in education for relevant stakeholders. In compiling this interim report, the WGDL has specifically aimed to:

- analyse and refine the concept of Data for Learning and establish coherent definitions within a lifelong learning and right-to-education perspective;
- examine differences across data for learners and learning, data for teachers and teaching, and data for administrators, managers, regulators and policy-makers;
- map the education data ecosystem and define a rationale for an ecosystem approach, the need for whole-of-government approaches and stakeholder dialogue, as well as multilateral cooperation and solidarity on Data for Learning;
- understand both the opportunities and risks associated with using data as a tool for teaching and learning, including using artificial intelligence (AI) and big data;

- propose a list of overarching recommendations for governments, including ministries and national regulatory and financing agencies, local education agencies, international organizations, donors, the private sector, NGOs, and civil society, investing in data systems and data-driven educational approaches.

The interim report is organized into five parts.

Part 1 explores, refines and defines Data for Learning, discusses why data are important to learning, and introduces the idea of Data for Learning as a double-edged sword.

Part 2 then explores the education data ecosystem, beginning by describing data uses of the past before describing the key uses and users of learning data in the digital age. The section concludes by examining the learning data ecosystem along the lines of the three strands of the WGD4L.

Part 3 delves deeper into the challenges and risks associated with Data for Learning, outlining key data issues that need to be accounted for to ensure fair and equitable implementation of Data for Learning policies.

Part 4 synthesizes the discussion of potential benefits and risks of Data for Learning, culminating with draft principles that should guide the development of policies for data use in education.

The final section, **Part 5**, outlines the future trajectory of the work of the WGD4L during the second year of its operation, including the production of a visualization of the education data ecosystem and the need to build an understanding of the diverse interactions that occur between multiple types of data and datasets, data producers and users, and stakeholders across the individual, local and global levels.

Examining the education data ecosystem

Introduction

Data ecosystems are expanding and evolving, propelled by the digital transformation of the Fourth Industrial Revolution and the acceleration of digitalization due to the COVID-19 pandemic. New information industries, powered by the technological capabilities of cloud computing, AI or the Internet of Things, increasingly influence decision-making across contexts. The data revolution has rapidly changed how and what services are produced and delivered, not only within industrial sectors but also within social sectors such as education, health and social security. The growing dominance of digital data in the educational landscape is influencing how learning is designed, discussed and delivered.

Datafication and digitalization are seeping into more and more aspects of education systems. Technological innovations have increased the ease with which data are captured, stored, processed and monitored. Education systems around the world, in different ways and at different levels of sophistication, have explored using such data-fuelled technologies to improve learning, teaching, administration, planning and management. Data-driven interventions in education – as well as popular discourse around precision education and personalized learning – grew dramatically alongside digital and hybrid learning during the period of COVID-19 disruptions. The prominence of these data-centred learning models has endured even as the pandemic wanes and most schools have reopened for in-person teaching and learning.

Despite the potential of data use in education to improve the policies, programmes and learning experiences at all levels of a system, **data in education is a double-edged sword**. On the one hand, data analytics can enhance the ability to evaluate multiple dimensions of learner and teacher competencies, facilitating the recognition and transferability of records across ecosystems, thereby improving both management decision-making and flexible learning pathways. Data provides new possibilities for monitoring and improving transparency in management information systems (MIS), which can be used as tools for empowerment of teachers, parents, schools and systems, and thereby improve the quality and effectiveness of planning and governance. On the other hand, data can be misused to the detriment of learner privacy and rights. The pace of developing data regulations and data literacies is slower than that of the expansion of data-fuelled technologies. Experiments with neurotechnology, biometric wearables, facial recognition systems and genomics give rise to new concerns about undue surveillance, deterministic algorithmic interference or data profiling (UNESCO, 2022*a*).

Furthermore, as increasing data practices sweep across multiple sectors including education, it becomes all too easy to lean too heavily on the insights data can provide, even though there are disparities between that which is easily captured by data and that which is not. Failing to fully grasp the limits of data as a technology for measuring and informing and incorporating that which lies beyond them into decision-making processes will undermine data's transformational potential for education. In fact, failing to understand the inherent partiality of data could lead to Data for Learning ecosystems causing more problems than they will solve.

The risks of Data for Learning are drawn along lines of existing inequalities and are exacerbated by the digital divide. Disconnected learners, marginalized learners, and women and girls risk being under-represented in datasets. As many educational technology (EdTech) programmes and platforms rely on large datasets to develop their tools, the misrepresentation or invisibility of certain learners in these datasets may result not only in ineffective data-driven tools, but also in a reproduction of broader social inequalities. Furthermore, systems with lower infrastructural or institutional digital capacities may prevent teachers, administrators and managers from tapping into the potential benefits of data. It is likewise in lower-capacity contexts where the digital divide often correlates with insufficient legislation or public awareness of how one's data can be used, which endangers the data sovereignty and security of the most vulnerable.

Box 1. Defining the digital divide and the data divide

The preliminary recommendations of this interim report are grounded in an awareness of the **digital divide** and **data divide** between and within countries, and the belief that transforming education does not only rely on high-tech solutions available in high-resource contexts. Many regions of the world are digitally disconnected or have limited digital infrastructures, which results in uneven potential for the digitalization of education. Under the digital divide lies a deepening data divide, which is defined as the gap that exists between those who can take advantage of the opportunities offered by digital data and those who are further left behind.

Although this interim report focuses primarily on the transformative power of digital data in education, Data for Learning extends beyond the digital including, for example, mobile, paper-based, radio and television data used for educational purposes. Moreover, the report also acknowledges that there are myriad types of data that are integral to teaching and learning processes that may not be captured accurately or completely in digitized data systems, and that pressures to collect and analyse data on all teaching and learning processes may lead to negative learner- and teacher-facing experiences and outcomes, some of which will be explored in this report. Finally, this interim report believes that the value of small, localized datasets should not be ignored in the face of big data, as learning data in qualitative, disaggregated forms also carry great potential for improving teaching and learning experiences.

The United Nations Road Map for Digital Cooperation raises concerns that issues of data are deeply intertwined with the digital divide, and calls for international cooperation to enable education data ecosystems to develop equitably and with due diligence. Today, only a few countries have adopted the frameworks and legislation needed to address concerns connected to privacy protection, regulatory control, data security and integrity. Few, too, have made the financial investments needed in infrastructure to leverage the educational, social and economic benefits of the data frontier. Protecting learner data will require sustained action to share knowledge and establish norms and standards for data security and use (UNESCO, 2022*a*). This will only be possible through public-private partnerships between data infrastructure and capacity providers and governments, for which the Broadband Commission has long served as an exemplar.

Lifelong learning and right-to-education perspectives

The WGD4L sets a broad scope, situating Data for Learning within broader **lifelong learning** and **right-to-education** perspectives. As such, learning should be understood not solely as a formal activity affecting children within the walls of a school building. Likewise, learners are not only children, but rather people of all ages and localities engaged in educational and training opportunities. The seventh International Conference on Adult Education made clear that the global community is taking steps towards affirming the right to education throughout life (UIL, 2022). Data play a critical role in supporting individual and informal learning pathways for adult learners and tertiary-level learners, especially with regard to supporting reskilling and upskilling for the future of work. However, given that school-based learning is the primary pursuit of most children and youth, the WGDL gives specific attention to the rights-based issues affecting young people's relationship to learning data.

From a human rights perspective, and in all settings, considerations of protection, privacy and security must be at the core of efforts to share data on vulnerable and at-risk groups, and to support accountability while protecting the safety of communities. The collection and use of data from learning spaces needs to be aligned with national data policies and regulatory frameworks, with consideration for all available legal protections. However, approaching Data for Learning from a rights perspective not only concerns the safety and security of learner information. As data and learning analytics play an increasingly dominant role in the discourse on individual learning pathways, especially for adult learners and those engaged in non-formal learning, the double-edged nature of Data for Learning is once again on display.

Against this backdrop, and in an effort to assist Broadband Commission members in their ongoing dialogue, the WGD4L aims to strengthen the foundation for further discussions by filling existing information gaps and mapping the different issues at stake in this debate. Three key issues include: (1) examining the **potential assets** and potential dangers of Data for Learning, (2) understanding the **drivers** and **barriers** to investment in Data for Learning, and (3) **anticipating the development** of the data ecosystem and its impact on education systems. This interim report strives to address each of these issues, and in doing so, the WGD4L will promote international cooperation, solidarity, and shared knowledge on safely and sustainably harnessing data to support lifelong learning pathways.

1

Defining Data for Learning

Developing a Data for Learning ecosystem that will deftly exploit the new possibilities offered by the digital transformation of education, while also avoiding the potential pitfalls and challenges related to the datafication of education, is a challenge that requires multilateral support and input. In this regard, a particular aim of this special interim report of WGDL is to define an accepted and common taxonomy for WGDL members and wider relevant stakeholders. The following section of the report sets out to achieve this objective.

Deconstructing digital data

Before examining the potential assets and dangers of Data for Learning, it is important to begin by defining what we mean by "data". The concept of data is often used synonymously with information or evidence, but given the volume, variety and velocity of digital data sources available in today's world, the word deserves a precise definition. In this interim report, we use data to mean **sets of discrete items of information such as numbers, text, images or sounds that are collected, cleaned, formatted, stored and shared, and used for analysis, calculation, inference and application**.

The United Nations Secretary-General's report *Our Common Agenda* stresses the Global Digital Compact to connect all people to the Internet, including all schools, avoid Internet fragmentation, protect data, apply human rights online, and introduce accountability criteria for discrimination and misleading content, while promoting the regulation of AI with digital commons as a global public good (United Nations, 2021).

> Data may be collected "by hand" or collected automatically; may be human-readable (such as handwritten figures in a ledger book), machine-readable (such as lists of numbers stored in an electronic ledger), or both; it may also be structured (such as databases of numbers), semi-structured, or unstructured (such as voice recordings). Datasets can be small or large, the latter categorized as big data. In other words, data may be thought of as a general purpose technology, like the steam engine, electricity or the Internet, which could "touch all aspects of societies and economies. But such sweeping changes are not automatic. The productivity value of the steam engine and electricity was realized decades after they were first introduced". (World Bank, 2021)

Like the game-changing use of steam power to turn the wheels of the Industrial Revolution, new uses of data have catalysed a "data revolution" driven by a "data economy" where value is derived from accessing, gathering, organizing and controlling information. This revolution is fuelled by a combination of big data, technological advances, the increasing use of AI, and growing numbers of individuals with basic or advanced data skills. With this revolution has risen a recognition that new approaches to data production and use can yield more dynamic insights that were previously unattainable with slower, traditional methods of information management.

Data and data analysis are in themselves nothing new. However, what has recently changed centres on scale and power: the vast amount of data that is constantly being created by emergent web, mobile and digital technologies in virtually all areas of human activity, and the powerful computers and data analytic techniques now available, have sharply reduced the costs of collecting, storing and using data. Today's mobile phones are more powerful than supercomputers were only 40 years ago (Bookman, 2017).

The vast amounts of data being created and collected are often referred to as "big data", which is distinguished from other data by exhibiting the so-called "V attributes" as follows:

Table 1. The V attributes of big data.

Volume	The size of the dataset is very large.
Variety	The different types of data are generated from multiple sources, needing to be cross-referenced and combined in order to be fully exploited.
Velocity	The data may be generated at a rapid rate.
Veracity	The data may be incomplete, influencing the precision of inferences made from it.
Volatility	The data being collected or inferred may become less relevant over time.
Value	The ability to extract value from such data while complying with given time, human and technical resource constraints.

Source: Adapted from du Boulay *et al*. (2018, p. 269).

Big data also "refers to things that one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more" (Mayer-Schonberger and Cukier, 2013, p. 6).

However, "big" is often conflated with "beautiful" or "better", usually without explicit justification:

> We tend to prioritize large data sets instead of small data sets. We believe bigger data sets will tell us more and will provide more accurate information than smaller data sets. But that is not always true … it's not binary – you can have big and small. Both tell you something, but neither tells you everything. The second thing is we tend to think that collecting more data will solve the problem. Here's a problem, let's collect some data on it and then we'll get closer to a solution. But in some cases, that is actually not the best thing to do … We call this the paradox of exposure.
> (Klein, 2020)

Countries around the world are investing heavily in their data-processing infrastructures in the hopes that they can transform the ever-growing piles of big data into actionable, real-time information. Big data is increasingly described as the fuel that drives businesses and organizations forward, so far as to claim it "the world's most valuable resource" (The Economist, 2017). However, without being broken down, analysed and transformed into usable information, data are crude and has little value. In fact, the processing costs – both financial and environmental – of transforming data into useful information are high. Big datasets are rarely democratized, and thus often must be purchased by those seeking to train machine-learning models that consume copious amounts of energy to deliver effective insights. There can be "data spills" (Thorp, 2012) when personal information is inadvertently leaked, giving rise to privacy and security concerns, which are even more serious when considering young learners' data.

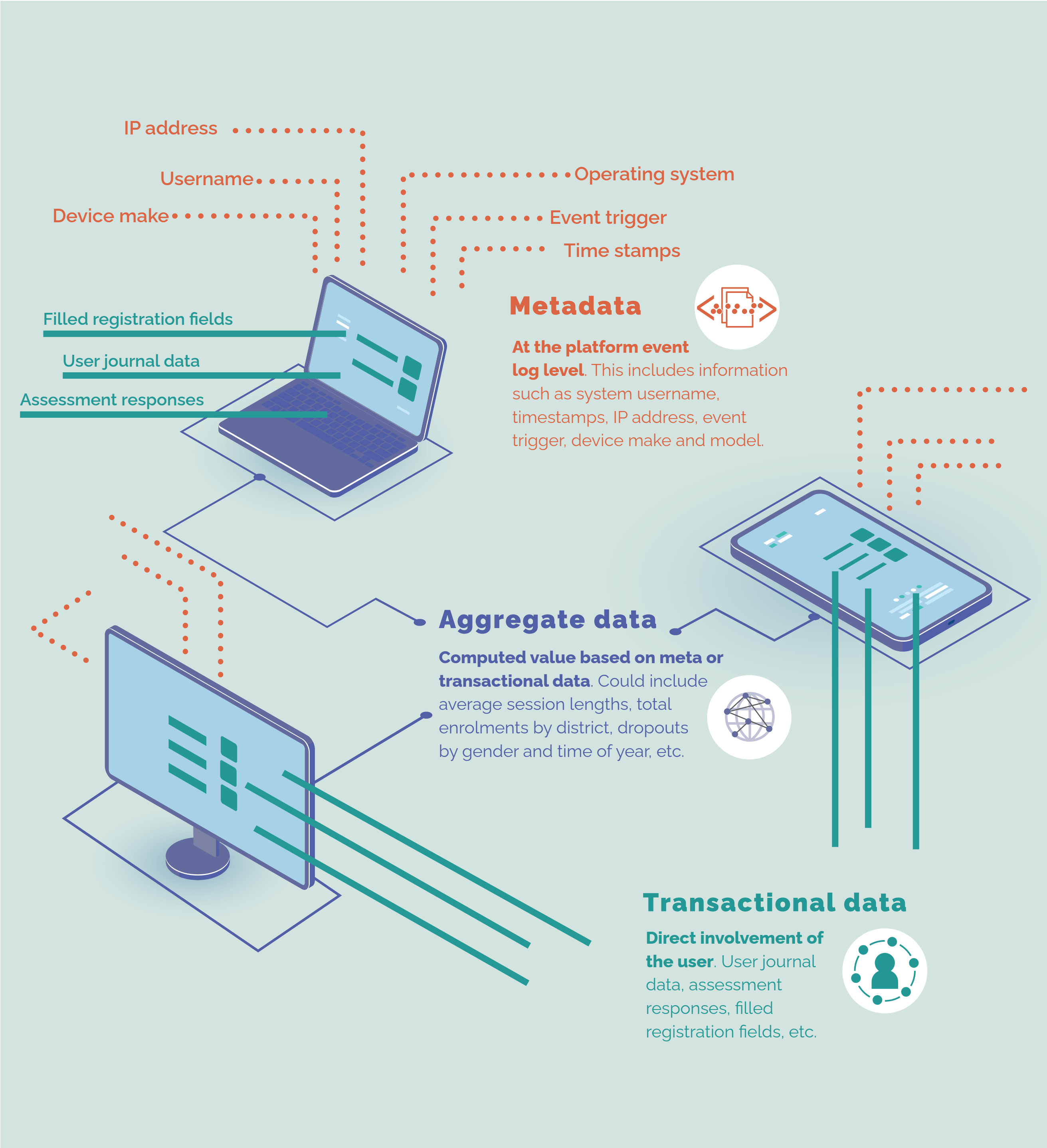
Given that data must be broken down to be understood, this report breaks down data into four categories to explore its potential affordances and risks: **metadata, digital transactional data, aggregate data** and **synthetic data**. These four types are explained in the table below, along with their potential uses in educational contexts.

Table 2. Types of digital data and common uses in learning spaces.

Type of data	Description	Examples	Potential uses for learning
Metadata	Metadata is data about data. Metadata provides information about other data, including descriptive summaries of information ("descriptive metadata") about a resource to aid discovery and identification.	Any data streams that do not involve the end users directly, e.g. digital platform logs, digital traces, clickstream data and smartphone sensor data.	Often used to operationalize and understand knowledge, cognitive strategies, and behavioural processes in order to personalize and enhance instruction andlearning.
Transactional data	Information captured and recorded about an event, which typically includes time, numerical values, and references to one or more objects.	Any data expressions directly generated by users, e.g. journalling, social media posts and online discussion forum comments.	Often used by institutions to understand how users interact with a website (time stamps on site traffic, popularity of topics, language used in comments, etc.). Often analysed with natural language processing techniques to relate linguistic features to cognitive, social, behavioural and affective processes.
Aggregate data	Individual-level data from multiple sources that is combined and summarized for the purposes of examining trends, making comparisons, reporting, etc.	Institutional data, student demographic data, graduation and enrolment rates, school standardized test performance scores, etc.	Often used to inform administrative decision-making; can be used to improve course enrolments or student engagement through data analytics, as well as cases of AI-powered use like course guidance systems and predictive systems.
Synthetic data	*Can exist at any of the above levels.* Data that mimics real-world data, generated by using sophisticated AI models to create whole new datasets from scratch.	Any dataset that does not exist in the real world, which can be applied to mimic any type of data, from insurance data (Hann, 2021), to self-driving vehicles (Behzadi, 2021), or even patient health care records (Walonoski *et al.*, 2017). Developers can train cars on virtual streets and can supply synthetic human faces on demand.	Often used to supplement or supplant real-world data with "better", "cheaper" or "bigger" datasets constructed using AI (Koperniak, 2017; Lohr, 2018) to (1) lower the cost of developing helpful AI algorithms, (2) improve the diversity of datasets to counter implicit bias or invisible data in "real" data, and (3) provide better privacy protections and lower the use of sensitive personal data, such as children's data.

Source: Authors.

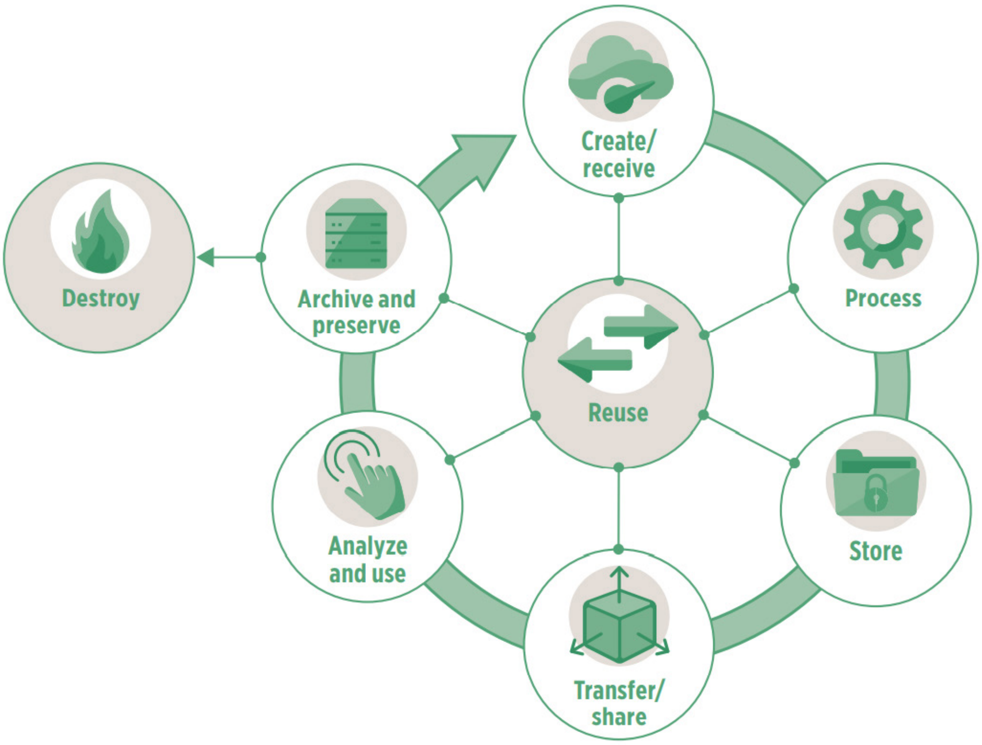
Figure 1. Visualizing a typography of data.



The emergence of big data in education, as evidenced by the variety of uses – both real and potential – outlined in Table 2, can be attributed to digital learning innovations and integrations. EdTech tools enable new pedagogical possibilities and, with their volume, velocity and variety, all the "big data" they generate represent a high-value perspective on learner behaviour for addressing questions that were either costly or even impossible to answer before these data sources were available.

Big data, and data in general, is reused, combined, shared, interpreted and reinterpreted in multiple ways and for different purposes, by different actors, without being depleted, as illustrated by the World Bank's "data lifecycle".

Figure 2. The data lifecycle.



Source: World Bank (2021).

Data do not exist in any unmediated or neutral form. Instead, all data are created from particular social, economic, technocentric, political or national perspectives, all of which needs to be accounted for when conclusions are drawn and outcomes implemented. As we are living in the age of the data economy, efforts to ensure agency, access, and awareness regarding data use, and to ensure the democratization of bias-free datasets, are of critical importance. Just as this section has broken down data into four categories, the subsequent section will break down Data for Learning into three categories to better understand the many dimensions of its use in education systems.

Why is data important for learning?

In July 2021, UNESCO, United Nations Children's Fund (UNICEF) and the World Bank joined forces to make the **Learning Data Compact** to ensure that all countries have at least one quality measure of learning by 2025. Such a compact signals that the global community is increasingly recognizing the importance of data in improving the quality of learning, teaching, administration, management and governance. Data plays a key role in determining progress towards the **Sustainable Development Goals** (SDGs), and in particular in understanding the toll of the COVID-19 pandemic on progress towards SDG 4, "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". It also opened the world's eyes to the vast data gaps that render the needs of many vulnerable learners invisible across regions; in high-income countries, only a narrow range of learner data may be collected, and in low- and middle-income countries, learning data may not be collected frequently or at all (UNESCO Institute for Statistics [UIS], 2021).

Data sources are limited by the **digital divide** between and within countries. In countries with higher capacities for data production and use, data can smooth the divisions and transitions between central, subnational and school levels. When used ethically and effectively, data users and stakeholders can make evidence-based decisions, plan strategically towards long-term goals, deliver services effectively and adjust strategies as situations evolve. Overall, Data for Learning can be categorized into three distinct although overlapping groups: (1) **data for administration, planning and governance**, (2) **data for teachers and teaching**, and (3) data for learners and learning. Each of these categories is described in the following subsections.

Box 2. Ceibal: Inclusive practices and the equitable use of data to enhance management and learning in Uruguay

Ceibal is a national initiative of Uruguay's Ministry of Education and Culture, which provides every student and teacher in primary and lower secondary public education with a personal computer, Internet access in schools, a national learning management system (CREA), and a comprehensive set of educational resources, including content, pedagogical services and programmes (Ceibal, n.d.*a*; Ceibal, n.d.*b*). The initiative has implemented a single login system for all its platforms, which helps to globally monitor data on user behaviour and trace users and user groups both when they access CREA and when they use specific platforms and resources. This enables the use and display of data to generate accessible, timely and valuable information for Ceibal's decision-makers, the National Administration of Public Education, school leaders, teachers, and the education community at large.

Ceibal's work in this regard enables a broad range of data use across various educational contexts. Regarding **education management**, Ceibal has created a school monitoring system, which brings together data produced by Ceibal and the National Administration of Public Education and makes it easier for school management teams to display and access valuable information including attendance records, assigned computers, computers being repaired, and use of learning platforms, among other indicators. Ceibal is also working on the generation and systematization of data, which it combines with strategies based on behavioural sciences to develop **behavioural interventions** aimed at solving problems in education. This has included the development of an intervention plan, which made it possible to increase attendance from 63 per cent in 2021 to 69 per cent in 2022 (Ceibal, 2022*a*). Another area of implementation relates to **using data to improve the ownership of learning platforms** such as the MATEC education project, which Ceibal is implementing to improve the teaching of mathematics through personalized student learning (Ceibal, 2022*b*). Using **data to help teaching and learning**, Ceibal makes adaptive learning platforms available to teachers and students including a platform used for learning mathematics at secondary and technical education levels. The tool uses AI to offer each student insight into their individual progress in the different topics covered, and allows the teacher to monitor each student, identify support needs and make timely interventions. Finally, Ceibal also seeks to analyse data on user demand to seek solutions to problems encountered by users accessing its platforms to **improve user experience and service**.

Data for administration, planning and governance

In many countries, but not all, data about education has long been collected, mostly in a conventional manner including administrative and analogue data collection processes and reporting. This has often drawn on multiple sources to include data such as the number of schools, teachers and learners (both those in and out of school), learner-teacher ratios, and school expenditure; the ages of the learners, their gender, their socio-economic status, whether they have special needs, their first language, the qualifications they achieve, the number of years they spend in school, and whether they complete primary and/or secondary education; and teacher qualifications and school inspection outcomes. Beyond formal education, data are also collected on the number of learners in vocational or higher education institutions, as well as about young people who are not in employment, education or training, or those who participate in non-formal or lifelong learning opportunities.

Learning management systems (LMS), education management information systems (EMIS), and school, university, and technical and vocational education and training (TVET) MIS, all which have their own distinct data requirements, are in a state of dramatic evolution. Many have grown in scope, integration and interoperability with the information systems of other sectors, primed to reach beyond the walls of the school building to support the collection of real-time, process-capturing and learner-centred data (UNESCO, 2022*b*). When these data are combined into aggregate data, it has the potential to help understanding of the mechanisms of specific policy effects and to address policy-relevant issues. For example, by connecting aggregate administrative data and individual transactional data from a digital learning platform, one can "unveil nuances" about educational inequities, and inform planning and management actions in faster feedback cycles (Fischer *et al*., 2020, p. 132). Data in this category include the many types identified earlier, such as traditional census data and data about schools, teachers and learners, which are used to inform, for example, the allocation of resources (capital and human), manage learner recruitment and monitor school effectiveness.

Data for teachers and teaching

Data for teachers and teaching, on the other hand, include data collected in classrooms from teaching and learning processes, assessments and data drawn from learner interactions with EdTech. Assessment data can be collected by hand, recording the achievements of learners in typical pen-and-paper tests, but might also be collected automatically in EdTech systems that include assessment functionalities, and may be underpinned by powerful data analytics and AI. For example, teachers may analyse a student's digital learning report to identify a particular skill the student may need additional support with. By looking at trends in class performance, they could differentiate their instruction and scaffold their lesson plans accordingly. They may use data-driven tools to lower the burden of administrative tasks, such as attendance, tracking homework assignments or entering assessments into gradebooks.

Data for learners and learning

Data gathered on learners by teachers in the classroom – both qualitative and quantitative – through observations, assessments and testing, and increasingly through EdTech applications, can have benefits for instructional practice, helping to fine-tune classroom pedagogies and generating a better level of individualized differentiation for learners. Learning data, in conjunction with other dimensions of education quality such as context, teaching and learning environment, and learner characteristics, reveal the factors that most affect learning outcomes. By revealing gaps in student achievement and service provision, learning data can be used to identify those groups which are currently underserved or are underperforming, and be used to hold education systems accountable for the use of resources.

Table 3. Sources of learner data.

Source	Example
National sources (In many countries, local education agencies and bureaux are also active sources providing accountability, enrolment and other demographic data.)	• Large scale learning assessments • EMIS databases • Multi-year sector plan documents and medium-term expenditure frameworks, operational documents (operational plan documents, budgets, mission reports, minutes of coordination meetings) • Financial data collected through financial management and reporting systems
Partner-facilitated sources	• Surveys and rapid assessments in the education and other sectors (including health, water, sanitation and hygiene, etc.) • Outcomes of multistakeholder joint monitoring and review exercises • Outcomes of decision-making within multistakeholder policy dialogue forum • Data available through partner reporting • Data available through education and learning ecosystems

Technology-facilitated sources	• Data available through LMS such as Anthology, Blackboard, Moodle, etc. • Data collected through online portals and open education resources (OERs)
Data collected directly from teachers and learners	• Qualitative data routinely collected by teachers through their teaching practice and multimodal assessment of learner progress • Data collected directly from learners through use of EdTech in the classroom • Data collected from stakeholders and beneficiaries through WhatsApp networks, social media groups, online and offline communication networks groups, etc.
Elements of learning that data systems may not be able to capture	• Elements of learning that data systems may not be able to capture reliably and ethically • Understanding, curiosity, imagination, creativity, thoughtfulness and collaborative processes of learning • Student frustrations, disappointments, missed learning opportunities, anxieties about learning, chilling effects of surveillance and workarounds to proctoring services • Learner social-emotional interactions, isolation and engagement/ disengagement • Data on informal, self-directed learning • Peer interactions, both within and outside classrooms • Social processes involved in navigating the learning spaces and material

Source: Authors, adapted in part from Global Partnership for Education (2021).

What learning data can we *not* collect?

Digital data collected in education are **partial** and cannot present a complete picture of all teaching and learning processes. Increasingly, they present interaction data from electronic systems but not other aspects of learning, such as reading, creating projects, talking to teachers, etc. However, the partial data that is collected and analysed can come to be taken as representing the whole, such that if a phenomenon does not appear in the data, it is either unimportant or effectively does not exist. EdTech is only capable of capturing data when the learner interacts with the EdTech system, which often constitutes only a **small part of the learning experience**. For example, no data are captured when the learner is involved in collaborative learning or project-based learning; when the learner is reading a paper book or involved in learning outdoors; when the learner is writing a poem, painting a picture or performing in a play; when they are learning work-oriented skills, such as mechanics, hairdressing or for the hospitality industry; or when they are engaging with one another, for example in discussions between themselves, or with their teachers. All these factors, although complex, are tacitly considered by most experienced teachers during their day-to-day interactions with their learners, but remain as **qualitative, contextualized data.**

If the subjects that are more visible in data systems are those which are more easily quantifiable, there is a risk that over time, certain subjects – for example, interpretive subjects such as the arts or the humanities – may receive less attention or value within education systems. Moreover, data captured about the **transition from education to the world of work**, work-based learning, or informal and non-formal lifelong learning, may not be integrated into information management practices in education systems. These activities all constitute aspects of learning and can contribute substantially to the learner's education experience, as well as to the health and success of the education system. If data from outside the formal education sector is invisible within education systems, then a deeper understanding of what learning experiences best lead to individual and societal flourishing is impossible. While data-driven policies tend to be based directly on a surface reading of the data, data-informed policies are inferred from a contextualized and critical interpretation, one that **balances the data with human behaviour and shared principles**.

Box 3. The importance of qualitative data

Data existing in a qualitative/disaggregated form, especially involving classroom interactions, the level of student involvement, their interest and anxieties, etc., can be effectively used by teachers and schools to improve the learning process. This type of data can also positively influence the design of assistive technologies aimed at supporting the needs of all learners, especially marginalized learners and learners with disabilities. Evidence of this can be found in the Ludic Design for Accessibility framework, designed at Microsoft Research India, and currently being adopted by the Computational Thinking for Persons with Visual Impairment (VICT) project. VICT aims to make science, technology, engineering and mathematics education accessible to students with visual impairments in India (Ludic Design for Accountability, n.d.; Microsoft, 2017).

The play-based pedagogical approach, implemented jointly by the non-profit enterprise Vision Empower, Microsoft, and the Centre for Accessibility in the Global South at the International Institute of Information Technology Bangalore, has introduced a range of accessible games. These include traditional games that use tangible artefacts, and specially designed accessible card games, to help children with blindness develop foundational numeracy and computational skills (Vision Empower Trust, n.d.).

Throughout the academic year, detailed observations and rich qualitative data are gathered from classrooms, with analysis of this data at local levels informing the introduction of novel methods into the learning process, placing a greater emphasis on children's participation and enjoyment. Including this qualitative data in the analysis of the programme has helped to drive success and scale the initiative from partnering with three schools in the 2019/20 academic year to over 100 schools at the time of writing.

Data for Learning as a double-edged sword

Ethical, thoughtful and innovative uses of data are likely to play a key role in the transformation of schools, learning processes, teaching methods, EdTech and the financing of education. In September 2022, at the time of the publication of this interim report, the United Nations will convene the **Transforming Education Summit**, which seeks to reignite a collective commitment to education and lifelong learning as a pre-eminent public good and as a key pillar of a new social contract. The desired outcome is to find new ways to target inequalities that have long prevented access to inclusive, quality education for all. This interim report on leveraging data to improve learning experiences is a timely contribution to the summit's goal of sharing knowledge, practices and resources to transform education to be relevant and responsive in the digital era.

Both the Transforming Education Summit and this interim report fall in the wake of over two years of **COVID-19** school disruptions. The experience of the pandemic is a reminder of the centrality of data and evidence for rapid decision-making during emergencies and crises. It also underscores the potential uses of data for building resilience mechanisms into education systems in anticipation of future shocks. Countries quickly gathered data through their public and partner networks to identify digital divides and school needs and design their education response plans accordingly. While context and resources were obviously critical in COVID-19 response capacities, the ability to gather and use data was a key differentiator in how quickly countries were able to get classrooms, students and staff online to resume schooling.

However, the uncontrolled expansion of data use in education risks undermining the value of education as a **public good** and a universal human right. The primary stakeholders in the expansion of EdTech are philanthropic foundations, education publishers, venture capitalists and tech companies themselves (Regan and Khwaja, 2019). EdTech could contribute to a reductive view of learning that values only that which can be numerically measured, tracked, and standardized, undermining the belief that all learning, including digital learning, is socially situated. Data-driven technologies are not simply unbiased teaching tools, but part of wider social systems within countries at vastly different stages of digital data culture and capacity development. Therefore, without critical engagement of the public in determining the long-term directions of Data for Learning, we may see a reduction in the control of teachers over their classrooms and an expanded market for student data, especially in lower-resource contexts.

Democratizing Data for Learning

The United Nations Road Map for Digital Cooperation invites countries to undertake a concerted global effort to encourage and invest in the creation of **digital public goods:** open-source software, open data, open AI models, open standards and open content. These digital public goods should adhere to privacy and other applicable laws and best practices, do no harm, and help the attainment of the SDGs. The UNESCO International Commission on Education (UNESCO, 2021*b*) considers that the best strategy for directing digital transformation towards supporting education as a common good is to ensure its democratization within a robust public sphere. The Commission considers that the continued development of digital technologies in education in directions guided by sustainability, justice and inclusion will require action from governments, support from civil society, and a broad public commitment to treating education not as an arena for profiteering, but as a space for public investment in a sustainable, just and peaceful future (UNESCO, 2021*b*, p. 112).

The Rewired Global Declaration on Connectivity in Education also highlights the double-edged sword nature of Data for Learning, stating that it "should help improve teaching and learning rather than merely document and control it" and that "students need freedom to take risks and make mistakes in online and offline environments built on trust and good will" (UNESCO and Dubai Cares, 2021, p. 6).

> **The ease of data capture, storage, and surveillance in digital spaces must be a primary concern for education. It should help improve teaching and learning rather than merely document and control it … Proper rules and protocols are needed to protect the rights of learners, particularly children. Education is a site of experimentation and identity formation, and students need freedom to take risks and make mistakes in online and offline environments built on trust and good will. An ethic of transparency and "do no harm" should guide data policies … Educational institutions should work to assure individuals own and control their personal data, and, in the case of children, families should be actively involved in decision-making. When possible, learners should be able to "opt-out" of data capture and still retain full access to educational opportunities.**
> (UNESCO and Dubai Cares, 2021, p. 6)

As this section has explored, Data for Learning offers much potential to positively transform education systems and processes. Clearly, however, the ability to capture, analyse and apply data in learning contexts comes with complex and numerous responsibilities. The uses of advanced data analytics in education remain a double-edged sword. As such, in order to successfully develop inclusive and equitable education ecosystems, the sword must cut both ways. Data must be wielded with care, including remaining cognizant of the inherent partiality of data, which is too often assumed to represent the whole, when basing decisions upon it. Instead, this partial data needs to be seen for what it is: a part of a picture. Only then can its analyses be truly representative.

Although this section has touched on several issues that must not be forgotten when attempting to implement Data for Learning ecosystems, the report will further investigate the challenges and risks associated with Data for Learning later. The following section, however, will examine the Data for Learning ecosystem from a lifelong learning perspective to help further develop and define the platform upon which we can attempt to build the future of education.

2

Examining the education data ecosystem

As Part 1 described, "data" is not a new technology, nor is it a new technology in the field of education. Teachers and education practitioners have long been recording and utilizing data in learning environments. However, Data for Learning holds a different meaning today, and holds new-found promise – and risk – to transform education for all learners. The breadth and depth of educational data now available invites many possibilities to transform education in many ways, with potentially positive and negative consequences.

Building on the common taxonomy developed in Part 1, Part 2 seeks to map a shared understanding of the Data for Learning ecosystem. It will outline the ecosystem users and the value of data to each actor, exploring how their interests are represented and the legal frameworks they fall under. In addition to exploring the WGDL's different strands in detail, the following section will also highlight potentially game-changing initiatives that could signpost the way towards the development of equitable and inclusive educational data ecosystems.

Ecosystem uses and users

Data collection, exchange and analysis all often involve better understanding of the various domains, as well as the way the data was actually generated. For data governance frameworks to be applicable across various learning contexts, it seems crucial to distinguish between the three domains of data:

- The **personal** domain covers all data relating to an identified, natural, or identifiable individual (personal data) for which data subjects have data rights.

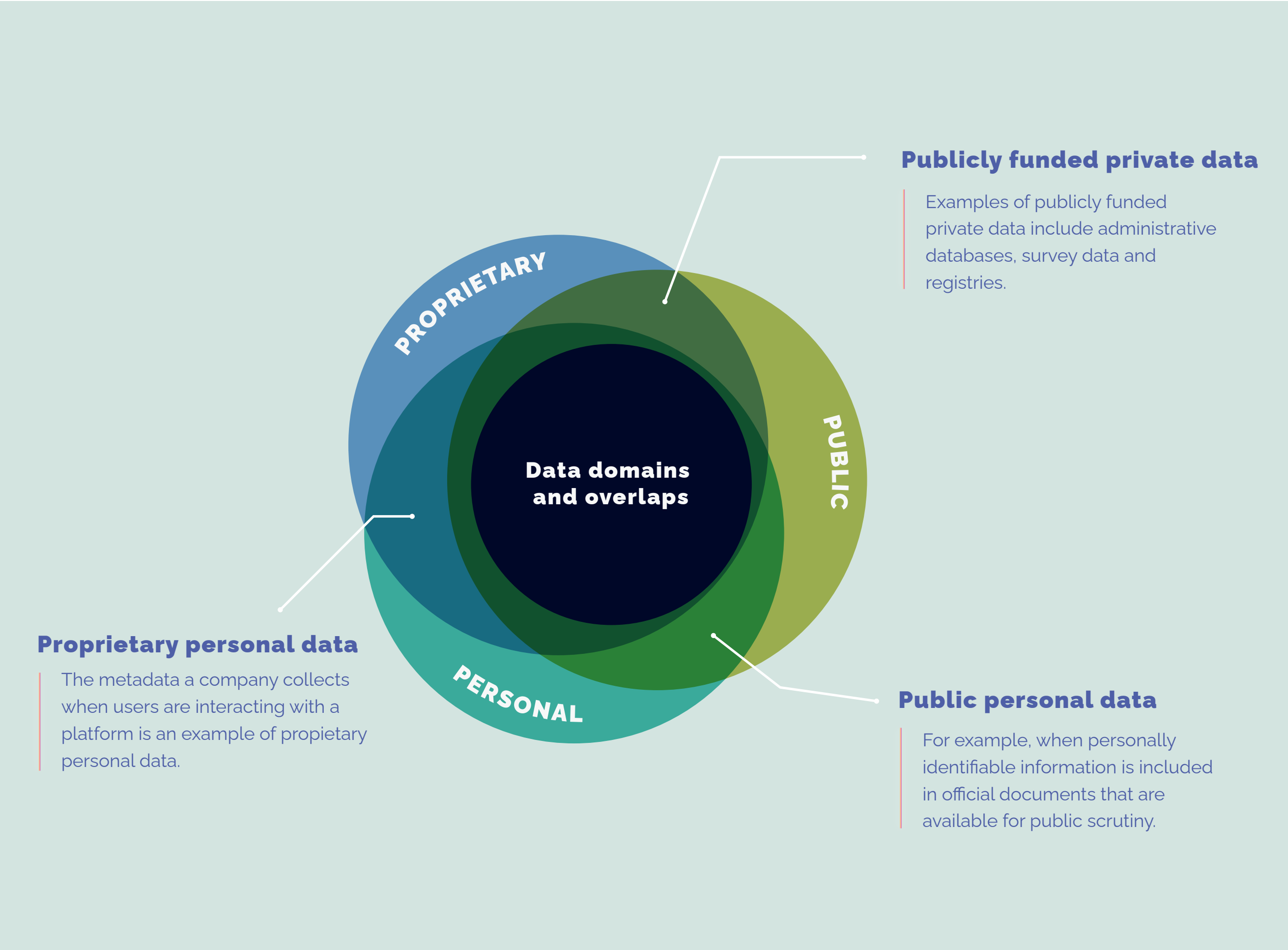
- The **proprietary** domain is typically protected by intellectual property rights (IPR) (including copyright and trade secrets), or by other access and control rights (provided by legal contracts, cyber-criminal law, etc.). There is typically an economic interest to exclude others.

- The **public** domain covers all data that are not protected by IPR or any other rights with similar effects, and therefore lie in the "public domain" (understood more broadly than to be free from copyright protection), thus certain types of such data are free to access and reuse.

These domains often overlap in real-world scenarios, and are also typically subject to different data governance frameworks that can affect each of them differently. For instance, privacy regulatory frameworks typically govern the **personal domain**, while the **proprietary domain** may not be subject to any specific regulatory framework, being mostly governed through contractual frameworks, or in some specific instances covered by IPR. The distinction between the personal domain and the proprietary domain, however, does not help differentiate how different stakeholders contribute to **data co-creation**, as multiple stakeholders are often involved in the contribution, collection and control of data.

Given domain overlaps, as well as the involvement of multiple stakeholders, it is no surprise that **data governance** is often perceived as complex from a legal and regulatory perspective. This is especially true where **cross-border data flows** are concerned. Currently, privacy and data portability rights vary significantly between countries, which adopt different approaches to personal and proprietary data for individuals and companies. **Data portability** aims to empower individual learners and give them more control rights over their personal data, but it remains unclear what type of data falls within the scope of cross-border initiatives.

Figure 3. Data domains and overlaps.



Source: Authors.

It is of paramount importance that the data governance or ownership frameworks applied in the education sector consider the interplay within data domains and data categorizations (personal, proprietary and public; metadata, transactional, aggregate). This interplay is relevant for the governance of data and data flows, for three reasons:

1. It helps determine the level of awareness that data subjects can have about the privacy impact of the data collection and process, which is critical when assessing the privacy risks associated with data collection and the level of control data subjects can be expected to have.

2. It reflects the contribution of various stakeholders to data creation, and therefore their rights and interests in accessing and using the data.

3. It helps identify the geographic location and jurisdiction based on data generation and collection, and it can therefore help determine the applicable legal and regulatory frameworks (Organisation for Economic Co-operation and Development [OECD], 2019).

The understanding of these domain and categorization interplays could help uncover opportunities and risks associated with data flows in teaching and learning contexts. For example, **data brokers** have emerged to monetize personal-domain metadata and transactional data to provide key insights on the learners often without proper consent or regulatory oversight. Third parties may obtain purchased or licensed data based on commercial (licensing) contracts (e.g. when data are acquired from data brokers) or other non-commercial means (e.g. when acquired via open government initiatives). As a result, contractual and other legal obligations may affect the reuse and sharing of the data.

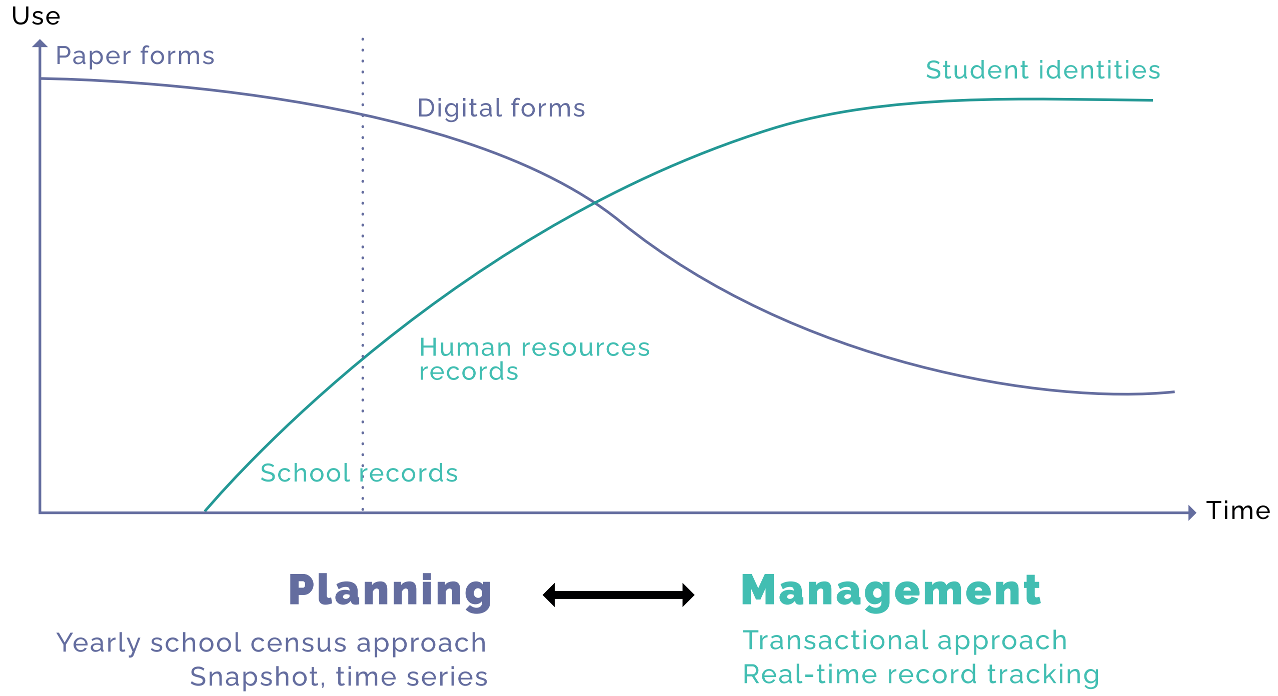
Ecosystem strands

Strand 1: Data infrastructure and learning ecosystems

Data infrastructure and architecture

Countries around the world are in the process of reimagining data architecture to enable more agile and accessible data gathering and sharing alongside more locally distributed data management processes and Data for Learning applications. Data ecosystems facilitated through accessible and integrated data management platforms are at the heart of this infrastructure. These digital platforms seek to enable powerful analytics that support real-time, data-driven decision-making at all levels, from day-to-day management of operations to strategic planning functions (UNESCO, 2021).

Figure 4. EMIS transformation over time, from management to planning.



Source: Authors.

In systems in which schools and learners have less access to digital technology, learning data may take the form of paper-based, mobile, radio or television data. The extent of data use is also impacted by the level of digitalization of MIS. In lower-income contexts, student data may be limited to that which ministry officials have experience managing, such as census or assessment data (UNESCOb, 2022).

Table 4. EMIS typology survey.

Regions	Paper (% of countries)	Standalone electronic mode (% of countries)	Online interface (% of countries)
Sub-Saharan Africa	81	31	19
Arab States	39	31	46

South and West Asia	33	22	78
Central Asia	33	33	67
East Asia	67	67	67
Pacific	75	75	25
Latin America and Caribbean	42	38	71
Central and Eastern Europe	0	13	100
Average	53	36	51

Source: UIS (2020).

The level of interoperability and openness of a system, as well as the digital capacities of its users, have great impact on the uses of Data for Learning. If the data architecture of EMIS (including higher education and TVET) is interoperable with that of school and LMS, or indeed with those of other government sectors, then teaching and learning data may achieve new levels of contextualization, thereby introducing improved decision-making for system management and sector planning. Education system data are not always integrated from non-traditional sources outside of formal and government-assisted non-formal systems, but they are becoming increasingly so (UNESCO, 2021*b*). **Cross-sectoral dataset interoperability** is used in some education management frameworks to reveal new insights to support holistic system management.

Learning analytics increase the potential usability of data gathered across state and non-state sources, and across formal and non-formal learning spaces and settings. AI can analyse census, administrative and learning process data, alongside structured and unstructured data gathered from Internet navigation, social networks, network devices, surveys, rapid assessments and more. AI might also enable a series of predictive markers to flag deviation from equitable educational provision and outcomes. Certain countries have invested in advanced data analytic platforms for the education sector to improve system management, planning, teaching and learning, examples of which are illustrated in the following table.

Table 5. Country use of data analytics to support learning, teaching and management.

South Africa's EMIS

As part of the National Development Plan 2030, the Department of Basic Education (DBE) implemented the South African School Administration and Management System (SA-SAMS). The out-of-the-box open EMIS has been designed for use across the South African education sector as a unified open data platform to support standardized education policy implementation across the nation and all regions.

SA-SAMS collects and reports data on education systems and provides real-time, validated datato help deliver improved data-driven educational decisions. The system helps over 10 000 schools across South Africa manage and administer systems including human resources, learner and parent information, governance, curriculum data, timetabling and more. The DBE reports that SA-SAMS has more than 15 000 daily users including principals, educators, administrative staff and DBE-supported staff (Thutong, 2022).

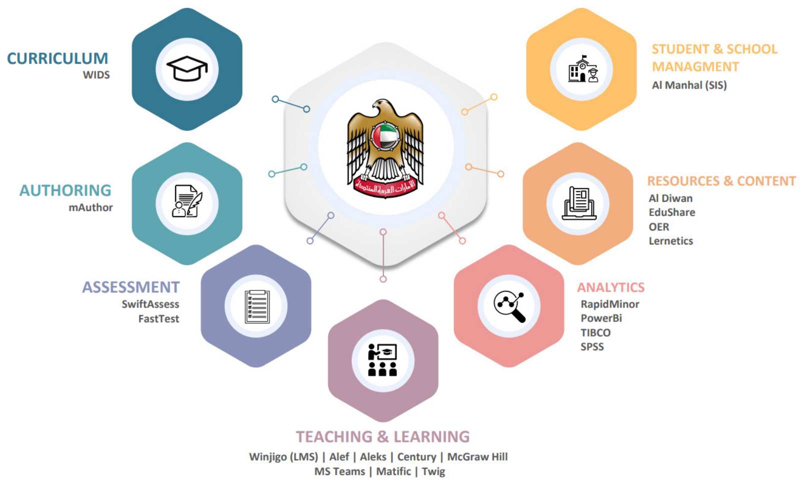
As well as improving operational administration and management at the school level, the mandatory reporting system also collates learner, teacher, management, and administration data to support operations across the entire education sector. In this way, SA-SAMS sits at the heart of South Africa's education modernization efforts, supporting tactical operations at the district level and strategic actions at the provincial level, while also informing policy decisions at the DBE.

United Arab Emirates' EMIS

Within its Ministry of Education, the United Arab Emirates has established a data analytics section dedicated to developing machine-learning algorithms in support of strategic studies on the country's education system. The Ministry of Education has rolled out a platform that is available to 1 200 schools and more than 70 higher education institutions, reaching over 1.2 million students. This data analytics system reports on curricula, teachers' professional development, learning resources, financing, operations, performance reports, teacher, student and parent feedback, and scores from international assessments like the OECD Programme for International Student Assessment and the Trends in International Mathematics and Science Study (G20 Education Working Group Report, 2022).

The United Arab Emirates Ecosystem for Education includes a unified student record, virtual schools, student portfolios, access to emerging technologies, personalized learning and optimized information technology (IT). Phase 1 focused on LMS and school information system (SIS) integration (attendance, grades, behaviour, timetable, surveys, events). Phase 2 involved developing a learning management container that was interoperable with other systems. Phase 3 brought all stakeholder information through the integrated and advanced learning platform to a unified data system.

Figure 5. The United Arab Emirates Ecosystem for Education.



Source: Adapted from UNESCO (2022).

The learning core pictured above brings everything together using Microsoft 365 and can integrate with different systems featuring varying levels of maturity. The core is integrated with external systems (ministries of health, HR, blockchain). The learning resource tool records every event that happens in the ecosystem. A future phase 4 will include AI, augmented and virtual reality (AR/VR), Internet of Things, and other functions to enable adaptive content, differentiation, personalized learning, and teacher use of integrated third-party tools.

Sierra Leone's Digital School Census

The Digital School Census in Sierra Leone is helping the Ministry of Education to make evidence-based financial allocation decisions as part of the Free Quality School Education programme for the most underserved communities, based on the accurate collection of enrolment and infrastructure data for all 11 000 primary schools in the country (including pictures, GPS coordinates, data on absenteeism and a teacher database). The conversion of the annual school census form to an open data kit format and the procurement of solar-powered tablets were key to the digital transition. Rapid data visualization formats are further helping to inform decision-making at the district and national levels.

Nepal's Equity Index

Launched in 2017, Nepal's Equity Index is an innovative tool covering the entire education sector. The Equity Index supports the Consolidated Equity Strategy for the School Education Sector in Nepal, adopted in 2014, which now underpins policies and interventions in the sector. The strategy relies on the Equity Index, which measures deprivation across districts.

Nepal's Equity Index is designed to capture data on disparities across the education sector. The index enables the Ministry of Education to use data to rank the prevalence of disparities in education outcomes (disparities in access, participation and learning outcomes). This then facilitates evidence-based planning and allocation of public resources to tackle the factors driving these disparities in a bid to reduce them. It also ensures that activities are undertaken to identify the barriers to accessing and/or staying in school, respond to the needs of children facing these barriers, and direct funds to where they are most needed.

Saudi Arabia's Unified Digital File programme

In Saudi Arabia, the Unified Digital File programme aims to develop a consolidated digital system that includes all student information and data. Digital files for students will include personal information, psychological, social and educational data, and a record of skills and knowledge to measure learning outcomes and to facilitate the use of diagnostic tools, including the detection of students at risk. Saudi Arabia's education data management and education data sharing initiatives aim to develop a system and platform for sharing education data electronically. The initiative seeks to increase transparency in education and support future decisions and plans.

Saudi Arabia's LMS, Madrasati ("My School"), works as a comprehensive e-learning management system and is linked to a national information system for students called Noor. The assigned digital content is aligned with learning goals and performance-monitoring dashboards. The system provides multiple educational tools to support the planning and implementation of educational processes synchronously and asynchronously. Madrasati is equipped with educational tools that promote a complete interactive educational journey to ensure quality of education and twenty-first-century skills acquisition, covering areas such as scheduling, learning objectives, virtual classrooms, an enrichment resources bank, and e-courses, learning paths, digital content and an e-learning dashboard. For optimal performance, these tools have been supported with practical guides and rich educational digital content that is scientifically grounded.

India's Rashtriya Uchchtar Shiksha Abhiyan (RUSA) to improve higher education access and equity and the National Digital Education Architecture (NDEAR)

RUSA is a centrally sponsored scheme operating in mission mode to fund state universities and colleges in India by using digital technology to enhance the efficiency of data collection, monitoring and evaluation. The schemes objectives include: a) improving access and equity in higher education by providing adequate opportunities for quality higher education to students from Scheduled Caste and Scheduled Tribe communities, and promoting inclusion of women, transgender people and people with disabilities; b) enhancing the overall quality of existing state higher education institutions by ensuring that all institutions conform to prescribed norms and standards; and c) enhancing employability by equipping students with the necessary skills and training relevant to the local economy.

India's **NDEAR** is an architectural blueprint for the country's educational ecosystem that defines a set of principles, standards and specifications, guidelines and policies to strengthen the digital infrastructure for education. NDEAR will provide a diverse education ecosystem architecture for the development of digital infrastructure. It also ensures data empowerment and protection of individual privacy and confidentiality, strictly adhering to India's data protection bill and laws.

Source: Authors, adapted in part from the G20 Education Working Group Report (2022).

Modern data architectures such as those described above rely on the recent transformation of data services. This transformation has shifted data processing from expensive, massive mainframe computers that served individual organizations to interconnected computers, introducing the concept of **cloud services**. The cloud is understood as the central hub for big data processing and analysis, enabling information to be accessed from anywhere. Recently, cloud computing has been eclipsed by **edge computing**, which enables data processing and data analysis to take place locally, eliminating the need for all data to be sent to a central cloud. Edge data centres generate a wealth of possibilities for locally enabled education actors to understand, manage and solve complex local problems using big data.

The learning transformations made possible by cloud-to-edge technology are visible in **OER platforms** and online learning portals that make it easier for teachers and learners to retrieve information and gain access to new content to support their learning. In the future, fifth generation (5G) cellular networks and AI, when combined with cloud-to-edge technologies, may allow for an even greater shift from centralized management towards a local culture of context-sensitive school management and customized applications tailored to a community's needs.

However, many countries do not yet have the connectivity or infrastructure required to incorporate the most promising, dynamic and safest elements of big data and machine learning, into their education data systems. Even high-income countries face challenges with their data system architecture, for example, in moving from cloud to edge computing or expanding existing digital infrastructure to remote communities. Where lower-resource countries decide to move towards integrated, open data ecosystems, significant investments will be needed to update their data architecture and cover the costs of data-migration projects and ecosystem maintenance.

In many countries, the lack of data on network coverage and connectivity is in itself a barrier to expanding such data systems and digital education opportunities. Tangible and exact data identifying which communities need connectivity are needed especially in peripheral and agricultural areas. Telecommunication companies, such as **Millicom** and **Ericsson**, work to bridge these rural-urban divides by identifying and connecting communities to enable the potential benefits of data-informed, digital learning. **Microsoft's EMIS project** supports education systems to develop modern data architectures for education reporting that respond to education ministry architecture.

Learning ecosystems. The experience of the COVID-19 pandemic has demonstrated some of the ways in which EdTech solutions can help to mitigate education inequities and unequal learning outcomes in other ways. Education technologies are increasingly integrated with advanced analytics, such as machine learning and predictive modelling, and as such, they promise to offer greater levels of individualized learning by allowing students to master concepts and progress at their own pace. In theory, teachers and educators learn from data generated by the software about students' progress and tailor lessons appropriately.

Unpacking the promise of personalized learning. Technologies that offer personalized learning programmes have dominated recent discussion on the future of digital technology in education. Companies such as Zearn, i-Ready and LearnZillion have built-in natural language processing, text-to-speech and speech-to-text for non-mother tongue and special needs students, which can allow students to practise their enunciation and writing without teacher supervision. Many blended learning policies are built on a belief that such programmes can offer more effective and personalized interventions than human teachers, who are limited in their ability to work one-on-one with many children in the same classroom.

Although self-led, digital learning platforms have been shown to support students with learning or sensory disabilities and to improve learning outcomes more effectively in lower-resource countries (Rodríguez-Segura, 2021), there is a fear that such approaches could undermine or undervalue the belief that teachers have a better understanding of students' learning progress than a digital programme (Williamson and Hogan, 2020). Additionally, it is imperative that personalized learning programmes are carefully vetted and rigorously reviewed before they are trusted in learning systems. Meyer *et al*. (2021) analysed the top 100 most downloaded apps marketed as educational for young children, and found that only seven scored in the high-quality category based on learning science, while over 50 per cent scored in the low-quality range. As such, EdTech companies may be selling the promise of personalized learning despite their services often having little understanding of best practices in pedagogy (Hirsh-Pasek *et al*., 2022; Meyer *et al*., 2021).

Supporting lifelong learning with digital credentials, open badges and blockchain. As young people and adults pursue learning throughout their careers and lifetimes, the processes of validation and certification of learning are becoming increasingly digitized, and digital credentials such as open badges are gaining in popularity. Digital competency badges are stored and available on any Internet-connected device, offering peace of mind that they are stored in a safe environment accessible via an encrypted connection. The **Insignias INTEF Open Badge Backpack**, for example, is automatically connected to various digital LMS (INTEF, n.d.). Digital badges can be imported across other open backpacks, and can also be shared on social networks (Twitter, Google+, Facebook) or shown on LinkedIn to complete professional profiles.

Another example is the **Digital Open Badge-Driven Learning project**, coordinated by the Oulu University of Applied Sciences and funded by the European Social Fund, to develop a nationwide open badge constellation, which enables the verification of adults' problem-solving skills in technology-rich environments by identifying and recognizing competences acquired outside the formal education system. The open badges created by the project will be piloted within different target groups in TVET and adult education, including preparatory training for TVET, integration training for migrants, adult students developing basic skills, and in upper secondary TVET.

For learner credentials, World Wide Web Consortium **verifiable credentials (VCs)** and verifiable presentation protocols are already operational and being implemented in the real world, in the learning credentials space, to support interoperability along with immutability of the records that are owned by users instead of third-party providers. The most notable large-scale implementations include the **EU's Education Verifiable Accreditation Records and the Digital Credentials Consortium** created by multiple universities, including the Massachusetts Institute of Technology (MIT), Harvard and Berkeley, that have developed and mobilized VC infrastructure for issuing, sharing and verifying digital credentials of academic achievements. In addition, many private players have already moved from purely identity solutions to education credentialing use cases and related solutions.

Countries such as Ethiopia, Estonia and Malta are using **blockchain** for credentialing or to create unique IDs for learners and teachers to help manage student transfers between schools, trace resources and track service delivery (UNESCO, 2022). For detailed examples of how countries are using digital data to support school-to-work transitions and to improve the accreditation of adult learning and upskilling, see Table 6.

Table 6. Using digital data to support school-to-work transitions and lifelong learning.

Digital Ethiopia 2025
A core component of Ethiopia's Digital Transformation Strategy (Digital Ethiopia 2025) is its partnership with IOHK, a global blockchain research and development company. Using a blockchain-based identification system, Ethiopia aims to give its 5 million students, 3 500 schools and 750 000 teachers unique records to verify performance, personalize learning and ultimately boost national education achievement and employment through data-driven planning (Parkin, 2021).
China's "24/365 All Day and All Year Round" online recruitment service
The programme targets online recruitment, providing online career development guidance and supporting the work of college employment counsellors. Over 1 million students have gathered information and secured jobs through the programme. The programme has cumulatively offered 23.3 million job vacancies. Registrations on the service platform number more than 10.17 million, and 86.56 million students have submitted their CVs through it. In addition, 42 large-scale job fairs have been hosted, offering 5.25 million jobs. The programme has been designed to offer employment-related cloud services for students, employers and school staff with the aim of ensuring a stable and smooth education-to-work transition in China.
Republic of Korea's Supporting Vocational Education in Secondary School programme
The programme provides vocational education opportunities that match student desires and aptitudes, and fosters technical talent and practical skills by delivering a curriculum that is highly relevant to industry. In addition, the provision of work experience and employment support through field training enables students who wish to enter the job market directly after graduating from vocational high school to find high-quality jobs. To create workforces with entry-level skills in new industries (e.g. digital) and respond to the changing job environment, there is a need to incorporate digital technology throughout vocational education and to increase students' understanding and utilization of digital technology. The curriculum includes programmes such as handling digital application devices and creating and utilizing content via digital applications. In particular, digital applications are used for experiments and fieldwork in vocational high schools.
Canada's Saskatchewan Ministry of Advanced Education
In Saskatchewan, micro-credentials are recognized as short knowledge, skills and/or competency-based programmes that should have clear, articulated assessments and demonstrated connections to the labour market and lifelong learning. Sasketchewan's *Guide to Micro-Credentials* is designed to provide learners, post-secondary institutions, employers and industry groups with a foundational understanding of micro-credentials and meaningful ways to use micro-credential programming. The Ministry has also created a landing page where learners can find more information on micro-credentials and see which institutions are currently offering them. Several post-secondary institutions are using digital platforms to create and issue micro-credentials such as digital credentials, with several adopting the MyCreds[1] digital credential platform. Micro-credentials issued as digital credentials give the earner full control over sharing the credential in any way they deem appropriate. The digital credential also allows the viewer (e.g. an employer) to seamlessly review all of the relevant information in the credential (e.g. assessments or skills/competencies earned).

Source: Authors, adapted in part from the G20 Education Working Group Report (2022).

1 See https://mycreds.ca/.

Strand 2: Data skills and competence framework for life and work

Within any typical information management system, capacities for data collection, accuracy, analysis and interpretation at a high level are needed. However, the human capacity to engage with high-quality digital data, delivered in real time through relevant technological platforms, is also essential for system functionality and effective decision-making. Over the past few decades, the increasing diversity of data sources from stakeholders within the education community has not always been reflected in top level decision-making. Government actors are not routinely leveraging multiple sources for their strategic planning and decision-making processes. This is partly due to the education system's low absorption capacity for integrating and applying the data that has been produced, and to weak human resource capacities for adjusting management and operational processes to improve service delivery.

Agenda 2030, Global Citizenship Education and Education for Sustainable Development, endorses the concept of building a just world through education to ensure that all learners acquire the knowledge and skills needed to promote sustainable development. These include data literacy and the key competencies included in a twenty-first-century skills framework, such as:

- **Information and communication skills:** These include, but are not limited to, analysing, accessing, managing, integrating and creating information; and understanding, managing and creating effective oral, written and multimedia communications in a variety of forms and contexts.
- **Thinking and problem-solving skills:** These include using reasoning to understand and make complex decisions; understanding the interconnections among systems; being able to identify, analyse, frame and solve problems; and fostering creativity and intellectual curiosity by developing, communicating and implementing new ideas while staying open to new perspectives.
- **Interpersonal and self-directional skills:** These skills cover teamwork and leadership; the ability to adapt to different roles and responsibilities; empathy; respecting diverse perspectives; monitoring one's own understanding and learning needs; locating appropriate resources; transferring learning between domains; exercising personal responsibility and flexibility in personal, workplace and community contexts; setting and meeting goals; tolerating ambiguity; acting responsibly with the interests of the larger community in mind; and ethical behaviour in personal, workplace and community contexts.

These competencies are critically aligned with a Data for Learning conversation, since any effort to expand data skills should be grounded in a framework that holistically considers ethics, local community needs and climate change. As such, an SDG-aligned approach to Data for Learning will provide learners with knowledge to promote social inclusion, find peace and cooperation, and raise awareness of how to protect our societies from emerging diseases, loss of biodiversity, large-scale pollution and the climate crisis.

Box 4. Mission 4.7 and data competencies

In 2015, all 193 Member States of the United Nations adopted Agenda 2030 and its 17 SDGs, a universal and interdisciplinary framework to promote prosperity, people and planet. Agenda 2030 calls for a drastic shift in the way that governments and society pursue economic development, while accounting for the effects of development on social inclusion and the environment. SDG Target 4.7 aims to ensure that all learners acquire the knowledge and skills needed to promote sustainable development.

Data for Learning

To support education systems in implementing the kind of transformative education championed by SDG 4.7 in everyday curricula, Mission 4.7 helps countries to translate these competencies and skills into grade-level appropriate guidelines that can be used by curriculum developers and educators to contextualize the guidelines for their environments, integrating them into their existing curricula or adopting revised curricula, as showcased in Mission 4.7's Guiding Principles Platform (Mission 4.7, n.d.*b*). The platform includes the framework guidelines that overlap with SDG 4.7, including Education for Sustainable Development and Global Citizenship. These frameworks have informed the development of our learning expectations for inculcating the kinds of knowledge, skills and attitudes necessary to achieve SDG 4.7, the same competencies that will ultimately inform learners' successful engagement with data about and for sustainable development.

Source: Mission 4.7 (n.d.*a*).

The Broadband Commission posits that low **digital literacy** is one of the main causes of digital exclusion, and it is often among the top answers when people are surveyed about why they do not use the Internet. According to the latest available data from ITU, less than 40 per cent of the population in 40 per cent of the countries reporting carried out at least one of the activities considered a basic skill (e.g. sending an e-mail with an attachment). Only 23 per cent of countries reported more than 60 per cent of the population having at least one basic information and communication technology (ICT) skill.[2] These indicators are based on a belief in learning-by-doing, or that by using technology, especially with the support of a skilled instructor, one will acquire digital skills. The same logic could be applied to data literacy. If the learner participates in the production and analysis of personal data and treats their own data as a way of critically exploring their beliefs, values, responses and social identities, then they will improve their data literacy and move "beyond the typical schooling practices of restating and critique" (Sheridan and Rowsell, 2010, p. 111, cited in Pangrazio, 2016).

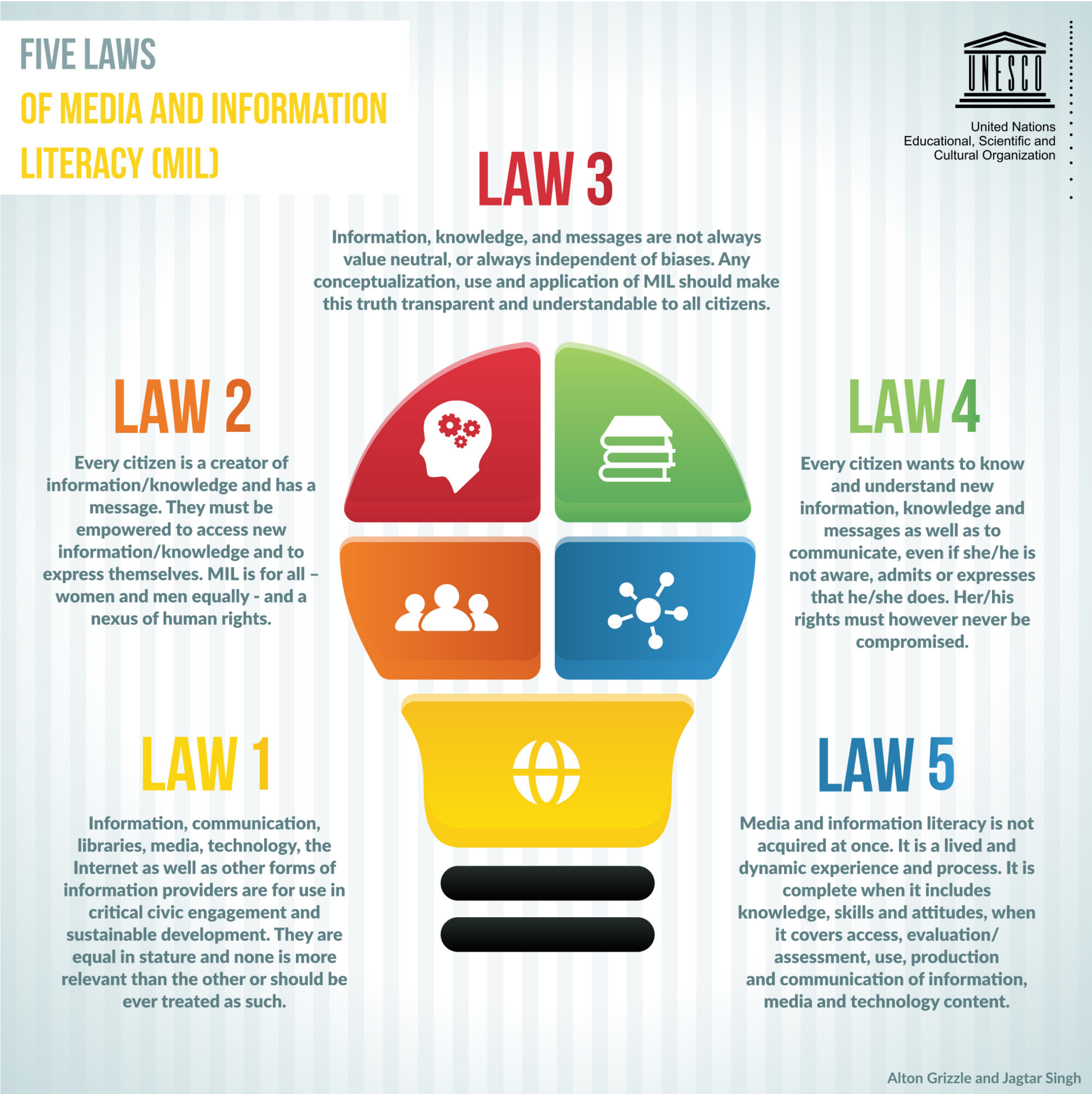
Data literacy is a multilayered concept comprising a combination of technical-statistical and analytic-narrative skills and has much in common with digital literacy. It is defined by the UIS as

> **...an individual's ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital technologies.**
> **(UNESCO, 2018*a*).**

Different organizations suggest definitions and taxonomies related to data literacy and skills. The Broadband Commission (2021a) report on hybrid learning noted that organizations from different sectors have specifically developed data literacy and competency frameworks at the global, regional, sub-regional and national levels. At the level of international government organizations, the UNESCO Media and Information Literacy Framework enables people's critical competencies, with the goal of empowering citizens to be creators of information/knowledge through lived, dynamic experiences, as indicated by its "five laws", set out as follows:

2 However, lack of data is itself a barrier to measuring digital literacy and data literacy. There are government gaps in basic facts and figures on young children's access to and use of technology in low- and middle-income countries, and so there are limitations to using these indicators to gauge digital competencies (Livingstone *et al*., 2019; Erstad *et al*., 2019, pp. 79-92).

Figure 6. The five laws of media and information literacy



Source: UNESCO (2018*b*)

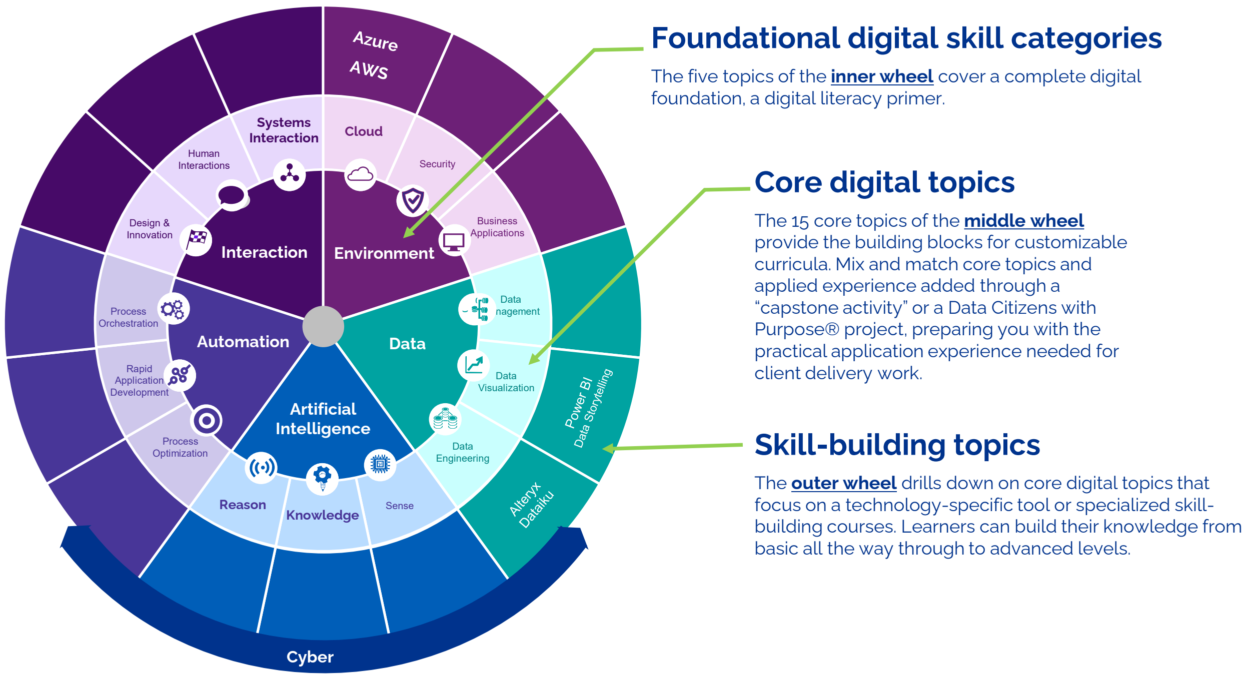
From the **civil society perspective**, the Open Data Institute (ODI) has produced a data skills framework that is widely used by governments and companies alike. Defining data literacy as the "ability to think critically about data in different contexts and examine the impact of different approaches when collecting, using and sharing data and information" (Tarrant, 2021), ODI separates data literacy into six categories: introductions, publishing, management, business, analysis and leadership, as visualized below.

Figure 7. ODI data skills framework

Source: ODI (2020)

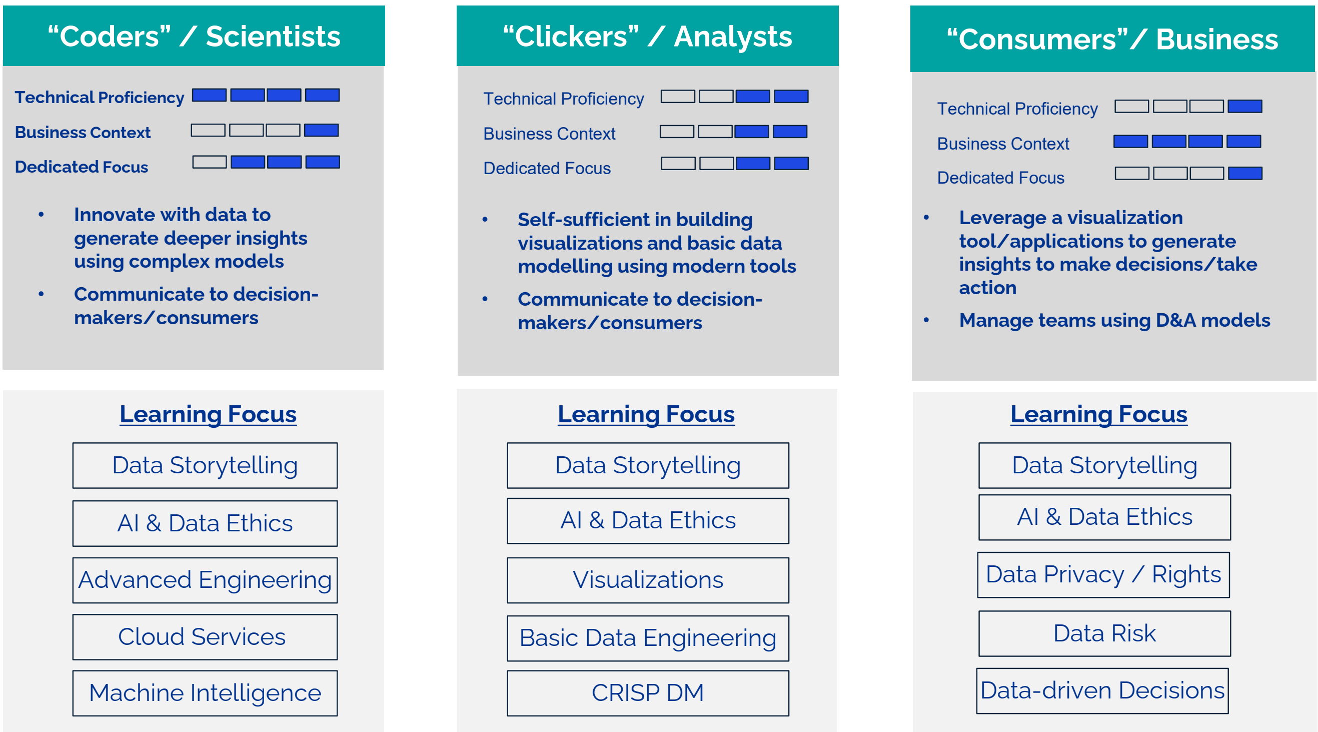
From the **corporate perspective**, companies have also developed their own skills frameworks to scaffold workforce development and growth through upskilling and reskilling. KPMG, a multinational professional services network, believes that **data literacy is not something that is just for data scientists**, it is for everyone. The company therefore developed a data literacy programme that helps its professionals to become comfortable working with data and analytics. KPMG recognizes that data literacy and digital literacy are interrelated, and that one cannot develop those skills without also considering digital skills. The company's broader skills circle, as shown in Figure 7, attempts to give a taxonomy to a range of skills that progresses into the outer rings, where one encounters actual tools.

Figure 8. KPMG's taxonomy of digital data skills.



Source: Presentation by Robert Parr during the WGD4L meeting on 5 July 2022.

Figure 9. KPMG's data skill needs across three personas.



Source: Presentation by Robert Parr during the WGD4L meeting on 5 July 2022.

To gauge how people progress through this circle, KPMG classifies individuals into three types of learner to determine the required level of data skills: **consumers, clickers and coders**. The common denominator across these three tiers is the standard goal of public engagement for democracy and empowerment through critical understanding of data uses and abuses. At the highest level, **coders** are the data scientists on the frontier of generating insights using complex models. Their learning focus is data storytelling, as they tend to have a technical background and may need to develop skills to better communicate complex insights derived from machine learning and deep learning. **clickers**, or data analysts, are those whose primary job is not to code, but rather to apply data science in business contexts. Their learning focus may be on how to weave data together from different sources credibly and ethically based on an understanding of basic methodologies for data mining, modelling and visualizing. Finally, **consumers** need training in storytelling, AI and data ethics, and data privacy and rights, for though they may not be proficient data scientists, they need accountability as managers and to understand the deeper questions to ask of data.

Coursera, a provider of massive open online courses (MOOCs) offers a different example of a **skills provider perspective**. Coursera's *Global Skills Report 2022* categorizes skills into three domains: business, technology and data science. In the data science category, Coursera identifies seven broad data skill types, which focus on capturing and using data to generate business decision-making or power underlying products and services (Coursera, 2022). The seven categories, along with relevant examples, are outlined in Table 7.

Table 7. Coursera's data science skills taxonomy.

Data science skill	Description	Example
Data management	Comprises everything related to managing and accessing data for reporting, analysis and model building.	Cloud application programming interfaces (APIs), Hadoop
Data visualization	Involves the creation and study of visual representations of data to communicate information clearly and efficiently.	Tableau, plotting data
Machine learning	Creates algorithms and statistical models that computer systems can use to perform a specific task without explicit instructions.	Multitask learning, deep learning
Mathematics	The study of numbers and their relationships, applying these principles to models of real phenomena.	Calculus, linear algebra
Statistical programming	Set of programming languages and tools used to create statistical models and algorithms.	R, Python
Statistics	Deals with all aspects of data collection, organization, analysis, interpretation and presentation.	Regression, A/B testing
Data analysis	Process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making.	Exploratory data analysis, spatial data analysis

Source: Adapted from Coursera (2022, pp. 39-40)

Although these two examples reflect corporate perspectives, they offer valuable lessons and approaches for formal educational institutions looking to design relevant data skills frameworks at various levels in the education sector. Educational organizations should pay particular attention to dimensions of equity in their data literacy programmes, in order to address the under-representation of women and individuals from the Global South in the field of data science. As data skills and digital skills are closely related, systems that currently offer little or no digital connectivity should consider an approach that combines digital and data skills into one taxonomy, such as the one used by KPMG. When thinking about skilling for future work, educational institutions should also consider that Coders also face the risk of being replaced by outsourcing or automation, whereas Clicker analysts and roles that require critical thinking may have higher job security due to ongoing demand.

To successfully apply a digital skills training programme like KPMG's in a classroom setting or formal system, active learning experiences are required. Schools could partner with open-source data initiatives to allow learners to develop their data skills using real-world problems that they care about. Learning without these raw materials is limited, a fact which further underscores the importance of **open data** as an **open educational resource** that can place cost-effective learning materials in the hands of educators around the world. If such processes and materials were combined into a structured curriculum, this approach to improving data literacy by focusing on the classroom level could be scaled up for global impact. Educational institutions around the world have caught on to this idea and are already adapting national or international frameworks into locally relevant initiatives for both teachers and learners, as demonstrated in the table below.

Table 8. Regional and national data literacy initiatives in education systems.

Australia's Data Literacy for Schools Leaders programme.

In the context of the Data Literacy for Learning strategy, Australia's Victorian Academy of Teaching and Leadership has developed a Data Literacy for School Leaders programme (Victorian Academy of Teaching and Leadership, 2022). The programme builds school leaders' understanding of the range of key datasets available through the Victorian Department of Education and Training and helps school leaders make evidence-informed decisions to improve student outcomes. The specific objectives are to be able: (1) to lead a data-informed conversation about school improvement; (2) to analyse key datasets that inform planning for school-wide improvement; and (3) to identify and explore key themes emerging from school datasets.

United States' CTE CyberNet programme

In the United States, the Department of Education leads and finances a large-scale data literacy programme for education leaders and teachers (United States Department of Education, 2021). The programme aims to rapidly expand the capacity of high school teachers to teach cybersecurity career and technical education (CTE) courses, thereby preparing students by giving them the range of knowledge, skills and abilities required to enter cybersecurity career and educational pathways. CTE CyberNet is a network of teachers' professional development intensive academies led by two-year and four-year post-secondary institutions designated by the National Security Agency (NSA) as Centers of Academic Excellence. The CTE CyberNet initiative was strategically designed through an ecosystem approach both to develop the cybersecurity skills of high school teachers and to create local support networks that increase the sustainability of these academies. CTE CyberNet has been supported nationally by private sector companies such as Microsoft, Offensive Security (Kali Linux) and Mastercard; local chambers of commerce and industry associations; and local employers, all of which have contributed to the development and sustainability of local ecosystems.

Code for Africa

Code for Africa and the World Bank have partnered with 22 institutions, including national statistical agencies, civic media and civil society organizations, universities and journalism institutes, to strengthen the culture of data use through data literacy training. With teams operating in 21 countries, Code for Africa's programmes aim to enable data access and use for government officials at all levels, as well as stakeholders in civil society, the private sector and academia. These activities led to the development of locally relevant MOOCs, with lessons on topics from data visualization to reporting on human trafficking, to strengthen learning opportunities within the region (Hammer *et al*., 2021). At the global level, the UNESCO International Institute for Education Planning (IIEP) offers a wide range of training courses, including data literacy dimensions, for education planners and school leaders (UNESCO IIEP, 2022).

Saudi Arabia's Madrasti programme and curriculum revision

Saudi Arabia's programme to revise and update its curriculum to enhance basic and future skills features programming from the first grade of primary school, including teaching digital safety, protection against threats and personal data management. The Madrasati digital platform provides a high-security centralized system that offers on-demand access to shared resources and data to personal computers and other devices.

Italy's National Digital School Plan

Italy's National Digital School Plan includes investment in the most innovative digital technologies for teaching coding, devices for robotics, virtual reality and inclusive education, to be used in at least 100 000 primary and secondary school classes.

Japan's Society 5.0

The Japanese Ministry of Education, Culture, Sports, Science and Technology has a JPY 9 billion policy programme to promote the realization of Society 5.0 and the development of highly specialized human resources for the post-COVID era in higher education institutions. The objectives include a series of policy and exchange programmes on enhancing mathematics, data science and AI.

Source: Authors, adapted in part from the G20 Education Working Group (2022).

Strand 3: Ethics, governance, national sovereignty and cross-border data flow regulation

The UNESCO Recommendation on the Ethics of AI provides a dedicated policy action area to offer guidance around ethical governance and stewardship. It recommends data governance mechanisms that are inclusive, transparent, multidisciplinary, multistakeholder and multilateral, which includes the possibility of mitigation and redress of harm across borders.

Today, many other frameworks are available to data science workers, students, researchers and industry professionals seeking to learn about and engage with the ethical dimensions of their work. For example, the ODI's Data Ethics Canvas, the United Kingdom Government's Data Ethics Framework, the Ethics Framework by Machine Intelligence Garage or lifecycle-based approaches. The variety and large number of approaches to data and computing ethics have inspired studies to map them and understand their similarities, what makes each one unique, their methods of implementation and their varied utility to different audiences (Ayling and Chapman, 2021; Jobin *et al*., 2019; Morley *et al*.,2020).

Data governance in the context of learning should include aspects of anticipation, effective protection, monitoring of impact, enforcement and redress to ensure that human rights and fundamental freedoms and the rule of law are respected in the digital world and in the physical world. Such encoded mechanisms and actions should include remediation mechanisms by design provided by companies in the private and public sectors. The auditability and traceability of data systems should be promoted to this end.

In addition, it is critical to strengthen the institutional capacities within the learning ecosystem to consider forms of soft governance, such as a certification mechanism for data governance frameworks and mutual recognition of their certification, according to the sensitivity of the domain in which they are used, the expected impact on human rights, the environment and ecosystems, and other ethical considerations. These mechanisms should include a regular monitoring component to ensure system robustness and continued integrity, as well as adherence to ethical guidelines over the entire lifecycle of the data flow, requiring recertification, if necessary.

Data-driven systems within the context of learning would be helped by seamless discoverability and the sharing of key insights in the form of governance initiatives, good examples of collaborative practices involving data-driven systems, and national and international technical and methodological guidelines as data sharing across borders advances, and more. Data governance models should put in place mechanisms to require all relevant actors to disclose and address any kind of bias in the outcomes of data-driven technologies or models, whether by design or by negligence, for example to ensure that training datasets for AI systems do not foster cultural, economic or social inequalities, prejudice, the spreading of disinformation and misinformation, or the disruption of freedom of expression and access to information. Particular attention should be given to regions where data are scarce.

We should encourage policy-makers to implement policies that promote and increase diversity and inclusiveness to ensure equal access to data-driven technologies and their benefits, particularly for marginalized groups. Data governance models should, where necessary, introduce liability frameworks or clarify the interpretation of existing frameworks to ensure the attribution of accountability for the outcomes and functioning of data-driven systems. Those who design and implement data governance models are further encouraged to use mechanisms such as policy prototypes and regulatory sandboxes to accelerate the development of laws, regulations and policies, including regular reviews thereof, in line with the rapid development of new technologies and to ensure that laws and regulations can be tested in a safe environment before being officially adopted.

Box 5. The ethical imperative for multilateral, cross-border cooperation

Researchers at Google have shown how current notions of algorithmic fairness, which mostly stem from United States-centric research, are often not compatible and may even be harmful when deployed in other countries and contexts (Sambasivan *et al.*, 2021). Based on interviews with dozens of activists, academic experts and legal authorities, the Google report highlighted three ways that the deployment of AI in India may require different methods for achieving justice and fairness from those required in the United States and other Western countries: (1) dataset bias due to the much broader digital divide that exists in the country, (2) the perceived inaccessibility of civil rights judicial recourse felt by many Indians, and (3) an eagerness among India's political classes to embrace new technologies such as AI in an uncritical manner.

In fact, numerous studies from around the world have examined the disproportionate influence Western ideals currently enjoy over AI ethics in general, making them unrepresentative in many of the contexts where they are being deployed (Sambasivan and Holbrook, 2019; Shin, 2019). With numerous EdTech providers and educational platforms spanning multiple countries, this risk of "AI colonialism" clearly demonstrates the need for cross-border cooperation among data and AI practitioners to ensure that the diverse experiences of all global citizens, including learners and students, are understood by the algorithms with which they will be interacting. A good example of what this type of work could look like comes from the Institute of Electrical and Electronics Engineers Standards Association, which has been investigating the effects that Buddhist- Ubuntu- and Shinto-inspired ethics systems could have on improving AI ethics systems (Institute of Electrical and Electronics Engineers, 2019).

On one level, the emergence of big data, data learning ecosystems and technological leaps suggest value creation in relation to more efficient information management and more transparent governance in the education sector. As stakeholders, partners and education actors consider the benefits of the latest data trends, they must be aware of laws regarding data protection in relation to each new type of data and the contexts in which it is used, and put safeguarding measures in place. As the Capgemini Institute has underlined, it is essential that data "happens in compliance with all local regulations and guidelines, and in an anonymous and aggregate manner (especially for personal data)" (Capgemini, 2021).

During the early days of the COVID-19 pandemic, governments and data owners such as telecommunication companies demonstrated a willingness to relax privacy rules for a range of passive locational data from mobile phone use, automatically recorded by mobile network operators or smartphones apps. This resulted in investments in legal and other agreements with mobile phone operators or smartphone app owners. It was highly regulated, however, and specifically intended to help track the spread of the pandemic. From a human rights perspective, and in all settings, considerations of protection, privacy and security must be at the core of efforts to share data on vulnerable and at-risk groups and support accountability while protecting the safety of communities. Accountability for data collection and use needs to be aligned with national data policies and consider all available legal protections.

The sharing of learner, teacher and private citizen data through multisourced datasets and the EdTech sector has raised concerns for a number of years. These ethical concerns mainly relate to data security, safety, tracing and potential biases ingrained in the data collection. When government-managed datasets are shared across national borders, additional concerns arise. The expanding use of EdTech to support digital and hybrid learning also highlights the need for strong ethical and legal frameworks, with stakeholders and beneficiaries involved in decision-making around how data are stored, processed and used (Frontier Technologies Hub, 2019).

Box 6. Data cooperatives

Data cooperatives could function as intermediaries who negotiate with companies and other entities to establish guidelines around the use of our shared data; set limits on who can view, store, use or buy it; and route the benefits back to us, in dollar form, in kind, or through recognition and access. Data coalitions (in Taiwan, Province of China, and elsewhere) own the data generated by distributed air and water pollution sensors, data that the public, private and civil sectors have used to design interventions and products to mitigate pollution problems. Organizations like Wikipedia function as data commons in that many people contribute data and knowledge into the pool. In small ways, these cooperative structures are already succeeding, and similar models could be real game changers in the field of education if designed and governed well. Driver's Seat Cooperative, for example, is an app-based ride-hailing platform with a dataset shared by the drivers (Driver's Seat, 2022).

AI requires large amounts of data to operate, and the past few years have seen a significant expansion in global consensus and international regulation on the ethical use of AI in education. UNESCO organized an international conference on AI and education in Beijing, China, in 2019. The conference produced the first internationally agreed document offering guidance and recommendations on how best to harness AI technologies in support of the Education 2030 Agenda, the Beijing Consensus on Artificial Intelligence and Education (UNESCO, 2019). The Consensus calls for the integration and/or development of AI technologies and EMIS tools to make education management and provision more equitable, inclusive, open and personalized (p. 5). At the same time, the Consensus emphasizes the importance of developing AI applications in education that are free from gender bias and to ensuring that the data used for AI development are gender sensitive (p. 8). In addition, the Consensus has a dedicated section on the ethical use of education data and algorithms, highlighting some key challenges including: (1) biased AI, (2) the balance between open access and privacy, and (3) legal and ethical risks. The Consensus also calls for further research and regulatory frameworks. The Beijing declaration kick-started a two-year international process to construct a standard-setting instrument to operationalize the consensus. After a multistakeholder process led by an ad hoc expert group of 24 specialists in the field, UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence in November 2021 during the 41st session of the UNESCO General Conference (UNESCO, 2021*a*). Currently, this expert group is looking at this recommendation in the context of education.

AI is one of many forms of data in education, and despite the progress in this specific area, the ethical use of data for teaching and learning more broadly has not reached the same level of international consensus and regulation. However, the regulatory frameworks that were developed for more advanced uses of data in education carry important precedents for regulating the broader landscape of Data for Learning. Many of the principles put forth in the Ethics of AI recommendation, such as those related to inclusive, bias-free learning data models, remain true in the more general discourse on data use in education. However, some complex dimensions of the dynamic data ecosystem remain unaddressed when examined from a purely AI perspective. Such issues include the often opaque and overlapping flows of data between and within nations. Clarity over data ownership and transparency over data use are much needed in the current sphere of education data.

The open data movement

The open data movement strives to improve the clarity and accessibility of Data for Learning by opening up data ownership in the name of data democratization. Open data is a part of the "open family", which includes open-source software, open access, open data, and OERs (UNESCO and Commonwealth of Learning, 2019). Conceptually, open data systems reside under the umbrella of open

Data for Learning

governance, as the philosophy of open government data is accompanied by a growing body of policies that promote transparency and accountability by making government data available to all, and that value creation and innovation through encouraging the use and free distribution of public datasets.

Box 7. Open Data 500

Data, including open data, can possess commercial value. The commercialization of education data is one of the core points of conflict with maintaining the sanctity of education as a social public good. Open data, however, operate under the Creative Commons licences 4.0, which include a specific clause on non-commercial use. As such, the uses of open data depend on which Creative Commons licences are chosen for a given open dataset.

The possession of commercial value and the use of open data for educational empowerment and equity are not, however, mutually exclusive, as exemplified by the Open Data 500 initiative (The GovLab, n.d.). The Open Data 500 roundtables matched the tech industry with public agencies that release open data. The goal was to create 500 start-ups, but the end result was more than double this target. This example illustrates that openly licensed data can create much commercial value without compromising the same data for social access, use and reuse by any citizen. Such approaches have also been implemented by governments in Australia, Italy and the Republic of Korea.

Open data as a model for the governance of education data possesses many promising advantages. For example, the policies adopted to ensure standardization and good service delivery in the government sector apply to the education sector. From the management perspective, such transparency would empower students and parents to audit school performance, funding and supplies, as exemplified by open school report cards. To prevent negative side effects associated with school report card generation, such as undue prioritization of standardized testing or teacher performance metrics, communities should be participants in the data collection process so that they are aware of performance standards, are invited to comment on and update the data, and are empowered to bring attention to corruption or mismanagement.

However, the movement towards eliminating sovereign barriers entirely through open data is being met with a simultaneous push in the opposite direction towards closing data borders to maintain “digital sovereignty” over datasets, limiting the movement of data around the globe (McCabe and Satariano, 2022). Many of these efforts are made in the name of combating data colonialism, or the ownership of data from the Global South by companies in the Global North, who may siphon off or exploit data created in developing countries for profit (Couldry and Mejias, 2019).

Major data storage providers have faced recent barriers in the face of regulations that limit the ability to store data beyond national borders, precipitating the creation of public-private partnerships to establish data centres for local data storage. The education data ecosystem is significantly impacted by this, especially in the field of digital learning, as many LMS and EdTech providers are located in North America and their ability to store the learning data of users from countries with strict data border regulations impacts their ability to deliver online services whose functionality depends on the steady supply of big data to charge their AI-fuelled learning models.

However, if EdTech tools themselves close their data borders through private, proprietary ownership, then the education sector is barred from using this learning data for self-evaluation processes or critical engagement on the part of school stakeholders. Closed ecosystems create the opaque and overlapping data systems mentioned earlier, since they impede the ability of EdTech tools to effectively respond to teaching and learning needs on the ground. As such, the EdTech sector itself stunts the product improvement of truly adaptive and locally relevant EdTech tools that support education systems if it does not share its data in partnership with local governments, creating feedback loops that can more rapidly influence the real-time pedagogical and practical needs of teachers, learners and school leaders.

Box 8. DXtera

DXtera is an international NGO that builds open-source solutions and is focused on the possibility of creating open authentic data (DXtera Institute, 2022). Open authentic data would enable data agents to share anonymized or encrypted data, or even to simulate data to create large datasets that could be utilized by researchers and product developers to reduce bias and increase the efficacy of their tools. The goal is to support the market for new tools that can address inequity in education delivery by anonymizing, correcting and protecting data generated by educational entities across the lifelong learning spectrum.

On the other hand, openness as a guiding principle of the education data ecosystem is likewise a double-edged sword and is not without its risks. Simply put, a tension exists between transparency and accountability on one side and anonymity, privacy and personal ownership on the other. UNESCO's Minding the Data report (2022) argues that a balance must be struck between the use of technology to transform education and the protection of individual rights to privacy. The efforts towards encouraging open data in the name of improving teaching, learning and management processes are countered by an effort to tighten restrictions to protect the right to the privacy, safety and security of learners, especially young learners, in education spaces.

However, this tension can be eased with conceptual clarification over what open data is and is not. There is concern that opening up data opens the door for personal and private data, especially children's data, to be commercialized, exposed or compromised. However, open data is not a type of information data. Rather, it is a way to deliver and socialize data. Openness in data relates to (a) the legal condition defined in an open licence that stipulates the permissions for others to use and reuse that data, and (b) the technical specifications regarding open and fair principles for that data to connect and interoperate with other systems, platforms and databases (Ruijer *et al*., 2020).

Both aspects create a shared context that widens access to data and promotes diversity in the use and reuse of that data. In the Data for Learning context, this may favour more innovative pedagogical practices using datasets, improved engagement and participation in data-informed dialogue, and greater efficiency and cost-effectiveness in processes related to education management and reform that thereby promote the system-wide acquisition of data literacy and higher order data skills.

Following the examination of the different strands of the Data for Learning ecosystem and the exploration of different country-level implementations of data analytics programmes in support of education, we now have a clearer picture of what Data for Learning means. This includes the users in the ecosystem and the different data governance frameworks that regulate how these diverse domains interact with and pull against the interests of each other. As well as legislative frameworks, this section has also clearly outlined the required skills and competence frameworks that need to be implemented if effective Data for Learning ecosystems are to be realized. These competencies range across the ecosystem from digital and data literacy frameworks to broader twenty-first-century skill frameworks that must be critically aligned to all attempts to successfully implement equitable and inclusive Data for Learning ecosystems.

Finally, the section has also explored the need for ethical frameworks that can successfully traverse the difficulties surrounding key tensions in the data ecosystem such as those between national sovereignty and the free flow of data across international borders, or between the sharing of learner and teacher data and individual privacy and security rights. As this section has clearly demonstrated, Data for Learning ecosystems are fraught with tensions and idiosyncrasies that go beyond those found in other data ecosystems because of the sensitive nature of the users involved and the complexity of teaching and learning processes. However, the identification of these tensions and the exploration of relevant frameworks that could help manage them mark a key step towards realizing equitable and inclusive Data for Learning ecosystems.

Therefore, to further build on the progress of this section, the following section of the report will concentrate on key challenges and risks that should not be overlooked when implementing and developing Data for Learning ecosystems.

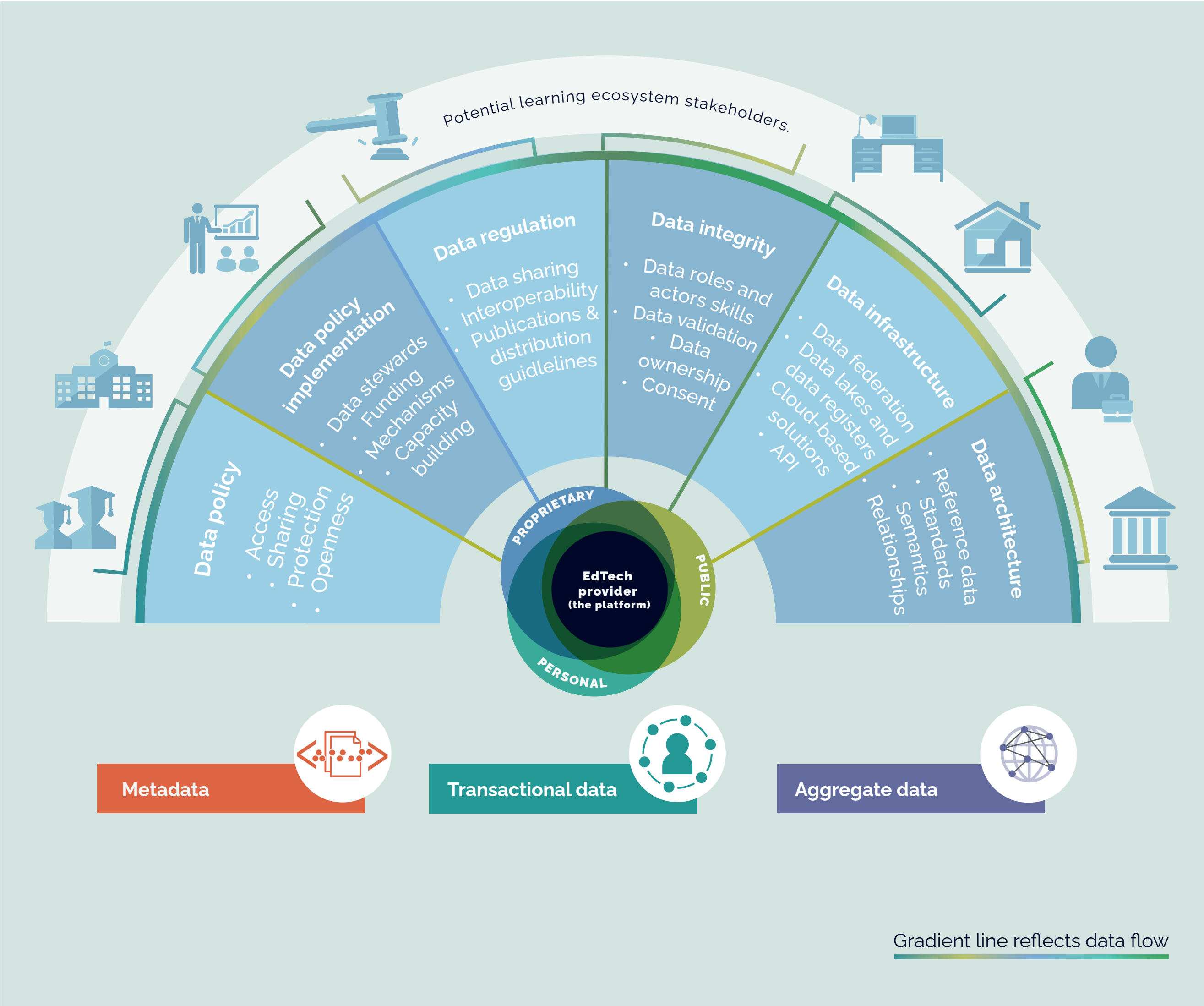
3

Data for Learning challenges and risks

In addition to those identified earlier, many of the challenges and risks related to data for education, especially big data for education, remain radically underexplored. Rather than simply capturing all available data, far more research needs to be conducted to establish what educational data are needed to address real educational problems, how that data might best be captured, and how best to analyse such data. Nonetheless, this section of the report will, as a starting point, endeavour to highlight other key challenges and risks that must not be overlooked or forgotten.

Governance, ownership and the common good

Figure 10. Learning ecosystem: Multistakeholder data governance and data ownership.



Source: Authors.

Above is a visualization of a learning ecosystem, which considers data governance, data ownership and the interplay between stakeholders. At the circular core is **ownership** (public, private, proprietary). This is then enforced by the wheel of **governance**, which has six key dimensions that stakeholders must consider: policy, policy implementation, regulation, integrity, infrastructure and architecture. In the white layer above are the relevant stakeholders, which range from regulatory to educational bodies and individuals.

Data offer immense potential value to drive transparent governance and management of education systems, empowering teachers and creating personalized learning experiences, assessments and certifications. However, data collection is also mixed up with economic and political power. Data accumulation can lead to a concentration of power, raising the possibility for data to be misused in ways that will harm learners. How do we reconcile data captured by commercial players ultimately for profit, with data captured and analysed for the common good?

The use of the terms "common good" and "public good" in relation to matters of education is a debated topic with diverse perspectives. Rita Locatelli, UNESCO Chair on Human Rights and Ethics of International Cooperation, reviewed UNESCO's position and stated that the "notion of education as a public good underlines the primary responsibility of the state in ensuring the right to education for all, in safeguarding social justice and the public interest in education", to underscore the core democratic values of equality and inclusion (2018, p. 2). However, the growing influence of non-state and for-profit actors in all levels of educational provision around the world poses problems for this definition. As such, the term "common good" has gained recent popularity in educational spheres, for the concept of the "commons" refers to developing innovative educational governance approaches that oppose the marketization of education goods and services.

Table 9. Public and common good.

Education as a public good
UNESCO, along with many other civil society and United Nations bodies,[3] has used the term "public good" to reaffirm a humanistic vision of education, justify the need to safeguard public interest and "reject calls for increased privatization or commercialization in education" (UNESCO and CCNGO, 2015, p. 5).
Education as a common good
Education as a common good introduces the values of innovation and community cooperation, or as Locatelli (2018, p. 11) states, "envisages new and innovative education institutions that can improve quality and efficiency thanks to the empowerment and the greater cooperation with the forces present in society." The sociocultural concept centres community justice and well-being over individual socio-economic investment.
Data for Learning as a common, public good
Data for Learning in this context, thus, should embody the democratically governed values implied by "public good" and the sociocultural, innovative, collaborative values implied by "common good". For examples of how open data may be considered an embodiment of this concept of Data for Learning as a common and public good, see **Open Data** (UNESCO, n.d.).

Source: Authors.

Digital data architecture, and a country's ability to process large amounts of data, are closely tied to the broader **digital divide**. To date, the primary drivers and stakeholders in expanding data frontiers in both the business and education spheres are resource-rich global businesses, philanthropic foundations, education publishers, venture capitalists and technology companies themselves. Under what conditions might their support generate value for resource-starved education systems rather than

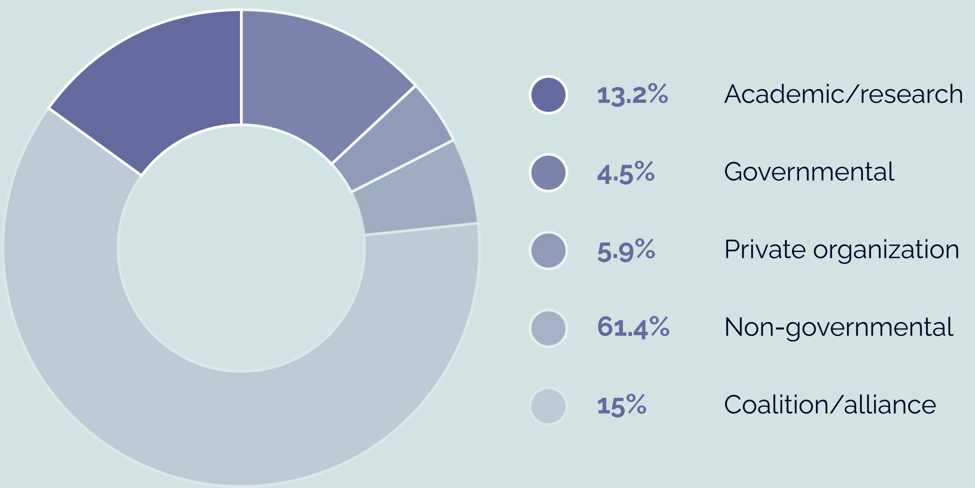
3 These include, among others, the Global Campaign for Education, the Right to Education Initiative, the Global Initiative for Economic, Social and Cultural Rights, and Education International (Locatelli, 2018, p.2)

undermine education as a common, public good and universal human right? Complexities to consider include manoeuvring intellectual property legislation, patenting, licensing, or lack of control over, visibility of or access to search and recommendation models for learning resources. In addition, governments must consider their role in protecting and expanding the **"knowledge commons"**, or the shared resources, information, data and content that is collectively owned and managed by a community of users.

Box 9. Datasphere Initiative

The Datasphere Initiative conducted a gap analysis of the geographic scope of data governance, existing norms and frameworks. It is accompanied by an interactive dashboard with a specific education governance filter (Datasphere Initiative, 2022). Of the more than 260 organizations mapped in the data governance ecosystem atlas, over 95 per cent are NGOs.

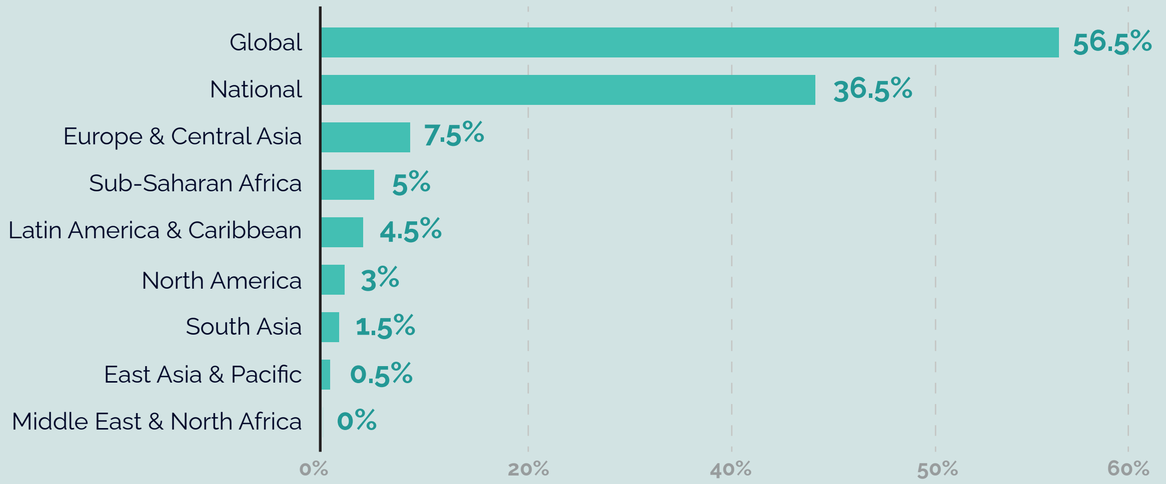
Figure 11. Types of organization identified by the Datasphere Initiative.



Source: Adapted from Datasphere Initiative (2022).

The activities of 41 IGOs (including the OECD and the United Nations) were included in the project, and most do not advocate a particular way of governing data. Major findings include that while more than half of the organizations have an intended global reach and impact, the majority, or 62 per cent (137 of 220 organizations), are headquartered in countries of the Global North.

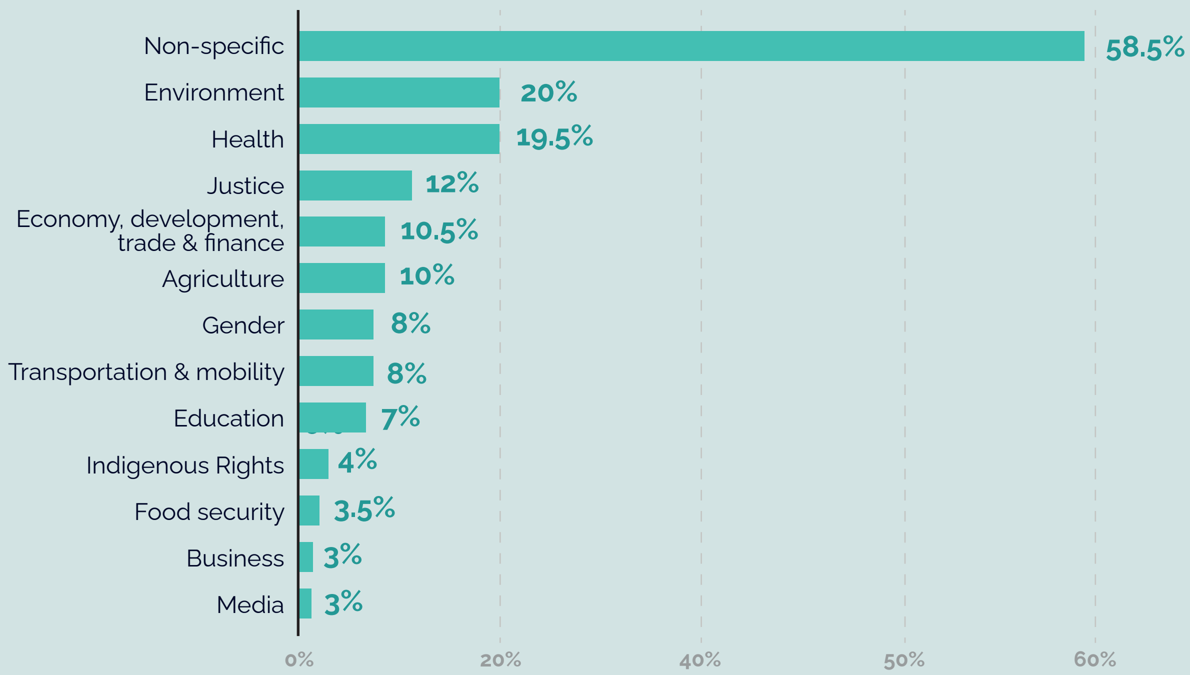
Figure 12. Geographic scope of organizations identified by the Datasphere Initiative.



Source: Adapted from Datasphere Initiative (2022).
Note: Geographic scope is an overlapping category. Organizations are counted more than once if their scope covers more than one region.

The primary focus of most organizations is on data topics related to health and the environment, with less focus on education, as evidenced by the mere 7 per cent of the 260 organizations profiled.

Figure 13. Sectors covered by organizations identified by the Datasphere Initative.



Source: Adapted from Datasphere Initiative (2022).
Note: Geographic scope is an overlapping category. Organizations are counted more than once if their scope covers more than one region.

This analysis confirms the lack of international regulation and makes clear the case for increased action and cooperation on education data governance across regions and sectors.

As indicated by the Datasphere Initiative's findings of minimal international cooperation on education data governance and regulation, it is clear that Data for Learning is far from a common, public good in the current climate. Commercial models for data use and ownership dominate education systems around the world, limiting users' freedom of choice to determine how their data are used. This is because educational data have value beyond that offered to learners, teachers, administrators and policy-makers. Learners' data also hold value for a variety of other stakeholders, from EdTech providers to data brokers, corporations, political parties, employers, social scientists, advertisers and more.

Where private actors step in to fill infrastructure and capacity gaps, this impacts data sovereignty and security, and is often accompanied by insufficient legislation or public awareness of how people's data will be used. Marginalized learners are often subject to more surveillance and abuse of data (e.g. Center for Democracy & Technology, 2019), giving rise to data colonialism, which refers to efforts to prevent or mitigate the consequences of data created in developing countries being siphoned off and exploited by EdTech or other technology providers in other countries. Another clear danger is that corporate-influenced reforms of education management systems may create a market for student data and reduce teachers' control over their classrooms without the critical engagement of governance bodies and the general public.

Given these risks, how do we resolve who owns the intellectual property relating to the data? Is it the EdTech provider or the learner? If the learner usually owns the intellectual property of anything else they create, such as a poem or a painting, how do we reconcile this with the claims of software providers, data brokers, EdTech providers and so on? What role can cross-border, multilateral normative instruments play in setting the agenda and establishing accepted ethical principles to be applied to both public and private institutions, given that student data plays a growing role in business plans?

Transparency, explainability, accountability and trustworthiness

Data transparency refers to the availability of knowledge regarding the use of data, including access to information concerning data ownership and consent. Beyond this, however, transparency also relates to the use of said data. When are the data being used, how are the data being used, by whom and for what purpose?

Something to consider is that transparency often sits at the opposite end of potential tensions with other principles such as privacy, safety and security. Meaningful transparency also implies something beyond simply dumping complete training datasets, which are akin to foreign languages to the layperson, onto the individuals in need of safeguarding. Transparency should aim to provide appropriate information to addressees about the use of metadata. Furthermore, transparency could introduce issues of fair compensation, especially with regard to AI technologies.

On this note, the **explainability** of data-informed systems refers to how easy it is to understand the input, output and functioning of each aspect of data processing (often algorithmic) and how it contributes to the outcome of the systems. Thus, explainability is closely related to the issue of transparency described above, as sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context. Transparency and explainability are closely related to adequate responsibility and accountability measures, as well as to the trustworthiness of data systems.

For-profit interests in student data may undermine trust in education systems. Higher education institutions around the world have expanded their use of big data analytics tools (Rubel and Jones, 2016), underscoring the need for increased transparency and explainability to maintain trust between learners and their higher education institutions. For an examination of the relationship between awareness of the use of learning analytics and trust in higher education, see Box 10.

Box 10. Learning analytics and trust in higher education

Studies of university students in France, Germany, Sweden, the United Kingdom and the United States reveal important insights into trust in institutional use of learning data. May *et al*. (2017) tracked the use of EdTech data at three universities in France and one in Germany and uncovered concerns related to privacy and to outcomes. Learners' opinions, behaviours and even learning outcomes were impacted by knowing how their data were being tracked and used. Another study, of university students from Sweden (Mutimukwe *et al*., 2022) illustrates the importance of informed consent relating to trust. Students' perceived privacy risk and privacy control clearly predicted their trust in learning analytics, as evidenced by their non-self-disclosure behaviours.

A study of students at the Open University in the United Kingdom in 2018 found similar results. The students who thought they had little control yet perceived high potential risks were less likely to exchange their personal data for the supposed benefits of learning analytics, revealing an important relationship between awareness of data use and actions to protect their personal data (Slade *et al*., 2019). The study also revealed that students inherently trust their institution to use their data ethically. Would this still be the case if institutions were open and transparent about how learning analytics were being applied to student data?

Given the strong relationship between awareness of the use of data analytics and increased individual action to protect personal data, colleges and universities have "a special responsibility to their students" to earn the trust of their students by openly disclosing algorithmic or predictive analytics (Jones *et al*., 2020). To capture student voices in this dialogue, Jones *et al*. conducted over 100 interviews with students from eight universities in the United States and found that students saw potential in learning analytics but felt that they should be educated about their institutions' analytic practices and should be treated as partners in designing Data for Learning strategies.

Similarly, without accountability, the unethical misuse of educational data becomes a serious concern. There need to be consequences for the misuse of data to incentivize and promote the protection of all learners' data, as well as their security and privacy. Again, this accountability relies on the transparency outlined above. Who is using data, why are they using it, how are they using it, and do they have consent?

For the key stakeholders in the learning ecosystem to make data-driven decisions based on accurate analysis, the underlying data must be trustworthy. Therefore, it is critical that data trustworthiness issues, which also include data quality, provenance and lineage, be investigated for organizational data sharing, situation assessment, multisource data integration and numerous other functions to support decision-makers and analysts (Bertino *et al.*, 2009). In general, the problem of providing trustworthy data to users and applications is an inherently difficult one, which often depends on the application and data semantics as well as on the data collection modalities, context and situation.

Box 11. Frameworks and approaches for ensuring trustworthy data

Data trustworthiness is inextricably linked to the concept of data quality. Data may be considered to be of high quality "if they are fit for their intended uses in operations, decision-making and planning" (Kerr *et al.*, 2007). Alternatively, data are deemed to be of high quality if they correctly represent the real-world construct to which they refer. There are a number of theoretical frameworks for understanding data quality. One framework seeks to integrate the **product perspective** (conformance to specifications) and the **service perspective** (meeting consumers' expectations). Another framework is based in semiotics to evaluate the quality of the form, meaning and use of the data. One highly theoretical approach analyses the ontological nature of information systems to define data quality rigorously (Price and Shanks, 2005). In addition to these more theoretical investigations, a considerable amount of research on data quality has been devoted to investigating and describing various categories of desirable attributes (or dimensions) of data.

Another promising framework for assuring information trustworthiness is based on a comprehensive framework composed of two key elements: **trust scores** and **data confidence** policy (Bertino *et al.*, 2009). Trust scores, associated with all data items, indicate the trustworthiness of each item. Trust scores can be used for data comparison or ranking. They can also be used together with other factors (e.g. information about contexts and situations, past data history) to make decisions about how to use data items. A framework that provides a trust score computation method could be based on the concept of data provenance, as provenance gives important evidence about the origin of the data, that is, where and how the data are generated. The second element of the framework is the notion of confidence policy. This kind of policy specifies a range of trust scores that a data item, or set of data items, must have for use by the application or task. It is important to note that the required range of trust scores depends on the purpose for which the data have to be used.

In many cases, it is crucial to provide analysts and processing applications not only with the needed data, but also with a universal annotation that indicates how much the input data can be trusted. This task is particularly challenging, especially when large amounts of data are generated and continuously transmitted across the system. Furthermore, solutions for increasing data trustworthiness, such as those specifically targeting data quality, may be expensive and may require access to data sources that have access restrictions due to data sensitivity (Bertino and Lim, 2010).

Interoperability

Information in and across educational organizations is distributed over different servers and stored in diverse databases. Although the discovery of information is an immense help in developing value-added services, data often remains siloed, inhabiting static spaces in the ecosystem, a space where the data may be difficult to access and analyse in the pursuit of improved learning outcomes. Despite this, many EdTech vendors often treat interoperability as an afterthought, while prominent data-related business models prefer to keep users captive by preventing the easy sharing of data across systems.

Furthermore, interoperability among education stakeholders needs to be considered as part of a multilayered approach that encompasses technical, semantic, organizational and legal aspects. **Technical** interoperability involves technical issues such as linking exchange protocols, file types, formats and services. **Semantic** interoperability entails giving precise meaning to the exchanged information so that it is preserved and understood by all parties.

Organizational interoperability refers to coordinated processes in which different organizations achieve a previously agreed and mutually beneficial goal. **Legal** interoperability involves aligned legislation so that exchanged data are accorded proper legal weight (EU Commission, 2017). Interoperability describes how well systems, applications and the data within them interact within an organization. When it comes to educational institutions, interoperable technology can facilitate student access to a broader range of tools, and enable schools to more easily share and manage data systems.

The frameworks proposed in New Zealand by Wakefield (2017) outline the depth of consideration that should be given to educational transformation in an era of increasing data practices. This involves both embracing the use of technology to improve efficiency and the availability of data across the education system as a whole, and including the technologies associated with the Fourth Industrial Revolution and related skills in the curriculum. Wakefield highlights the importance of the student as the central point in a web of stakeholders, with the modern technologies described above forming core methods of engagement at multiple points across the system. Another example of interoperability frameworks in action comes from South Africa, where JET Education Services proposes a set of recommendations to ensure interoperability within the education system (see below).

Table 10. JET recommendations for ensuring interoperability in education systems.

Key recommendation 1
A focus on technical or narrow interoperability is insufficient, and broader semantic frameworks and organizational interoperability must be considered. The purpose of an interoperable system is much broader than the aggregation of data. Therefore, from the onset of the development or implementation of such a system, factors such as the creation and use of data must also be considered, and the methods and means through which organizations and individuals will engage in the system should be defined.
Key recommendation 2
Invest substantial time in initial preparation. Initial preparation includes not only stakeholder engagement but also marketing and advocacy, and investments should be made in both initial and ongoing advocacy for and marketing of the benefits of the ecosystem or platform. Understanding the purpose and intended outcome of the ecosystem is vital at this stage.
Key recommendation 3
Explore and leverage available systems. The benefits of joining an existing community, in most cases, will far outweigh the benefits of creating a bespoke system. However, this should be done subsequent to robust internal consultation in which the needs of users and the intended outcomes of the system are determined to ensure the needs of beneficiaries are met. It is notable that even in cases where specific objectives may not be reached, it is still likely to be more effective to leverage available education standards which are collaborative in nature and allow for individual application planning interface (API) development on top of a standard or core offering. An open suite with a component architecture is thus recommended.
Key recommendation 4
Work backwards from the needs of users to design the system and its components and allow ample opportunities for innovation to come from and in collaboration with users. In designing and building the system, agile development should be a primary focus and the system should be built in iterative cycles. A responsive system is adaptive and innovative and allows for connection, integration and collaboration. It is recommended that a customer discovery and validation process be undertaken in the first instance, where assumptions and hypotheses are tested.

Key recommendation 5

Continuously improve through innovative cycles of development. The strongest systems will allow for innovation to come from a broad community and will encourage and reward the sharing and open-source nature of tools developed by community members. At the same time, it is imperative to not only rely on community members but to ensure that ongoing maintenance and redesign are adequately budgeted for and/or funded on an ongoing basis.

Key recommendation 6

Plan for future updates, additions and adjustments to data. It may seem obvious, but it is of paramount importance to consider the ongoing collection of data required by education systems, a consideration which may be overlooked by technical experts. Systems should have a sound and agreed-upon semantic architecture, and whatever tools/APIs are associated with the project should have the functionality to manage updated or replaced datasets or databases without a redesign. Leveraging international and/or national standards can significantly reduce the initial build time and cost.

Key recommendation 7

Define clear parameters for ownership and use of data, platforms and data standards. Concerns about ethical access and use of data must be resolved. Consideration must, therefore, be given to ownership and access to data as well as data portability. It is advisable that individuals are positioned as the primary owners of their data, and systems must work to ensure the portability and beneficial use of such data.

Key recommendation 8

Schedule data releases, encourage the use of APIs and charge for custom requests. Discrete requests from a central management agency can be accommodated, but these should require a financial commitment to prevent abuse and further encourage use of the available APIs and platform functions.

Key recommendation 9

Plan for investments in capacitation for all beneficiary groups. In the post-school education and training (PSET) system, this includes government departments, PSET institutions (including higher education institutions, qualifications authorities, sector authorities, etc.), labour market and work integrated learning representatives, PSET staff and students.

Key recommendation 10

Consider from the onset who will hold a system developed long-term. This body may be different from a governance structure and must be sufficiently capacitated and funded. It is more efficient if the unit can leverage additional resources and expertise for specific development tasks. Notably, a central managing organization does not have to assume ownership of the data or the outputs of the data; in some cases, the organization is responsible for only the development process.

Source: Shiohira and Dale-Jones (2019).

Interoperable technology in education delivers capabilities, such as seamless data sharing, that can help move towards more individualized learning at scale and could potentially reshape the learning environments of the future. How then, do we enhance interoperability between different EdTech providers, or between different countries, so that the data can genuinely inform theories about teaching and learning and contribute towards advancing SDG 4?

Inclusion, diversity and fairness

We can learn to prevent the misuse of data in education by learning from misuse in other sectors. Organizations increasingly rely on algorithms to help make decisions that impact people's lives, including who gets a bank loan (Unitas 360, n.d.), a job (Metz, 2020) or jail time (Zhu, 2020). Public backlash has led to proposals like the Algorithmic Accountability Act (Jones Day, 2019), which would require the United States Government to develop rules that mitigate algorithmic bias and provide ways for citizens to appeal automated decisions.

It is well known that digital proctoring methods such as tracking eye movements are commonly operationalized by for-profit tech companies to measure engagement with products and advertising. What happens if we define engagement in terms of eye movement in learning spaces? Some EdTech companies have already attempted to quantify traditionally qualitative data on social and emotional factors by deploying AI facial coding algorithms that learn about facial expressions of emotion. When evaluating such data for a learning model according to the United Nations Convention on the Rights of the Child (UNCRC, 1989), significant ethical, legal and socio-emotional risks were revealed (McStay, 2020).

All algorithms are prone to some degree of error. In large EdTech companies, even a tiny error can have a large impact. Every effort should be made to audit such systems for fairness, make sure the trade-offs between flexibility and efficiency are transparent and treat individuals with compassion and respect. The impact could be even more detrimental if algorithms are applied within education in an opaque, ill-designed data architecture.

Box 12. Challenges on the horizon: Synthetic data

Engineers generate synthetic data based on a smaller sample of real data that is labelled with all the aspects deemed relevant for the AI models to train on, and a set of rules that seek to counteract any obvious, known biases in the original dataset. However, taking too optimistic a view of synthetic data for creating data ecosystems without bias or privacy issues overlooks the reality that it is still individuals who are making decisions about what data to include and exclude, and how to analyse the data, with those choices based on what is deemed important or relevant by those individuals. If people are making decisions on which of these datasets should be built, which problems they should solve, and what real-world data they should be based on, we will never be able to fully remove bias. Furthermore, as synthetic data are based on smaller samples, such data may not only reproduce the patterns and biases drawn from the data but amplify them too.

In a worst-case scenario, we could get an echo chamber effect, whereby AI feeds the AI and models that develop and control key aspects of our world – the information we consume, the digital worlds we frequent, the recommended learning paths and learning products we receive – which increasingly respond to an internal logic divorced from the reality we inhabit. If the synthetic dataset is not grounded in (or perhaps made from) a rigorous understanding of the most recent underlying human phenomenon – such as the differences between what people say and do, or the unexpected influence of tangential variables in our lives on the actions we take – it risks simulating a social world that short-changes reality in ways that could cause real harm to individuals. And this is before we even begin to contemplate more nefarious uses of synthetic data, such as deepfakes or misinformation on a massive scale.

Synthetic data also raise complicated issues relating to privacy and consent. From the legal perspective, as synthetic data are often not data relating to a natural person (under the General Data Protection Regulation [GDPR]), or to a particular consumer or household (under the California Consumer Privacy Act) or to an individual (under the Health Insurance Portability and Accountability Act), then they are not considered to be personally identifiable or sensitive information. Such data are therefore outside the scope of these privacy laws. As such, synthetic data are not held to the same accountability principles and profit from being detached from "natural" individuals or legal structures. In the absence of clear associations, synthetic data rights should be rethought and reprotected by design and regulations, especially those concerning consent, privacy, accountability and explainability. Otherwise, the unethical adoption of synthetic data in education could go unchecked before scaling becomes irreversible.

At the local level, Member States should promote equitable representation between rural and urban areas, and among all persons regardless of ethnicity, gender, age, language, religion, political opinion, national origin, social origin, economic or social condition of birth, or disability, and any other grounds, in terms of access to and participation in the data-driven system lifecycle (UNESCO, 2021a). At the international level, the most technologically advanced countries have a responsibility to show solidarity with the least advanced to ensure that the benefits of data-driven technologies are shared such that access to and participation in the data-driven system lifecycle for the latter contributes to a fairer world order regarding information, communication, culture, education, research, and socio-economic and political stability. Furthermore, digital and knowledge divides within and between countries need to be addressed throughout the digital system lifecycle, including in terms of access and quality of access to technology and data, in accordance with relevant national, regional and international legal frameworks, as well as in terms of connectivity, knowledge and skills, and meaningful participation of the affected communities, such that every person is treated equitably.

The international community is coming alive to this reality, with scholars and experts warning against blindly trusting the supposed neutrality of data. For example, the Beijing Consensus on Artificial Intelligence and Education recommends that stakeholders "be cognizant that AI applications can impose different kinds of bias that are inherent in the data on which the technology is trained and which it uses as input, as well as in the way that the processes and algorithms are constructed and used" (UNESCO, 2019). Similar multilateral understandings need to be built and developed regarding Data for Learning ecosystems too.

Box 13. Unfair data practices: UK A-level grading during COVID-19

The COVID-19 pandemic caused major disruption to education systems all over the world affecting both high- and low-income countries. One such country was the United Kingdom, which saw nationwide school closures at a time when students would normally be sitting end-of-year exams (United Kingdom Department of Education, 2020). This included A-level exams for students aged 16-18, the results of which directly affect the higher education opportunities available to students.

When the government asked teachers to assign A-level grades in place of exam results, there were more higher grades than usual. Therefore, in an attempt to objectively standardize the grades of all students, the UK Government turned to the Office of Qualifications and Examinations Regulation (Ofqual), which used an A-level grading algorithm to assign grades to A-level students (Berridge, 2020).

The result saw 40 per cent of students receiving lower grades from the Ofqual grading algorithm than teacher assessments had indicated they would receive. Furthermore, there were striking trends between students who received lower grades compared to students whose grades remained on par with teacher assessments or even improved upon them. Many more students from state schools saw their grades cut, therefore limiting their prospective higher education opportunities, while students from independent and private schools saw their grades improving (Allegretti, 2020).

The results caused an outcry in the United Kingdom and sparked protests up and down the country (Castle, 2020). In response, the government U-turned on the move to put A-level assessments in the hands of the Ofqual grading algorithm, and fell back on teacher assessments whenever they were higher than those awarded by the algorithm (Taylor, 2020).

Accuracy, completeness and reliability

The heavy cost and capacity requirements of data collection, cleaning and analysis prevents some contexts from developing a data culture that would improve general trust in the reliability and usefulness of data due to higher data literacy and exposure in education settings. The complexity of assessing and auditing the reliability and accuracy of data analysis and modelling for insights could be heavily context dependent. Drawing inaccurate insights in the context of recommending one course as opposed to another is of a different order of sensitivity to inaccurately grading a child in a consequential exam. A calibration that involves assessing production readiness needs contextual assessment. In systems where learning data are readily available, these data represent only limited aspects of the past, yet are often used to predict the future. This can limit outcomes and individual development if not done properly and well.

In this context, some of the main barriers to unlocking the beneficial potential of Data for Learning relates to the availability and quality of the data, which can undermine accuracy, completeness and reliability. Currently, many education data systems around the world lack the ability to capture real-time data on teaching and learning, and this prevents school systems from monitoring participation, progress and outcomes. It is important, therefore, to consider how to enhance the analysis of data that may typically be messy, restricted or incomplete. On the other hand, the pressure to collect as much data as possible to present a more holistic picture of student progress is not without its risks. As evidenced by the study in Box 14, pressures to produce complete and reliable data can create a culture of mistrust in teacher judgement, pressuring pedagogical proof in all aspects of student learning.

Box 14. Pressure for "infinite data collection" seen as detrimental to student learning and teacher trust in Queensland, Australia

A 2021 study of teachers in Queensland, Australia (Daliri-Ngametua *et al.*, 2021) revealed a perception that teachers were constantly engaged in collecting student data to justify their teaching practices with "hard evidence" of student growth (p. 7). This created a culture of mistrust in teachers' professional judgement from both school leaders and parents. Schools increased auditing and surveillance, and parents pressured teachers to provide them with material, measurable data to prove claims about learner performance, classroom behaviour and teaching practices.

The volume of data that Queensland teachers were expected to collect was perceived as "detrimental to student learning because it took away the time and energy needed to focus on providing substantive learning opportunities to students" (p. 9). As a consequence, teaching became less about "doing" and more about performing progress, which in turn, diminished the power of teachers' professional judgement and strengthened the sanctity of explicit, quantifiable student learning metrics, which reflect only a partial view of learning experiences and outcomes (Hardy, 2021).

Literacy and accessibility

Another key challenge that cannot be overlooked concerns learners' abilities to take advantage of the new digital learning possibilities available to them. Digital skills and literacy are key components of the digital divide that can cut learners off from the lifelong learning opportunities their human rights afford them. Digital literacy has been defined by UNESCO as:

> the ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital technologies for employment, decent jobs and entrepreneurship. It includes competences that are variously referred to as computer literacy, ICT literacy, information literacy and media literacy.
> (UNESCO, 2018*a*)

Furthermore, digital literacy skills relate explicitly to SDG Target 4.4, with skills indicators 4.4.1 and 4.4.2 focusing on ICT skills and digital literacy skills. Beyond a definition, however, key questions relating to digital literacy remain. What are the skills required to engage with data at different levels of teaching, learning and education management? Furthermore, as they continue to evolve, how will these skills be defined and by whom? How will they be developed, assessed and certified? What would data proficiency look like? What is the impact of data choices (what, where and how data are collected) on students, teachers and systems?

Box 15. The costs of low levels of digital literacy among a population

In an attempt to ensure an efficient and effective roll-out, the COVID-19 vaccination campaign in Catalonia, Spain was administered via a system of SMS messaging and smartphone app notifications. Unfortunately, however, the lack of digital skills among certain segments of the population in Barcelona delayed the roll-out of the vaccine as citizens did not have the digital skills required to book an appointment through an app or via SMS (Rodríguez and Oliveres, 2021). The differences in the uptake between certain demographics and across varying neighbourhoods was stark, with the intersecting nature of the digital divide clearly on show as fewer citizens from poorer areas of the city signed up for the vaccination programme. However, the nature of the pandemic meant that, in this instance, not only did the lack of digital skills among certain subsets of the population marginalize them from public health efforts, it also had a degrading effect on wider public safety.

The training teachers receive to engage with student data in informed and productive ways is critically important. Frequently, school districts or systems will purchase a one-off, vendor-based training with a technology purchase that may include LMS, assessment and early warning systems, dashboards and other applications depending on the system's educational objectives (Mandinach and Gummer, 2021). Often these trainings focus on how to access data, rather than on how to use the data to inform pedagogy. If teachers are not supported by their schools to understand how to contextualize the information their students produce to inform their teaching practice, then the data risks being misused, leading to simplifications or misinterpretations.

Additionally, the issue of accessibility of data skills and literacies for individuals with special needs should be considered in the design of data systems. Data attached to students with disabilities is often of a sensitive nature, and some countries, such as the United States, have particular legal requirements related to sharing the data of learners with disabilities, as protected by the Individuals with Disabilities Education Act and the Children's Online Privacy Protection Act. Although such legal frameworks seek to prevent the abuse and misuse of student data relating to learners with special needs, obtaining genuinely informed consent with the understanding that sharing such data may negatively impact future educational or employment opportunities remains a challenge (Stahl and Karger, 2016).

Box 16. Understanding the relationship between literacy and transparency

Who bears the burden of informed consent?

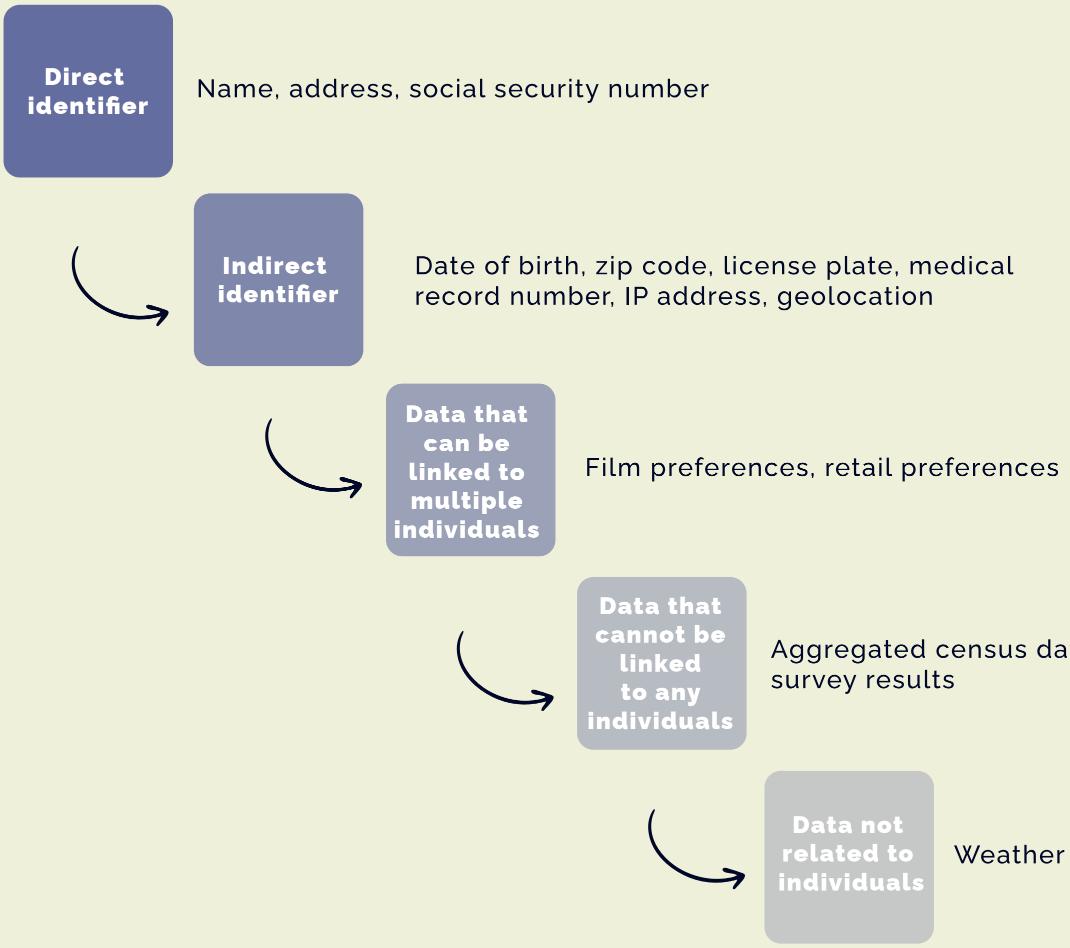
Transparency ends up as a form of free labour, where we are burdened with disinformation or misinformation but deprived of the capacity for meaningful corrective action. What results is a form of neoliberal "responsibilization", in which the public becomes burdened with duties it cannot possibly fulfil: to read every terms of service, understand every complex case of algorithmic harm, fact-check every piece of news. This shift in responsibility makes it, implicitly, our fault for lacking technological literacy or caring enough about privacy, never mind that vast amounts of money and resources are poured into obfuscating how our data are collected and used. This is the crux of the problem. Transparency is often valued as the great equalizer, a way to turn the tables on those in power and to correct the harms of technological systems. However, sometimes what you need to correct abuses of power is not more information, but rather a redistribution of power.

Privacy, consent, safety and security

Privacy. Privacy and security have become possibly the most obvious concerns relating to the collection of all types of personal data. Data can be compromised, exposing them to a plethora of risks and threats, such as identity theft and blackmail. Within education, however, there are further issues to consider. For example, sensitive dimensions of student identity, such as family income, special needs, counselling files, grades, addresses and contact information, can be uncovered if a school data management system is compromised. Even detailed aspects of students' socio-economic backgrounds can become visible if students participate in video classes that can put the interiors of their homes on display, and from this companies could accrue potentially marketable information.

Anonymizing student data at the individual level does not go far enough to protect learner privacy. If the security of the system as a whole is weak, then individual data can easily be triangulated by a hacker who combines multiple databases to enable the re-identification of individuals (Quinton and Reynolds, 2018). How do we ensure data privacy and anonymity when data mining makes it possible to de-anonymize apparently anonymized data?

Box 17. How data can be de-anonymized



According to the *Georgetown Law Technology Review*, there is a spectrum of identifiability that resembles a staircase when it comes to personal data (Lubarsky, 2017). The top of the spectrum includes directly identifiable data, such as an individual's name, social security number or phone numbers. The spectrum continues through indirect identifiers such as zip codes and dates of birth, data that could apply to multiple individuals including personal preferences, and so on. It is direct identifiers that can lead to data de-anonymization if they are discoverable in data that are supposed to have been de-anonymized. There are three main techniques for de-anonymizing data, which can all be used in tandem in attempts to de-anonymize data: combing datasets, pseudonym reversal and inadequate de-identification in the first place. Attempts to de-anonymize data can identify the real identities of individuals using direct identifiers.

Moreover, the integrity of data used for monitoring and credentialing processes can itself be called into question, as learner identity needs to be secure enough for digital credentials to be trustworthy. Important questions remain then, not just concerning the privacy and anonymity of learners, but also relating to the legitimacy of Data for Learning as a tool. The organization EdSAFE AI Alliance is attempting to address the application of data rights in education spaces, and is leading an international effort to develop benchmarks, standards and certifications to establish trust for the use of AI tools in education (EdSAFE AI Alliance, n.d.).

Consent. Another tension that stems from the protection of learners' data is the issue of meaningful consent. Although providing consent is currently the main feature of most personal data protection efforts, enabling meaningful consent remains one of the most difficult challenges in the digital context. It is possible to request consent using both opt-in and opt-out techniques, however, policies are often very long, hard to find and difficult to understand. As such, consent often only reflects the need to access the educational services rather than a true acceptance of the terms presented in the privacy policy. How do we ensure that learners are providing genuine consent that does not require them to read pages of dense text, and that is easily understood? How do we understand the temporality of consent? If a learner consents to their data being used at one specific moment, does that also mean they have consented to its reuse in the future, or to its inclusion in a larger dataset, the use for which was not disclosed at the moment the learner consented?

Box 18. Predicting challenges to consent in the education sector

In one study, researchers used AI to scan thousands of babysitters' profiles available on Facebook, Twitter and Instagram. The researchers then used data analytics to rate the "risk" posed by these babysitters to children. Parents were able to access these ratings and use them to decide whether or not to hire the babysitters (Harwel, 2018.). AI can similarly analyse people's speech and even their facial expressions to come up with appraisals of their trustworthiness and competence (Youyou *et al.*, 2015). These babysitters never gave consent for their data to be analysed in such a way, nor were they informed that this was occurring. Moreover, it is possible that many of these babysitters did not even know that researchers possessed the tools to automatically scan through their profiles and generate this type of risk score. While this type of research is not institutionally supported, one can imagine that such analysis might one day be used to predict how risky it is for an educational institute to recruit or promote a certain teacher or student. It is important to anticipate such concerns and plan for them accordingly.

Safety and security. When learner data are stored online it becomes much easier for malicious actors to access and exploit and, increasingly, breaches of this nature can expose the data of huge numbers of learners all at once. The question is, then, how do we prevent and mitigate data breaches? There are efforts to safeguard learners at both the local and international levels. At the local or school level, learner data should be protected from exploitation, but also from the threat of security breaches and school cyberattacks. At the system level, there is some debate as to whether these protections are ensured by existing international standards that apply to all individuals, such as the GDPR, or whether a specific international normative framework for the learning context is needed. If such a normative framework were developed to safeguard learner data, then it should include the voices of teachers, schools and local agencies, whose efforts should not be sidetracked by the introduction of new regulatory instruments.

The challenge of providing safety and security for learner data is deepened by socio-economic disparities, since the maturity of protections aligns with a country's economic development. Of the 76 countries reviewed by a recent MIT and Infosys study on regulatory frameworks for the use of cloud models, the two low-income countries ranked 71st (Uganda) and 76th (Ethiopia). Lower middle-income countries did not rank higher than 43rd (MIT Technology Review, 2022). This puts students in less wealthy areas at a disadvantage. On the supply side, international standards organizations that influence EdTech providers should be considered. In the United States, the Software and Information Industry Association has influence over the EdTech industry. Influencers such as the State Educational Technology Directors Association, the Consortium for School Networking and the International Society for Technology in Education also help to establish standards for EdTech development and have created human and digital forums that match student and educator needs with the capabilities of digital solutions. Developers would pay attention, creating products with data protection standards designed for the more developed countries, where revenue resides. The standards would therefore also apply to students in less developed nations, whose legal data protections may not be as heavily regulated.

Security breaches are becoming increasingly common. A McKinsey Global Institute (2021) report outlines that:

> **Breaches can occur during transfer of data, or at any institution involved in the open data ecosystem, such as a bank or fintech. For example, when data transfer is achieved via APIs, a hacker who breaches such an API can hijack any apps that use the interface to collect data.**

In the context of learning, this means that any data passing through a student information system or a digital learning programme faces the threat of data compromise. Naturally, with the use of digital platforms and programmes growing globally during the period of COVID-19 school disruptions, the number of cyberattacks and data breaches have also increased (Levin, 2021). For example, the publishing company Pearson, known for its textbooks, was fined USD 1 million to settle charges that it had "misled investors about a 2018 cyber intrusion involving the theft of millions of student records, including dates of births and e-mail addresses, and had inadequate disclosure controls and procedures" (United States Securities and Exchange Commission, 2021).

Box 19. Illuminate Education data breach in the United States

A data breach that occurred during a cyberattack on Illuminate Education in the United States in January 2022 is so far known to have affected nearly 2 million students across the country. The breach has affected learners and educators in five different states with over 820 000 students impacted in New York City alone (Kuykendall, 2022).

Officials from some of the affected school districts have stated that data included in the breach include learners' names, dates of birth, races or ethnicities, and test scores. Representatives from at least one of the districts indicated that the breach also included particularly sensitive information such as behavioural records, tardiness rates, information about disabilities and migrant status (Singer, 2022). Breaches of this nature can have long-lasting consequences for the affected learners, with sensitive and confidential information becoming available for public scrutiny.

Unfortunately, the full extent of the Illuminate Education data breach is yet to be known as the company, working with the stated aim of assisting education partners and educators reach new levels of student performance through the utilization of data, claims on its website to serve over 5 000 schools nationwide, covering a total of over 17 million enrolled students in the United States.

Many local, regional and national education bureaux simply do not have the human or capital resources to do this in a best-practice, sustained manner. Public schools may have fewer resources than private corporations to train their students and staff and to configure and secure school networks and devices to prevent online disruptions. As such, ensuring cyber protections is an equity issue, as learning environments with fewer resources are more vulnerable to attack.

Financial sustainability

The costs and consequences of expanding learning data models are key challenges to consider when weighing the potential benefits of these systems against the potential risks to environmental well-being. In parallel with concerns about data architecture, there are great concerns about the deployment and integration of new data practices where these must be incorporated into already under-resourced education systems. The adoption of emerging data practices would require major investments in hardware, software and human resources to support the shift away from centrally managed data architecture towards more vertically and horizontally integrated management systems.

The paradox is that the three countries with most expensive mobile data per gigabyte are in Africa. Globally, there is a 30 000 per cent difference between the cheapest data price and the most expensive (Ang, 2020), with India ranking the cheapest at USD 0.09 per gigabyte, and Malawi ranking the most expensive at USD 27.41 per gigabyte.

Box 20. Zero-rating education content

Zero-rating education content and access to Data for Learning is a measure that has been adopted by many countries during the COVID-19 pandemic. This has been an effort on the part of governments and telecommunication providers to ensure continuity of learning. The table below provides examples of zero-rating efforts.

Table 11. Zero-rating programmes across countries.

Country	Zero-rating programme
Democratic Republic of the Congo	Vodacom DRC has worked with the national government to offer a zero-rated education platform to all subscribers. The platform offers students content in mathematics, sciences, computer science, economics and finance
Ghana	MTN and Vodafone have zero-rated access to a number of educational sites. MTN offers subscribers a daily allowance of 500MB to explore government sites.
Jamaica	The Ministry of Education, Youth and Information partnered with One-On-One Educational Services and FLOW to zero-rate access to a national e-learning platform for two weeks.
Jordan	Internet service providers have zero-rated access to the Darsak e-learning platform between 6am and 4pm each day.
Kenya	Safaricom has zero-rated access to the Longhorn and Visuasa e-learning platforms. Students can have a daily allowance of 250MB to explore educational content for 60 days.
Malawi	The Ministry of Education Science and Technology and Telecom Networks Malawi have zero-rated access to lessons through the Ministry's website.
Paraguay	The government has an agreement with Microsoft to cover the e-learning needs of 1,200,000 students and 60,000 teachers at zero cost
Rawanda	Rwanda's Ministry of Education and Ministry of ICT and Innovation partnered with Airtel and MTN to zero-rate access to the government's e-learning platform.
South Africa	Vodacom, MTN, Telekom and C Cell have zero-rated access to e-learning platforms for current school, university and T-VET students.
United Republic of Tanzania	Vodacom has zero-rated access to the Shule Bora e-learning platform.
Zimbabwe	Econet has zero-rated access to the Ruzivo Digital Learning platform. The Zimbabwean government has officially endorsed this e-learning system.

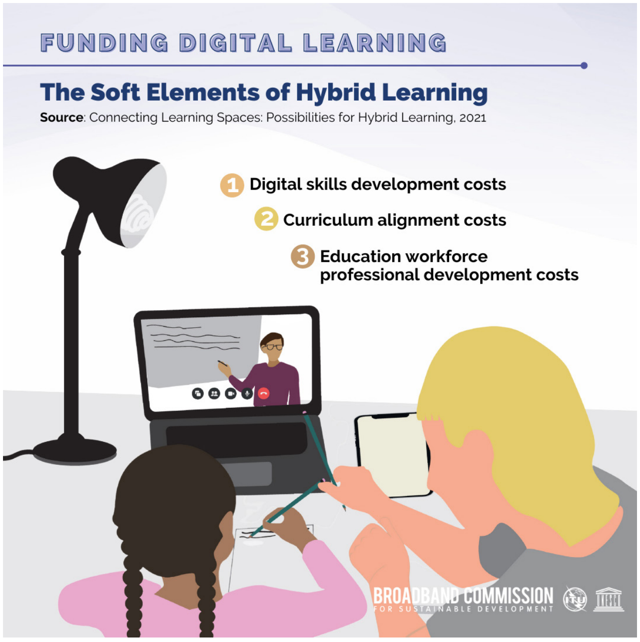
Source: McBurnie *et al*, (2020).

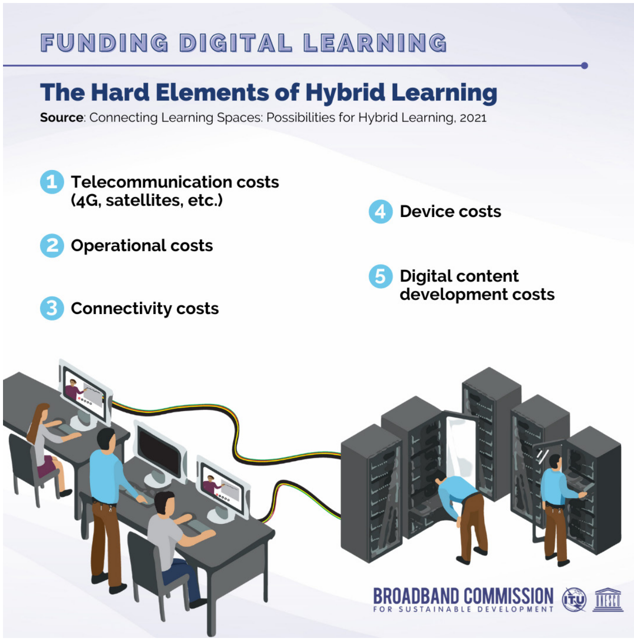
During the COVID-19 peak period, broad agreement was reached that zero-rating needs to be governed by a set of fair use principles to protect both the operators and the beneficiaries of zero-rating. These fair use principles would clearly define what students and learners have access to, and what counts as abuse. They would also offer detail around the point at which a service provider can cut off an individual subscriber when they have exceeded a certain level of data usage.

In some cases, governments are considering providing incentives of various sorts for Internet providers to maintain zero-rated access to education platforms, with "universal service funds" sometimes used to subsidize such practices. In other cases, telecom companies are committing to continuing these practices, whether as a way to differentiate themselves in competitive local markets or as part of their corporate social responsibility efforts (or some combination of the two).

As described in the Broadband Commission's (2021*b*) report on digital and hybrid learning, financing digital learning comprises both "hard" and "soft" elements, many of which are similar to the costs of financing Data for Learning. The figure below shows the hard costs, such as telecommunication, operational, connectivity, device and digital content development costs, and the soft costs, which include digital skills development, curriculum alignment and education workforce professional development.

Figure 14. Hard and soft elements of funding digital and hybrid learning.





Source: Broadband Commission (2021*a*).

When investigating the specifics of financing data systems, additional hard and soft costs are introduced. Hard elements include configuration and data storage requirements, hardware costs in the data hub, both capital and annual running costs, energy consumption and cooling water consumption requirements for data centres. Soft elements include database management, cloud-based computing, and training for data literacy and advanced data analytics. Additionally, how and where the software package was developed, either in-house or purchased from a vendor, influence costing. If developed in-house, the amount of time expended on development is a factor in the cost calculation, as is the administrative structure for data systems, in other words the maintenance and IT staff needed to keep the hardware and software operational. Thus, the human capacity needed to maintain data-rich systems comes with a large financial cost.

The adoption of certain frontier technologies raises additional issues for financing and planning departments. Even if countries could raise the funds, investments are often a complex set of moving targets that generate cumulative costs and require complicated forecasting and financing skills. A key feature of cloud-to-edge, for example, is the opaqueness of the costs involved. Cloud providers seemingly make it intentionally difficult to understand costs, meaning that a whole industry has recently emerged to help companies understand and compare cloud storage costs.[4]

Environmental sustainability

In addition to financing, the environmental costs of weaving data-informed learning into education spaces are often overshadowed by a conflation of digital transformation with the green transition. Selwyn (2021) takes a critical perspective on this by outlining the vast energy expenditure of the digital technology needed to sustain data-rich educational models. As Selwyn indicates, training a typical machine-learning model is estimated to emit the equivalent of around 300 000 kilograms of carbon dioxide, which is equivalent to the lifetime carbon emissions of five cars (Strubell *et al.*, 2019). Indeed, the technology industry is one of the largest greenhouse gas-emitting industries, and thus plans to expand data practices into all corners of education systems should consider processes to decrease the carbon footprint of data-processing activities (Belkhir and Elmeligi, 2018).

How, then, do we incentivize sustainable behaviour to ensure Data for Learning models support environmental well-being? From the social well-being perspective, it is important that investments in data infrastructure that supports the expansion of data-fuelled learning models – especially those that rely on big data – take the necessary measures to mitigate the environmental risk of energy-consuming big data analytics. Likewise, the reliance on such big data models must consider and protect those employed to clean big datasets, who may experience sensitive and traumatic material in the cleaning process (Perrigo, 2022). Therefore, the use of locally sourced, smaller datasets should also be considered to support a sustainable Data for Learning ecosystem.

4 For example, cloud storage costs are derived from a complicated formula based on: (1) the number of gigabytes stored, (2) how frequently the data are accessed and retrieved, (3) network bandwidth, (4) copy costs across multiple locations, such as America, Europe and Asia (particularly relevant for international enterprises), and (5) disaster recovery to move from on-premises storage to cloud and vice versa. All these factors accrue because there is a per-gigabyte cost each time servers in different domains communicate with each other, and another per-gigabyte cost to transfer data over the Internet.

4

Preliminary recommendations for Data for Learning

Figure 15. Data for Learning interim recommendations for further analysis.

There is a critical need to establish benchmarks and standards for utilizing data with appropriate agency. To do this, we must engage parties from across the entire ecosystem – including policy-makers, firms, organizations, education institutions and, most of all, learners, teachers, school staff and others – to build common definitions, practices and a visionary framework.

In 2021, the Data Futures Platform of the United Nations Development Programme published eight data principles and illustrated their alignment with existing international initiatives and frameworks. These principles are: (1) safeguard personal data, (2) uphold the highest ethical standards, (3) manage data responsibly, (4) make data open by default, (5) plan for reusability and interoperability, (6) empower people to work with data, (7) expand frontiers of data, and (8) be aware of limitations.

Building on this existing work and tying it to the broader mission of the Broadband Commission, this report underscores that, in the digital age, a high proportion of people are shut out of learning and economic activity due to barriers to digital and data equity. Data should be seen as a public good and Data for Learning should be seen as part of the vision that education is a public good that helps individuals reach their full potential.

We must harness the power of the digital revolution, including the data revolution, to ensure that equitable and quality education and lifelong learning are provided as a human right, with a particular focus on the most marginalized. This has been affirmed by the United Nations Secretary-General's Our Common Agenda, the United Nations Roadmap for Digital Cooperation, the Rewired Global Declaration on Connectivity for Education, and the International Commission on the Futures of Education. This issue is also at the centre of the Transforming Education Summit process.

Therefore, following the completion of the first year of its two-year remit, the WGDL proposes the following interim recommendations for further analysis to safely harness the power of Data for Learning:

1. **Develop and implement a whole-of-government vision and strategy on the use of Data for Learning, grounded in a rigorous understanding of the potential opportunities, benefits, limitations and risks.**

Data collection should always stem from specific desired outcomes, such as informing teaching practices and learning methodologies, and focus exclusively on the data required to achieve those specified outcomes. Furthermore, data should not be collected solely because it is technologically feasible or easy to do so, since collecting data without an intentional and transparent intended use can create a sense of surveillance that may limit the autonomy of teachers and learners, reproduce social inequalities or jeopardize individual security. For example, data collection should directly support the formulation of contextualized observations and recommendations.

There is also a need to unify interoperable data processes to bring nationally sourced and partner-sourced data into local analysis. A unified approach will not only help at the local level but also offer sector-wide support for decision-making at the district and provincial levels. For reasons including poor human resource capacities and the inability of education systems to fully integrate and apply learning data, many government actors do not fully leverage the many data sources available to them when making decisions and defining education policy. We must therefore facilitate multistakeholder partnerships to deliver the skills required to take full advantage of the new possibilities learning data can offer. Incentivizing a whole-of-government approach creates an environment conducive to overcoming the obstacles preventing many governments from efficiently utilizing data to support the public and common good for the benefit of learners.

Furthermore, inherent tensions exist between privacy and transparency, and there is a need to ensure that policy is data-informed rather than data-driven. Following the lead of the UNESCO Recommendation on the Ethics of AI, governance in the context of Data for Learning should therefore incorporate transparency and accountability as active aspects of protection, while also enabling effective monitoring of impact, enforcement and redress. This will require broad partnerships that can also reach beyond national borders to ensure effective regulation at the national and global levels, with an understanding of the inherent tensions in data ecosystems and the added sensitivity related to learning environments. Any strategy seeking to engage the true value of Data for Learning should also understand the limits of the data and seek to build contextualized, domain-informed and critical interpretations that balance data with common sense and shared values.

2. Establish a sustainable financing strategy for Data for Learning that benefits the public and protects learners' interests, and that is grounded in multistakeholder partnerships.

A key challenge standing in the way of educational data becoming a truly transformational technology for learning is financing. Collecting and processing educational data can be prohibitively expensive for smaller players such as schools. Duplicating data collection across disparate management systems is a costly and time-wasting process. Adopting a whole-of-government approach to fully appreciate the benefits of an inclusive and equitable lifelong learning ecosystem requires a sustainable financing model that covers both hard and soft elements. This will be expensive, with hard elements ranging from hardware and equipment, such as for data storage, in data centres to ongoing monthly and annual energy consumption, hardware cooling and other costs. The soft elements of data ecosystems cover a broader array of considerations, including database management costs, data and digital literacy training, software development and, where necessary, IT staff who can manage and maintain digital systems.

At the moment, many countries simply do not have the connectivity or infrastructure required to benefit from an advanced Data for Learning ecosystem. In fact, many countries do not even have tangible and exact data on network coverage and connectivity. Even high-income countries face challenges relating to data architecture when seeking to exploit frontier technologies, and continue to struggle to expand the existing digital infrastructure to remote communities. Without significant and continuing investment, implementing inclusive and equitable Data for Learning ecosystems that transform lifelong education opportunities for all will prove difficult if not impossible.

Furthermore, it is imperative that all sustainable financing mechanisms provide for the public good and deliver ongoing benefits to learners. The complexity and breadth of the hard and soft elements of effective Data for Learning ecosystems must come together to build powerful capabilities that can be harnessed for the good of education, rather than be seen as proprietary concerns. An example of how sustainable financing could help to achieve this would be support for the creation of open authentic data, which would be available to all parties and enable the continuous development of technological capabilities in the service of equitable education provision.

3. Strengthen critical data literacy and skills at all levels of the education system to spur improved regulation and inclusive innovation.

Data literacy and skills are required across the education ecosystem, from classroom learners to government ministers. Insufficient digital skills and literacy form a tangible pillar of the digital divide, preventing accessible digital transformation on a variety of scales. Many governments are currently unable to fully exploit the transformational potential that Data for Learning offers as they do not have sufficiently skilled human resources to do so, and many learners are unable to take advantage of the new digital education tools now available.

For data learning ecosystems to be fully inclusive, efforts should be made to strengthen data literacy, skills, including those of learners and students; teachers, educators, instructors and assistants; and administrators, governors, school leadership teams and policy-makers. Stakeholder support could follow the UNESCO framework, focusing on improving competencies including data and media literacy as well as

> ... the ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital technologies for employment, decent jobs and entrepreneurship.
> (UNESCO, 2018*a*).

Importantly, this includes empowering learners to collect data locally to facilitate autonomous pathways towards improving their understanding of local and global phenomena through the data they are able to collect and analyse themselves.

This will not, however, be a static endeavour. As data as a technology and tool for education continues to evolve, so will the skills required to understand and engage with data across all levels of the education ecosystem. The capacity building required to break down one of the pervasive pillars of the digital divide should therefore also address the need to understand an evolving landscape and the skills required to traverse it, and support infrastructure that can develop, assess and certify those skills. This will require the development of active assessment, measurement, monitoring and evaluation frameworks to ensure the continued and effective development of digital skills and literacy, even as the skills themselves evolve.

Beyond simply developing data collection, analysis, access and use skills, strengthening data literacy also sits at the heart of other key tensions such as data privacy, ownership, consent, transparency, accountability and security.

4. Prioritize the potential benefits of data to transform education by targeting education's enduring obstacles to assist informed and inclusive quality learning, teaching, management, planning and financing.

This interim report has identified key challenges and risks associated with Data for Learning practices, not least the simultaneous existence of more and bigger data than have ever been collected before and the unshakeable reality that educational data will only ever be able to capture certain aspects of educational practices. This dichotomy pulls at the potential effectiveness of inclusive and equitable Data for Learning ecosystems by shrouding information in data and at the same time offering key insights from analysis. In this environment, it is easy to lose sight of the real objectives of an endeavour, seeing only what the data shows us. It is important to remember that "data" is not information, and that information is not wisdom.

A Data for Learning ecosystem that delivers inclusive and equitable lifelong learning for all will work hand-in-hand with human understanding to ensure that data-informed policies govern with data rather than being governed by data. Consequently, it is important to focus on the benefits of data-informed learning for global efforts to transform education and tackle some of its most persistent problems. In this regard, and to build on the digital literacy outlined in the previous recommendation, policy-makers also need to understand the potential benefits Data for Learning policies offer, and the possible and plausible harms. This includes a strong, functional understanding of the legal frameworks in which Data for Learning policies will be considered, so as not to overlook unintended consequences and to fully comprehend associated direct and indirect costs. This will require collaboration between policy-makers and data science, education and privacy experts to support informed decision-making processes and effective Data for Learning policy-making.

Care must also be taken to ensure that the data itself is both inclusive and representative. Data governance models that support efforts to transform global education will require policies to promote and increase diversity and inclusiveness within datasets to protect against any bias or values that may exist in the contexts where the data were collected and produced. Furthermore, mechanisms for disclosing and combating any cultural, economic or social prejudices present in data, either by design or negligence, are vital, particularly in areas where data are scarce.

5. Harness multilateralism, solidarity and international cooperation to bridge the digital divide, nurture local data capacities and promote open authentic data for use by all parties to support more equitable education through the development of better tools including international standards and norms.

As this report has explored, addressing the challenges associated with inclusive data education ecosystems to fully reap the potential benefits of Data for Learning will require significant investment. However, the barrier to entry for countries with existing strong digital infrastructure and high levels of skilled human capital is much lower than for lower-resourced countries. Countries without strong digital infrastructure see contexts where resources are repeatedly consumed by collecting multiple instances of the same learning data that often exist in silos across education management systems.

Echoing this unfortunate reality are increasing claims of national data sovereignty that seek to push back on the open data movement and close down cross-border data flows. International solidarity and cooperation are vital if learners' human rights are not to be infringed in the pursuit of maintaining national sovereignty over datasets. This is particularly true in the cases where data sovereignty is being pursued as a defence against data colonialism, which sees companies and institutions in the Global North taking ownership over data from the Global South. There is a critical balance to be struck between regulating data ownership to prevent the exploitation or commercialization of learner data, particularly by distant companies that have no local economic impact, and incentivizing data openness to democratize data as an educational resource and improve transparency in education governance.

Openness is therefore a critical area, as open-licensed data create a more flexible data ecosystem, helping to diversify the use of Data for Learning. Data openness is a legal condition that guarantees permissions that can maximize the flexibility of public use and engagement of data, and unleash many uses that promote data literacy and skills building in an equitable way. Open data enables individuals and organizations to access and reuse data to innovate and collaborate, in a transparent context that allows citizens to work with governments to plan and monitor improved public services, while businesses untangle potential markets and new data-driven products.

Governments, particularly in developing countries, should embark upon strategies focusing on open data skills, policies and programmes that include awareness campaigns within government departments. These skills are indispensable for building and supporting the interoperability of the open data ecosystem and should be an essential part of it.

5

What's next? Towards a vision for Data for Learning

Following the first year of this unique two-year Broadband Commission WGDL, this interim report has taken important steps towards defining and refining the meaning of the term Data for Learning. This includes articulating a relevant taxonomy that can facilitate the realization of genuine benefits for the common good and proposing recommendations on promoting inclusive, equitable and successful Data for Learning ecosystems. However, despite this progress, more work needs to be done on visualizing the interplay that occurs between the different aspects, levels and tiers of data that make up the education data ecosystem.

As this interim report has shown, data as a technology is not simply a tool but rather a double-edged sword. It is critical to understand both the opportunities and risks that exist at the individual, local and global levels across the diverse tiers of data identified in the report. This understanding needs to incorporate datasets, producers, brokers, consumers and regulators, and be broad in scope, encompassing data infrastructure and learning ecosystems, data skills and competencies, and an awareness of data ethics when defining governance practices.

The complexity of the task at hand is apparent but this report, and the case studies presented in it, have made clear the transformational potential of this understanding to unlock greater learning experiences, improved teaching practices and more nuanced insights into educational governance and institutional administration. As the WGDL moves forward into its second year, the final section of this interim report raises some key questions and considerations to drive further discussion on the implications of Data for Learning for Working Group members and all relevant stakeholders.

Ethical tensions to consider

The following competing scenarios could help us to understand the complexities involved in the ethical implementation of universal principles within data-driven systems in the context of learning. The scenarios capture a range of issues that are either already salient or likely to grow in importance as we move forward.

1. Using data to improve the quality and efficiency of services while respecting the privacy and autonomy of individuals. Machine learning and big data are already being used to improve various public services (including healthcare, education and social care). These improvements could be hugely beneficial to citizens, but require large amounts of personal data, raising concerns about how best to protect privacy and ensure meaningful consent.

2. Using algorithms to make more accurate predictions and decisions versus ensuring fair and equal treatment. This tension arises when public or private bodies base decisions on predictions about the future behaviour of individuals (e.g. when probation officers estimate the risk of reoffending) and when they employ machine-learning algorithms to improve their predictions. These algorithms may improve accuracy overall but discriminate against specific subgroups for whom representative data are not available.

3. Reaping the benefits of increased personalization in the digital sphere versus enhancing solidarity and citizenship. This is extremely pertinent in the field of education and learning where "personalized learning at scale" is touted as a holy grail, and both companies and governments use personal data to tailor the learning pathways, messages, offers and services individuals see. This personalization can make it easier for individuals to find the right products and services for them, but differentiating between individuals in such fine-grained ways may threaten societal ideals of citizenship and solidarity.

4. Using automation to make people's lives more convenient and empowered versus promoting self-actualization and dignity. Automated solutions may genuinely improve people's lives by saving them time on mundane tasks that could be better spent on more rewarding activities, but they also risk disrupting some of the practices that are an important part of what makes us human. With automation we may see the gifts of arts, languages and science become more accessible to those who were excluded in the past, but we may also see widespread deskilling, ossification of practices, homogenization and loss of cultural diversity.

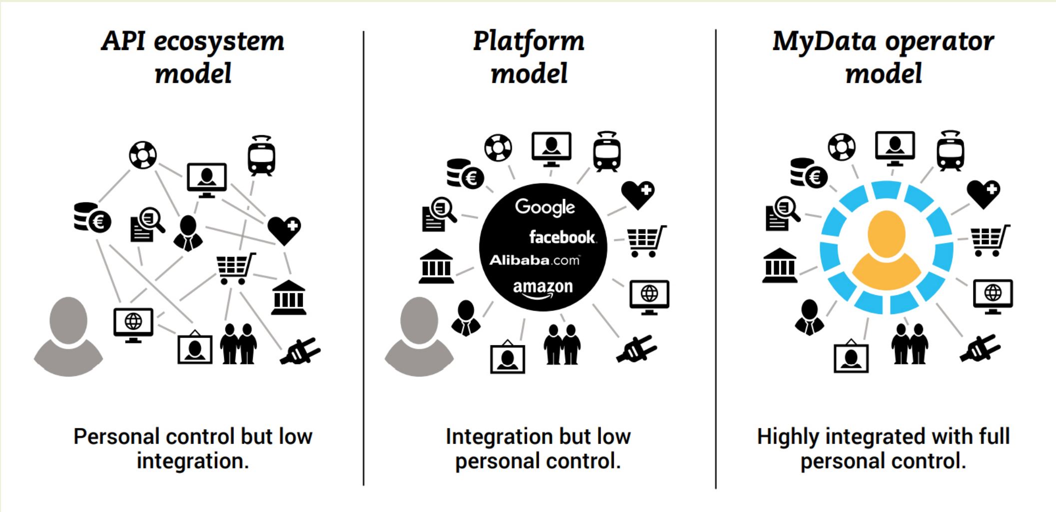
Understanding personal data

The fair use of personal data is one of the most critical issues for shaping a sustainable and prosperous digital society. Personal data have significant social, economic and practical value. Such data hold the key to improving a range of services and products provided by governments, companies and organizations. But personal data-based services must be built on mutual trust. Today, storing personal data can be seen as a liability, while having permission to use a piece of data is an asset. Companies who grasp this early enough are in a better position in the newly emerging data economy. Whole industries – including the energy, health and well-being, and finance sectors – are already being disrupted by this trend.

Box 21. Data governance: MyData (implemented by MyData Global and funded by the Ministry of Transport and Communications, Finland)

MyData is an umbrella term for a human-centric approach to personal data. The core idea is that individuals should be in control of data about them. The MyData approach aims to strengthen digital human rights while also opening up new opportunities for individuals to access the practical tools needed to exercise them. The minimum implementation of MyData is that an individual can download data about themselves in a machine-readable format for their own analysis.

Figure 16. Comparisons between MyData operator model and legacy API and platform models.



Source: Poikola *et al*. (2020).

Figure 16 shows that in the API ecosystem model (left), if the number of services increases, the number of connections will increase even faster. Centralizing data management to platforms (centre) facilitates application development, but there is no incentive for different platform players to seek interoperability. Compared to the platform model, the MyData operators' infrastructure (right) is robust and scalable because it is not dependent on any one organization providing the infrastructure.

MyData is an alternative vision which offers guiding technical principles for how we, as individuals, can have more control over the data trails we leave behind in our everyday actions. The core idea is that we should have an easier way to see where our personal data go, specify who can use the data, and alter these decisions over time. Legislation, regulation and technological changes can all contribute to the realization of MyData. Education systems should pay close attention to the evolution of this initiative and how they too can contribute to a realization of this vision.

Anticipating challenges to the future of Data for Learning

The most challenging issue for education systems will be to figure out the contours of Data for Learning and data management systems that allow for a balance between traditional methods and the benefits of emerging practices. There is a need for contextualized solutions that are forward-looking while remaining sensitive to capacity constraints. In low resource and fragile contexts, improvements in existing systems may be more effective than trying to adopt advanced technologies. Understanding the evolving dimensions of the right to education in an increasingly digital learning landscape deserves nuanced discussion. As discussed in Ravitch (2010), there may be concerns that have yet to emerge, or privacy issues that need to be resolved in relation to the implications of future arrangements.

Data for machine learning and unlearning. It is straightforward to delete a customer's data from a database and stop using it to train future models. But what about models that have already been trained using an individual's data? These are not necessarily safe; it is known that individual training data can be exfiltrated from models trained in standard ways via model inversion attacks (Veale *et al*., 2018). Regulators are still grappling with when a trained AI model should be considered to contain individuals' personal data in the training set and what the potential legal implications may be.

Data protection and privacy have been the subject of much discussion as more and more individuals come to realize just how much personal information they are sharing through the countless apps and websites they regularly visit. Many people are concerned. Recent government initiatives such as the EU's GDPR are designed to protect individuals' data privacy, with a core concept being "the right to be forgotten". The bad news is that it is generally difficult to revoke things that have already been shared online or to properly delete such data. Facebook, for example, launched an "Off-Facebook activity" tool – previously called "Clear history" – which the company says enables users to delete data that third-party apps and websites have shared with Facebook. But as the *MIT Technology Review* notes, "it's a bit misleading – Facebook isn't deleting any data from third parties, it's just de-linking it from its own data on you." Machine learning is increasingly viewed as exacerbating this privacy problem (Synced, 2020). Machine-learning applications are driven by data, and this can include collecting and analysing information such as personal e-mails or even medical records. Once fed into a machine-learning model, such data can be retained forever, putting users at risk of all sorts of privacy breaches.

Switching to a researcher's perspective, a concern is that if and when a data point is actually removed from a machine-learning training set, it may be necessary to retrain downstream models from scratch. In a new paper, researchers from the University of Toronto, the Vector Institute, and the University of Wisconsin-Madison propose SISA training (Bourtoule *et al*., 2020), a new framework that helps models "unlearn" information by reducing the number of updates that need to be computed when data points are removed. "The unprecedented scale at which machine learning is being applied on personal data motivates us to examine how this right to be forgotten can be efficiently implemented for machine-learning systems," the researchers explain. Having a model forget certain knowledge requires that some particular training points be made to have zero contribution to the model. But data points are often interdependent and can hardly be removed independently. Existing data also work continuously with newly added data to refine models. One solution is to understand how individual training points contribute to model parameter updates. But as previous studies have shown, this approach is only practical when the learning algorithm queries data in an order that has been decided prior to the start of learning. If a dataset is queried adaptively – meaning a given query depends on any queries made in the past – this approach becomes exponentially more challenging and thus can hardly scale to complex models such as deep neural networks. All this is to say that every new data-driven breakthrough could challenge previously assured safeguards. This is even more important and relevant in the case of lifelong learning, which often involves data relating to minors.

Conclusion: Areas for further analysis, data collection and policy dialogue

The above tensions are important and represent areas where exploring tensions is likely to be fruitful for data ethics. Going forward, further similar areas can and should be identified. As well as building an in-depth understanding of the interplay throughout and across the education data ecosystem, critical questions still need to be answered if Data for Learning is to truly contribute to the transformation of education. These questions include:

- Where data are being used to serve a particular goal or value, or for "social benefit" in general, what risks to other values are introduced?
- Where might uses of data-driven systems that benefit one group, or the whole population, have negative consequences for a specific subgroup? How do we balance the interests of different groups?
- Where might applications of data-driven systems that are beneficial in the near term introduce risks in the long term? How do we balance short- and long-term impacts on society?
- Where might future developments in data-driven systems, including AI, either enhance or threaten important values, depending on the direction they take? (Whittlestone *et al*., 2019)

Digital divides exist, as many people remain removed from the digital transformation of education for a variety of reasons, including connectivity, access to devices, and skills gaps. However, once these digital divides are bridged, a data divide emerges. Due to inequalities in capacity, different parts of the world are at very different stages of developing the necessary safeguards and protections to ensure that learner data are private, secure and protected from unethical commercialization or compromise. This data divide is drawn across socio-economic lines that leave learners in lower-income areas vulnerable to rights abuses, and even ignorant of the knowledge that they possess data rights. It is clear, therefore, that the data divide is a rights and equity issue, as well as an issue of values. Even if data are reliable, secured and trustworthy, they cannot capture a full picture of learning. If Data for Learning is to truly help transform education, they must be socially contextualized, used skilfully, safe and secure, and, above all, they must serve the primary purpose of improving teaching and learning experiences.

References

Allegretti, A. 2020. A-Level results: Government accused of "baking in" inequality with "boost" for private schools. Sky News. https://news.sky.com/story/35-of-a-level-results-downgraded-by-one-grade-figures-reveal-12048251

Ang, C. 2020. What does 1 GB of mobile data cost in every country? Visual Capitalist. https://www.visualcapitalist.com/cost-of-mobile-data-worldwide/

Ayling, J. and Chapman, A. 2021. Putting AI ethics to work: are the tools fit for purpose? *AI Ethics* 2, 405–429 (2022).

Behzadi, Y. 2021. Synthetic data to play a real role in enabling ADAS and autonomy. Automotive World. https://www.automotiveworld.com/articles/synthetic-data-to-play-a-real-role-in-enabling-adas-and-autonomy/

Belkhir, L. and Elmeligi. A. 2018. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *J. Clean. Prod*. Vol. 177, pp. 448–463.

Berridge, E. R. 2020. Impact of Covid-19 on summer exams. Statement made by Baroness Berridge on 23 March 2020. UK Parliament. https://questions-statements.parliament.uk/written-statements/detail/2020-03-23/HLWS170

Bertino, E., Dai, C., and Kantarcioglu, M. 2009. The challenge of assuring data trustworthiness. X. Zhou, H. Yokota, K. Deng, Q. Liu (eds.), *Database Systems for Advanced Applications* (Vol. 5463). Heidelberg, Springer Berlin, pp. 22–33.

Bertino, E. and Lim, H.-S. 2010. Assuring data trustworthiness – concepts and research challenges. W. Jonker and M. Petković (eds.), *Secure Data Management*. Heidelberg, Springer Berlin, pp. 1–12.

Bookman, S. 2017. 15 huge supercomputers that were less powerful than your smartphone. The Clever. https://www.theclever.com/15-huge-supercomputers-that-were-less-powerful-than-your-smartphone/

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D. and Papernot, N. 2020. *Machine Unlearning*. arXiv. http://arxiv.org/abs/1912.03817

Broadband Commission. 2011. Working Group on Multilingualism. https://broadbandcommission.org/working-groups/multilingualism-2011/

Broadband Commission. 2017. Working Group on Education, 2017. https://broadbandcommission.org/working-groups/education/

Broadband Commission. 2019. Working Group on Child Online Safety. https://broadbandcommission.org/working-groups/child-safety-online-2019/

Broadband Commission. 2020. Working Group on School Connectivity. https://broadbandcommission.org/working-groups/school-connectivity-2020/

Broadband Commission. 2021*a*. *Connecting Learning Spaces: Possibilities for Hybrid Learning*. https://broadbandcommission.org/wp-content/uploads/dlm_uploads/2021/09/Digital-Learning-Report-Broadband-Commission.pdf

Broadband Commission. 2021*b*. Working Group on Digital Learning. https://broadbandcommission.org/working-groups/digital-learning-2021/

Broadband Commission. 2022*a*. Working Group on AI Capacity Building. https://broadbandcommission.org/ai-capacity-building/

Broadband Commission. 2022b. Working Group on Data for Learning. https://broadbandcommission.org/data-for-learning/

Broadband Commission. n.d. Achieving the 2025 Advocacy Targets: Universal connectivity, affordability, skills, access, equality and use. https://www.broadbandcommission.org/advocacy-targets/

Capgemini. 2021. Data ecosystems on the rise. https://www.capgemini.com/insights/expert-perspectives/collaborative-data-ecosystems/

Castellanos, S. 2021. Fake it to make it: companies beef up AI models with synthetic data. Wall Street Journal. https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601

Castle, S. 2020. Boris Johnson retreats in a U.K. exam debacle. The New York Times. https://www.nytimes.com/2020/08/17/world/europe/england-college-exam-johnson.html

Ceibal. 2022*a*. Behavioural interventions to promote attendance of those enrolled in AcreditaCB 2022. Internal report from Ceibal, with data from National Administration of Public Education, Uruguay.

Center for Democracy & Technology. 2019. Technological school safety initiatives: considerations to protect all students. https://cdt.org/insights/technological-school-safety-initiatives-considerations-to-protect-all-students/

Ceibal. 2022*b*. MATEC: Una manera distinta de enseñar matemática en clase [MATEC: A different way to teach mathematics in class]. https://www.ceibal.edu.uy/es/articulo/matec-una-manera-distinta-de-ensenar-matematica-en-clase

Ceibal. n.d.*b*. What is Ceibal? https://www.ceibal.edu.uy/en/institucional

Ceibal. n.d.*a*. CREA. https://www.ceibal.edu.uy/crea

Chen, A. n.d. Know the difference between data-informed and versus data-driven. @AndrewChen. https://andrewchen.com/know-the-difference-between-data-informed-and-versus-data-driven/

Couldry, N. and Mejias, U. A. 2019. Data colonialism: rethinking big data's relation to the contemporary subject. *Telev. New Media*, Vol. 20(4), pp. 336–349.

Coursera. 2022. *Global Skills Report 2022*. https://www.coursera.org/skills-reports/global

Daliri-Ngametua, R., Creagh, S. and Hardy, I. 2021. Data, performativity and the erosion of trust in teachers. *Cambridge J. Educ*. Vol. 52(11), pp. 1–17

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datagen. 2022. *Synthetic Data: Key to Production-Ready AI in 2022*. https://datagen.tech/ai/synthetic-data-key-to-production-ready-ai-in-2022/

Datasphere Initiative. 2022. *Datasphere Governance Atlas: Mapping organizations in the data governance ecosystem*. https://www.thedatasphere.org/programs/intelligence-hub/datasphere-governance-atlas/

Devaux, E. 2021. List of synthetic data startups and companies – 2021. https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42

Driver's Seat. n.d. https://driversseat.co/

du Boulay, B., Poulovassilis, A., Holmes, W. and Mavrikis, M. 2018. What does the research say about how artificial intelligence and big data can close the achievement gap? R. Luckin (ed.), Enhancing Learning and Teaching with Technology. London, Institute of Education Press, pp. 256–285.

DXtera Institute. 2022. OAD in Ed Community. https://dxtera.org/oad-in-ed/

EdSAFE AI Alliance. n.d. https://www.edsafeai.org/

Erstad, O., Flewitt, R., Kümmerling-Meibauer, B. and Susana Pires Pereira, Í. 2019. *The Routledge Handbook of Digital Literacies in Early Childhood*. London, Routledge, pp.79–92.

EU Commission. 2017. *European Interoperability Framework – Implementation Strategy*. Brussels, European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3A52017DC0134

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R. and Warschauer, M. 2020. Mining big data in education: affordances and challenges. *Rev. Educ. Res.,* Vol. 44, No. 1, pp. 130–160.

Frontier Technologies Hub. 2019. Frontier Data Study: *Releasing the power of digital data for development*. London, Frontier Technologies Hub. https://www.frontiertechhub.org/resources/blog-post-title-two-8c4t5

G20 Education Working Group Report, 2022, forthcoming

G20 Indonesian Presidency Education Working Group. 2022 (forthcoming)

Global Partnership for Education. 2021. *Joint Education Sector Monitoring in the Context of COVID-19*. Washington DC, Paris and Brussels, Global Partnership for Education.

Hammer, C., Ottaviani, J. and Kumar, R. 2021. Strengthening data use culture in West African countries. https://blogs.worldbank.org/opendata/strengthening-data-use-culture-west-african-countries

Hann, T. 2021. Synthetic data enables insurers to get more value from AI. https://www.propertycasualty360.com/2021/12/24/synthetic-data-enables-insurers-to-get-more-value-from-ai

Hardy, I. 2021. The quandary of quantification: data, numbers and teachers' learning. *J. Educ. Policy*, Vol. 36(1), pp. 44–63.

Harwell, D. 2018. Wanted: The "perfect babysitter." Must pass AI scan for respect and attitude. The Washington Post. https://www.washingtonpost.com/technology/2018/11/16/wanted-perfect-babysitter-must-pass-ai-scan-respect-attitude/

Hirsh-Pasek, K., Zosh, J. M., Hadani, H. S., Golinkoff, R. M., Clark, K., Donohue., C. and Wartella, W. 2022. A whole new world: Education meets the metaverse. Brookings. https://www.brookings.edu/research/a-whole-new-world-education-meets-the-metaverse/

Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Classical Ethics in A/IS*. New York, IEEE. https://ethicsinaction.ieee.org/wp-content/uploads/ead1e_classical_ethics.pdf

INTEF. n.d. About Insignias INTEF. https://insignias.educacion.es/en/node/119

Jobin, A., Ienca, M. and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell*. Vol. 1, pp. 389–399.

Jones, K., Asher, A., Goben, A., Perry, M., Salo, D., Briney, K. and Robertshaw, M. 2020. "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. *J. Assoc. Inf. Sci*. Vol. 71.

Jones Day. 2019. Proposed Algorithmic Accountability Act targets bias in artificial intelligence. https://www.jonesday.com/en/insights/2019/06/proposed-algorithmic-accountability-act

Kerr, K., Norris, T., & Stockdale, R. (2007). Data quality information and decision making: A healthcare case study. W-G. Tan (ed.), *Proceedings of the 18th Australasian Conference on Information Systems Doctoral Consortium (ACIS 2007)*. University of Southern Queensland, Toowoomba, Australia.

Klein, L. 2020. There's no such thing as raw data. Feed. https://feedmagazine.tv/interviews/lauren-klein-theres-no-such-thing-as-raw-data/

Koperniak, S. 2017. Artificial data give the same results as real data – without compromising privacy. MIT News. https://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303

Kuykendall, K. 2022. Illuminate data breach spreads to fifth state as Oklahoma City notifies parents. The Journal. https://thejournal.com/Articles/2022/05/17/Illuminate-Data-Breach-Spreads-to-Fifth-State-as-Oklahoma-City-Notifies-Parents.aspx

Levin, Douglas A. 2021. *The State of K-12 Cybersecurity: 2020 Year in Review*. K12 Six. https://k12cybersecure.com/year-in-review/

Livingstone, S., Stoilova, M. and Nandagiri, R. 2019. Children's data and privacy online: Growing up in a digital age. An evidence review. London, London School of Economics and Political Science. https://eprints.lse.ac.uk/101283/1/Livingstone_childrens_data_and_privacy_online_evidence_review_published.pdf

Locatelli, R. 2018. Education as a public and common good: Reframing the governance of education in a changing context. *Education Research and Foresight Working Papers Series*, No. 22. https://unesdoc.unesco.org/ark:/48223/pf0000261614/PDF/261614eng.pdf.multi

Lohr, S. 2018. Facial recognition is accurate, if you're a white guy. The New York Times. https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Lubarsky, B. 2017. Technology Explainers: Re-Identification of "anonymized" data. Georgetown Law Technology Review. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/

Ludic Design for Accessibility. n.d. All for play, play for all. https://www.ludicdesign.org/

Mandinach, E. and Gummer, E. 2021. *The Ethical Use of Data in Education*. New York, Teachers College Press.

May, M., Iksal, S., and Usener, C. 2017. The side effect of learning analytics: an empirical study on e-learning technologies and user privacy. G. Costagliola, J. Uhomoibhi, S. Zvacek, B. McLaren (eds). Computers Supported Education. CSEDU 2016. *Communications in Computer and Information Science*, Vol 739. Springer, Cham.

Mayer-Schonberger, V. and Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London, John Murray.

McBurnie, C., Taskeen, A., Kaye, T. and Haßler, B. 2020. *Zero-rating educational content in low- and middle-income countries*. EDTech Hub. https://docs.opendeved.net/lib/4W3D35BT/download/3NBMJTSW/McBurnie

McCabe, D. and Satariano, A. 2022. The era of borderless data is ending. The New York Times. https://www.nytimes.com/2022/05/23/technology/data-privacy-laws.html

McKinsey Global Institute. 2021. Financial data unbound: The value of open data for individuals and institutions. https://www.mckinsey.com/industries/financial-services/our-insights/financial-data-unbound-the-value-of-open-data-for-individuals-and-institutions

McStay, A. 2020. Emotional AI and EdTech: serving the public good? *Learn. Media Technol.*, Vol. 45(3), pp. 270–283.

Metz, R. 2020. There's a new obstacle to landing a job after college: Getting approved by AI. CNN Business. https://edition.cnn.com/2020/01/15/tech/ai-job-interview/index.html

Meyer, M., Zosh, J. M., McLaren, C., Robb, M., McCaffery, H., Golinkoff, R. H., Hirsh-Pasek, K. and Radesky, J. 2021. How educational are "educational" apps for young children? App store content analysis using the four pillars of learning framework. *J. Child. Media* Vol. 15(4), pp. 526–548.

Microsoft. 2017. Ludic Design for Accessibility. https://www.microsoft.com/en-us/research/project/ludicdesign/

Mission 4.7. n.d.*a*. https://www.mission4point7.org/

Mission 4.7. n.d.*b*. Guiding Principles for Transformative Education. https://sites.google.com/view/mission47/guiding-principles

MIT Technology Review. 2022. Global Cloud Ecosystem Index 2022. https://www.technologyreview.com/2022/04/25/1051115/global-cloud-ecosystem-index-2022/

Mohamed, S., Png, M. T. and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philos. Technol.* Vol. 33, pp. 659–684.

Morley, J., Cowls, J., Taddeo, M. and Floridi, L. 2020. Ethical guidelines for COVID-19 tracing apps. Nature. https://www.nature.com/articles/d41586-020-01578-0

Mutimukwe, C., Viberg, O., Oberg, L.-M., and Cerratto-Pargman, T. 2022. Students' privacy concerns in learning analytics: Model development. *Br. J. Educ. Technol*. Vol. 53(4), pp. 932–951.

ODI. 2020. Data Skills Framework. https://theodi.org/article/data-skills-framework/

OECD. 2019. *Mapping Approaches to Data and Data Flows: Report of the G20 Digital Economy Task Force.* https://www.oecd.org/sti/mapping-approaches-to-data-and-data-flows.pdf

OHCHR. 2021. General Comment No. 25 (2021) on children's rights in relation to the digital environment. United Nations. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation

Pangrazio, L. (2016). Reconceptualising critical digital literacy. *Discourse: Studies in the Cultural Politics of Education* Vol. 37(2), pp. 1-12.

Parkin, D. 2021. Ethiopia overhauls its education system with IOHK blockchain partnership. City A.M. https://www.cityam.com/ethiopia-overhauls-its-education-system-with-iohk-blockchain-partnership/

Perrigo, B. 2022. Inside Facebook's African sweatshop. Time. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

Perry-Hazan, L. and Birnhack, M. 2019. Caught on camera: Teachers' surveillance in schools. *Teach. Teach. Educ.* Vol. 78, pp. 193–204.

Poikola, A., Kuikkaniemi, K., Kuittinen, O., Honko, H., Knuutila, A. and Lähteenoja, V. 2020. MyData – *An Introduction to Human-Centric Use of Personal Data.* Helsinki, Ministry of Transport and Communications. https://www.mydata.org/wp-content/uploads/2022/07/mydata-white-paper-english-2020-2.pdf

Price, R. and Shanks, G. 2005. Empirical Refinement of a Semiotic Information Quality Framework. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE.

Quinton, S., and Reynolds, N. 2020. Digital Research as a Phenomenon and a Method. *Understanding Research in the Digital Age*. London, Sage Publications.

Ravitch, D. 2010. The myth of charter schools. Brookings. https://www.brookings.edu/articles/the-myth-of-charter-schools/

Regan, P.M. and Khwaja, E.T. 2019. Mapping the political economy of education technology: A networks perspective. Policy Futures *Educ*., Vol. 17, No. 8, pp. 1000–1023.

Rodríguez, P. and Oliveres, V. 2021. La Vacunación También Va Por Barrios: El Mapa de Barcelona Que Muestra Diferencias Entre Ricos y Pobres [The vaccination roll-out also differs between neighbourhoods: The map of Barcelona showing differences between rich and poor]. elDiario. https://www.eldiario.es/catalunya/vacunacion-barrios-mapa-barcelona-muestra-diferencias-ricos-pobres_1_7953642.html

Rodríguez-Segura, D. 2021. EdTech in developing countries: a review of the evidence. *World Bank Res. Obs*. Vol. 37(2), pp. 171–203.

Ruijer, E., Grimmelikhuijsen, S., van den Berg, J., and Meijer, A. 2020. Open data work: understanding open data usage from a practice lens. *Int. Rev. Adm. Sci*., Vol. 86(1), pp. 3–19.

Sambasivan, N. and Holbrook, J. 2019. Toward responsible AI for the next billion users. Interactions. https://interactions.acm.org/archive/view/january-february-2019/toward-responsible-ai-for-the-next-billion-users

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T. and Prabhakaran, V. 2021. Re-imagining algorithmic fairness in India and beyond. Google Research. https://research.google/pubs/pub50002

Selwyn, N. 2021. Ed-Tech within limits: Anticipating educational technology in times of environmental crisis. *E-Learning and Digital Media*, Vol 18(5), pp. 496–510.

Sheridan, M.P. and Rowsell, J. 2010. *Design Literacies: Learning and Innovation in the Digital Age*. London, Routledge.

Shin, D. 2019. Toward fair, accountable, and transparent algorithms: case studies on algorithm initiatives in Korea and China. *Journal of the European Institute for Communication and Culture*, Vol. 26(3), pp. 274–290.

Shiohira, K. and Dale-Jones, B. 2019. *Interoperable Data Ecosystems: An international review to inform a South African innovation*. Johannesburg, JET Education Services. https://www.jet.org.za/resources/interoperable-data-ecosystems.pdf

SIIA. n.d. https://www.siia.net/

Singer, N. 2022. A cyberattack illuminates the shaky state of student privacy. The New York Times. https://www.nytimes.com/2022/07/31/business/student-privacy-illuminate-hack.html

Slade, S., Prinsloo, P., Khalil, M. 2019. Learning analytics at the intersections of student trust, disclosure and benefit. *Proceedings of the 9th International Conference on Learning Analytics and Knowledge (LAK19).* New York, Association for Computing Machinery, pp. 235–244.

South Africa. 2013. *Every Child Is a National Asset*. Pretoria, Department of Basic Education, Republic of South Africa. https://sasams.co.za/repo/modules/00.pdf

South Africa. 2022. SA-SAMS SA School Administration and Management System. https://sasams.co.za/

Stahl, W. and Karger, J. 2016. Student Data Privacy, Digital Learning, and Special Education: Challenges at the Intersection of policy and practice. *J. Spec. Educ. Leaders*., Vol 29(2), pp. 79–88.

Statista. 2022. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/

Strubell, E., Ganesh, A. and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. College of Information and Computer Sciences, University of Massachusetts Amherst. https://ezproxy-prd.bodleian.ox.ac.uk:2246/doi/full/10.1177/20427530211022951

Synced. 2020. Machine unlearning: fighting for the right to be forgotten. SyncedReview. https://medium.com/syncedreview/machine-unlearning-fighting-for-the-right-to-be-forgotten-c381f8a4acf5

Szczepański, M. 2020. Is data the new oil? European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/646117/EPRS_BRI(2020)646117_EN.pdf

Tarrant, D. 2021. Data literacy: What is it and how do we address it at the ODI? ODI. https://theodi.org/article/data-literacy-what-is-it-and-how-do-we-address-it-at-odi/

Taylor, R. 2020. Statement from Roger Taylor, Chair, Ofqual: How grades for GCSE, AS, A level, Extended Project Qualification and Advanced Extension Award in Maths will be awarded this summer. London, Ofqual. https://www.gov.uk/government/news/statement-from-roger-taylor-chair-ofqual

The Economist. 2017. The world's most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

The GovLab, n.d. Open Data 500 Global Network. https://thegovlab.org/static/files/od500-onepager-cropped.pdf

Thorp, J. 2012. Big data is not the new oil. Harvard Business Review. https://hbr.org/2012/11/data-humans-and-the-new-oil

UIL. 2022. *5th Global Report on Adult Learning and Education: Citizenship education: Empowering adults for change*. https://unesdoc.unesco.org/ark:/48223/pf0000381666

UIS. 2020. *Data Innovation for Producing SDG 4 Indicators: An EMIS metadata global analytical report*. http://uis.unesco.org/sites/default/files/documents/ip65-emis_typology-final_en2.pdf

UIS. 2021. Learning Data Compact – UNESCO, UNICEF, and the World Bank unite to end the learning data crisis. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2021/06/Learning-Data-Announcement.pdf

UNESCO. 2018*a*. *A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2*. http://uis.unesco.org/sites/default/files/documents/ip51-global-framework-reference-digital-literacy-skills-2018-en.pdf

UNESCO. 2018*b*. UNESCO launches Five Laws of Media and Information Literacy (MIL). http://en.unesco.kz/unesco-launches-five-laws-of-media-and-information-literacy-mil

UNESCO. 2019. *Beijing Consensus on Artificial Intelligence and Education*. https://unesdoc.unesco.org/ark:/48223/pf0000368303

UNESCO. 2021*a*. Recommendation on the Ethics of Artificial Intelligence. https://en.unesco.org/artificial-intelligence/ethics

UNESCO. 2021*b*. Reimagining Our Futures Together: A New Social Contract for Education. Paris, UNESCO.

UNESCO. 2022*a*. *Minding the data: Protecting learners' privacy and security*. https://unesdoc.unesco.org/ark:/48223/pf0000381494

UNESCO. 2022*b*. *Re-Imagining the Future of Education Management Information Systems: Ways Forward to Transform Education Data Systems to Support Inclusive, Quality Learning for All.* Paris, UNESCO.

UNESCO. n.d. Open Data. https://www.unesco.org/en/communication-information/open-solutions/open-data

UNESCO and Commonwealth of Learning. 2019. *Guidelines on the development of open educational resources policies*. https://unesdoc.unesco.org/ark:/48223/pf0000371129

UNESCO and CCNGO. 2015. *2015 NGO Forum Declaration: Towards the right to inclusive quality public education and lifelong learning beyond 2015*. https://unesdoc.unesco.org/ark:/48223/pf0000233243

UNESCO and Dubai Cares. 2021. *Rewired Global Declaration on Connectivity for Education*. Paris and Dubai, UNESCO and Dubai Cares.

UNESCO-IIEP. 2022. Three offices, one mission: Discover the new global campus. http://www.iiep.unesco.org/en/three-offices-one-mission-discover-new-global-campus-14146

UNESCO-MGIEP. 2022. *Reimagining Education: The International Science and Evidence-based Education Assessment*. https://mgiep.unesco.org/iseeareport

Unitas 360. n.d. The Use of AI in Loan Decisions – Unitas Financial Services. https://www.unitas360.com/blog/the-use-of-ai-in-loan-decisions

United Kingdom. 2020. Schools, colleges and early years settings to close. London, UK Department of Education. https://www.gov.uk/government/news/schools-colleges-and-early-years-settings-to-close

United Nations. 1989. Convention on the Rights of the Child. Entered into force 2 September 1990.

United Nations. 2021. *Our Common Agenda: Report of the Secretary-General*. New York, United Nations Publications.

United States Department of Education. 2021. Data literacy webinar. https://www.ed.gov/sites/default/files/documents/stem/20211015-data-literacy.pdf

United States Federal Trade Commission. 2021. FTC approves final administrative consent order against Amazon for withholding customer tips from Amazon Flex drivers. https://www.ftc.gov/news-events/news/press-releases/2021/06/ftc-approves-final-administrative-consent-order-against-amazon-withholding-customer-tips-amazon-flex

United States Securities and Exchange Commission. 2021. SEC charges Pearson Plc for misleading investors about cyber breach. https://www.sec.gov/enforce/33-10963-s

Veale, M., Binns, R. and Edwards, L. 2018. Algorithms that remember: Model inversion attacks and data protection law. *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, Vol. 376.

Victorian Academy of Teaching and Leadership. 2022. Data literacy for school leaders. https://www.academy.vic.gov.au/professional-learning/data-literacy-school-leaders

Vincent, J. 2021. Amazon delivery drivers have to consent to AI surveillance in their vans or lose their jobs. The Verge. https://www.theverge.com/2021/3/24/22347945/amazon-delivery-drivers-ai-surveillance-cameras-vans-consent-form

Vision Empower Trust. n.d. Vision Empower. http://visionempowertrust.in/

Wakefield, S. 2017. Education system digital strategy: Transforming education for the digital age. Presented at Tertiary ICT Conference, Wellington, 6–8 September 2017. https://vdocument.in/transforming-education-for-the-digital-transforming-education-for-the-digital-age.html

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Kudakwashe, D., Gallagher, T. and McLachlan, S. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* Vol. 25, No. 3, pp. 230–238.

Whittlestone, J., Nyrup, R., Alexandrova, A. and Cave, S. (2019). The role and limits of principles in AI Ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200.

Williamson, B. and Hogan, A. 2020. *Commercialisation* *and privatisation in/of education in the context of Covid-19*. Education International Research. https://www.researchgate.net/publication/343510376_Commercialisation_and_privatisation_inof_education_in_the_context_of_Covid-19

World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington D.C., World Bank.

World Economic Forum. 2015. A brief history of big data everyone should read. https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read

Youyou, W., Kosinski, M. and Stillwell, D. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112(4), pp. 1036–1040.

Zhu, M. 2020. An algorithmic jury: Using artificial intelligence to predict recidivism rates. Yale Scientific. https://www.yalescientific.org/2020/05/an-algorithmic-jury-using-artificial-intelligence-to-predict-recidivism-rates

BROADBAND COMMISSION
FOR SUSTAINABLE DEVELOPMENT

ITU

unesco