Annexure 7: Guidelines for Access to Data for MSMEs and Researchers, and promoting data sharing with and by public and private sectors

The Digital Economy Working Group (DEWG) affirmed in its 2024 Ministerial Declaration that data is a critical input for economic development. Data governance is a cross-cutting aspect of all the priority areas of 2025 DEWG under South Africa's Presidency for enabling public and private data value creation. As part of this, the focus here is on increasing the supply of data, particularly local language data sets, and addressing the inequitable distribution of access that inhibits r the development and deployment of AI systems. The OECD Compendium, introduced by the G20 Brazilian presidency in 2024, makes the case for data access and data sharing across public institutions and with the private sector in the public interest, including in AI foundation models and applications.

This background underscores the increasing social and economic value of data, as well as the need for access to and sharing of data.

- The intangible and non-rivalrous nature of data, which allows for its reuse without diminishing its value or preventing its use by anyone else, offers a wide range of means of access and sharing for purposes of public and private value creation. It means that the market and social value associated with the re-use or circulation of data is larger than the value of primary data use for the individual. Access to and sharing of data can maximise the re-use and therefore the value of data across organisations, sectors and economies.

- Data access is deeply implicated in the development and deployment of Artificial Intelligence (AI), warranting consideration of such access as a major part of data governance. The Global Partnership on Artificial Intelligence (GPAI) identifies access to data as a mechanism to redress the currently highly uneven distribution of opportunities associated with AI development for more equitable and just outcomes.

In the context of rapidly evolving advanced data-driven technologies such as AI systems, access to high-quality, diverse, and relevant local language data that is privacy-preserving and secure is crucial for the ethical and effective training of AI models. Balancing access to such data with the

protection of personal privacy is not at odds with accelerated efforts to enhance openness and availability that is necessary for public and private value creation in the digital economy.

Primary constraints on the local development of various forms of AI have been identified as the availability of data, computing and advanced data skills.

Among these constraints, a major challenge for many data scientists, especially on the African continent, is the lack of access to the local language data necessary to build and deploy small or large LLMs in indigenous languages. In a context where data is already lacking by virtue of many forms of indigenous knowledge not being extensively written, recorded or digitised, and billions of people not having access to the Internet and being invisible in the giants data sets scaped from the web or from social networks, the lack of access to data to build AI LLMs exacerbates existing geographical, cultural and economic inequalities.

The lack of access to the digital-ready linguistic data where it does exist inhibits the development of AI that could potentially boost the local economy and empower tech start-ups and developers.

Drawing from the issue brief produced for the G20 by Research ICT Africa and the University of Pretoria's Data Science for Social Impact Research Group and Data Science Law Lab, with support from UNESCO, it is evident that numerous options exist to facilitate data access in a balanced manner. Together, these can enable greater openness of defined datasets by default; help clarify conditions for compulsory disclosures; and encourage voluntary agreements through promoting equitable licensing systems.

The possibilities set out below are especially focused on unlocking public and private sector data as a public resource, particularly for researchers, start-ups and micro-, small- and medium-enterprises (MSMEs). Many data scientists, especially in Africa, lack access to local language data to build and deploy small or large LLMs in indigenous languages.

Shaping the points below are international experiences and normative standards for data sharing, such as referenced by the Global Digital Compact and the Governing AI for Humanity Report released by the UN Tech Envoy High-Level Advisory Board on Artificial Intelligence. UNESCO's 2023 guidelines cover public sector information and data access, while also relevant is the organisation's 2023 publication of Data sharing to foster information as a public good. The Open Government Partnership brings together 75 countries with interests in open systems of

governance. The US-EU Trade and Technology Council in 2023 adopted shared principles on access to data from online platforms for researchers, and the 2022 African Union Data Policy Framework includes high-level principles on data access, open data, and sharing of data for development, research, innovation and knowledge building. The OECD in 2021 agreed a Recommendation on Enhancing Access to and Sharing of Data, and in 2025 it published Enhancing Access to and Sharing of Data in the Age of Artificial Intelligence.

A. Actions for producing pro-access policy

1. Optimise legal, policy and regulatory systems (including self- and co-regulatory systems) to ensure a balance whereby human rights-based access to and sharing of data are facilitated.
2. Provide for tiered access frameworks in terms of degrees of data openness, applicable to different actors, while also ensuring that these contribute to equitable opportunities for marginalised communities and MSMEs.
3. Commission and support research into the barriers to availing and accessing data, and into mitigations thereof, and require impact assessments for major data-sharing initiatives with clear metrics and accountable reporting mechanisms.
4. Adopt a critical and judicious approach through recognising that data sets are not neutral but generated for particular purposes, which may not be suitable for different purposes, and that there are also substantive debates about data colonialism and data justice.
5. Create a clear and predictable legal regime for data access, including a range of licensing frameworks that cover purpose and use specifications, incorporating both limits and allowances for any further sharing, and which also include provisions for legal disclosure of data partnerships, along with penalties in the event of data breaches or purpose violations. Give civil law effect to Creative Commons licenses and other public open licences.
6. Promote harmonised data access approaches across G20 members and more broadly, given that addressing such issues within a single country is not enough to deal with the realities of transnational data holdings, storage and flows.

B. Innovating institutions to be fit for purpose

1. Promote institutionalised mapping of data holdings and data set inventories through encouraging or requiring the publishing of dataset descriptions and codebooks in both public and private sectors. Avoid inappropriate burdens on MSMEs.
2. Convene policy discussions to prioritise high-impact domains (for example, giving special attention to: local language broadcasting, including by public broadcasters; transportation with real-time transport data; and environmental monitoring).
3. Assess markets that commercially trade in data, and assess how these might be governed in the wider public interest of expanded data access.
4. Strengthen and safeguard the authority, autonomy and accountability ofof relevant regulators from external interference, and ensure independent adjudication mechanisms and procedures to validate data requests in terms of public interest criteria and for cases of compelled disclosures of defined data sets.

C. Driving public sector data sharing

1. Foster the uptake of data pooling within the public sector for both internal operations as well as public services, and specify data sharing provisions in regard to relevant procurement contracts with the private sector.
2. Ensure open data by default in public entities, and for restricted data, provide systems for considering requests by external actors and proposals for data partnerships.
3. Engage different levels of government, including municipalities, and establish linked-up public data repositories and trusts that can aggregate datasets from multiple agencies, such as those working in health, education, transportation, and other public services.
4. Foster data pools with non-state partners that are topic-specific and integrated, such as for transit planning, by assembling data from sources like telecoms, finance and fuel providers.

D. Governing access to private sector data

1. Mandate private companies operate free data-sharing agreements with public, private and civil society actors, regarding data that provides insights for immediate crises such as pandemics or environmental disasters, law enforcement and prevention of incitement to violence and monitoring systemic risks to rights as related to the business operations of the data holder. In parallel, encourage voluntary efforts for data sharing to knowledge building, public health, children's wellbeing, safety issues and environmental monitoring.
2. Create frameworks for compensated access for non-emergency, public interest uses. Develop systems for tiered compensation of data holders (for example, distinguishing non-profit research and commercial reuse of public-private datasets). Advance codes of conduct where companies agree to allow research access via independent third-party intermediaries and also commit to not taking adversarial action against public-interest researchers.
3. Implement frameworks that promote the use of privacy-enhancing technologies (PETs), such as differential privacy, federated learning, and secure multi-party computation, to enable responsible sharing while minimising risks of re-identification and misuse of sensitive data.

E. Addressing technical issues

1. As appropriate, issue directives on data infrastructure and architecture specifications, including technical access control mechanisms, applicable standards, and regulations.
2. Establish guidance for private sector data to be shared in standardised, machine-readable formats with clear metadata to ensure interoperability. This will facilitate seamless data integration for public interest research, crisis response, and regulatory oversight.
3. Oversee that there are effective cybersecurity frameworks which protect data integrity, as well as prevent unauthorised access, and build public trust in digital systems. Join international mechanisms for threat detection, response and mitigation of potential risks.

4. Promote technical options for data sharing, such as “Trusted Research Environments”. Popularise architectures like well-regulated sandboxes and experimental data repositories.

F. Promoting data quality

1. Encourage awareness and assessment of data in terms of potential bias and exclusions, which may affect the utility of the sharing thereof.
2. Ensure there is appropriate liability for problems regarding data completeness, consistency and reliability, and require public datasets to adhere to FAIR principles (Findable, Accessible, Interoperable, Reusable).
3. Incentivise activities related to dataset cleaning, pre-training and standardisation, in order to address undetected errors and exclusions in datasets.

G. Building data literacy

1. Through a range of interventions, promote awareness and skills around data literacy, especially amongst MSMEs, including digital start-ups, and amongst both researchers and data scientists, as well as regulators. Such measures include financial incentives and training support, as well as reduced-cost access to data sets, computing power and expertise.
2. Invest in education and training programs to enhance data literacy among citizens and civil society, so that they have improved agency and rights to access, manage and utilise data effectively.