Responsible data stewardship

Exploring how responsible data can unlock the potential of sensitive and previously closed data.

ODI Research
ADVANCING TRUST IN DATA

Contents

Introduction

Access to data is critical to tackling some of society's biggest challenges, such as the climate crisis and health inequalities. Once a by-product of industrial, commercial, consumer and other activities, data has become the lifeblood of the way the world works in the 21st century. Nowhere is this clearer than in the development of Artificial Intelligence (AI) – as without data, there is no AI.

Data stewardship, which we define as the 'collection, maintenance and sharing of data', is a crucial part of enabling data to be used effectively. The concept of data stewardship is a response to 'data hoarding' and 'data fearing' scenarios, where data is not shared and thus the benefits of sharing data are not realised. At the ODI, we've explored the topic of data stewardship through our data institutions programme. Data institutions are 'organisations that are stewarding data on behalf of others towards public, educational or charitable aims'.

We've mapped out the landscape of data institutions, considered how they can be financially sustainable, and measured their impact.

Over the last year, the ODI has been researching how organisations can steward data responsibly. We have worked with lots of data institutions that have demonstrated how to steward data effectively, however we have not had a systematic and reliable mechanism to collect evidence that identifies what constitutes responsible data stewardship. We researched what responsibility with data means to different actors through the data ecosystem, and developed a definition of responsible data stewardship to use in our work to help others design and practise responsible data collection, use and sharing. This research involved a thorough literature review of academic, grey literature and practical resources on responsible data, alongside two workshops with a stakeholder group of experts from the field of data governance and 18 interviews with experts and practitioners working with data from across a variety of sectors and geographies.

We define responsible data stewardship as:

An **iterative**, **systemic** process of ensuring that data is collected, used and shared for **public benefit**, **mitigating** the ways that data can produce **harm**, and addressing how it can **redress structural inequalities.**

We understand responsible data stewardship to go beyond compliance with local regulation, and beyond good data practices. It involves a holistic view of data stewardship and its impacts on society and the planet. In our prior research, we focused on understanding responsibility from a theoretical point of view. A key focus for this research is to build out the evidence base of examples that show what responsible data stewardship looks like in practice, via a set of case studies. We hope that these examples will increase the understanding of what constitutes responsible data stewardship as per our definition, and enable learning around how to deliver it.

Each case study sets out information about the organisation or initiative, the data that they steward, how it is stewarded, and the particular practices that we deem to be responsible. The case studies cover a diverse set of organisations, in lots of different shapes and sizes, and stewarding different types of data, across different sectors. They showcase different approaches to stewarding data responsibly across a wide spectrum of sectors, spanning health, humanitarian, labour equality, geostorytelling (using location and mapping to support oral storytelling traditions about places of significant meaning or value) and more.

We have tagged each of the case studies with the different aspects of responsible data stewardship, to demonstrate how each case study connects with our definition:

Iterative: Responsible data stewardship is a negotiated and reflective process. Because contexts vary over time, mitigations and approaches to collecting, maintaining and sharing data need to constantly evolve.

Systemic: The impact of data collection and use is rarely fully within the control of any one organisation, so organisations need to develop a systemic view of their data practices that links how choices made around data have impacts outside the organisation.

Public benefit: Stewarding data responsibly involves ensuring it is used and shared for the benefit of others, rather than only for the benefit of the organisation that holds it.

Reduce harm: Alongside seeking positive impact from the use of data, responsible data stewardship involves identifying and reducing harmful impacts to individuals and communities, which may mean going beyond legal requirements around privacy, security and transparency.

Redressing structural inequality: Data stewardship always occurs within a wider system of relationships, value exchanges and power imbalances, which have real-world consequences for data. Responsible data stewardship may involve meaningful communication with data subjects and other stakeholders, or adopting alternative forms of governance.

Terrastories

Suriname

Terrastories is a free open-source application designed for local communities to map, locate, protect and share oral traditions and stories related to specific geographical places, which would be typically shared offline. Terrastories originated from the mapping oral histories fieldwork conducted by the Amazon Conservation Team (ACT) and a network of volunteer programmers from Ruby for Good.

The app comprises an interactive map that allows users to navigate and click on various locations. Through these clicks, users can delve into the stories and oral traditions connected to each specific point on the map. A data point encompasses various data categories, including geographical details such as information about the region and topographical data, as well as cultural categories like historical and spiritual data.

The app comprises an interactive map that allows users to navigate and click on various locations. Through these clicks, users can delve into the stories and oral traditions connected to each specific point on the map. A data point encompasses various data categories, including geographical details such as information about the region and topographical data, as well as cultural categories like historical and spiritual data.

The data is collected from interviews with communities about their oral traditions, subject matters and geographical data, and the map is curated by users, who can add, edit or remove stories, as well as set differentiated access to them. These mechanisms have the objective to ensure, among other things, that the data is accurate, that the maps have been labelled properly and that the user permissions to view restricted content are well configured. Terrastories is free to install and use, the MIT data licence permits open modification and distribution.

Credit: Photo by Deb Dowd on Unsplash

Practices that demonstrate the responsible stewardship of data

ACT's methodology and the Terrastories platform contribute to reducing harm from insufficient or wrong data being collected, used and shared about Indigenous communities. They work as examples of how technology and data can be used to enable communities to leave a digital footprint about their place-based oral history storytelling traditions, to preserve history and educate future generations about their ancestral landscape.

Terrastories is designed to work entirely offline, so that remote communities can access the application without needing internet connectivity, demonstrating how this tool has directly been tailored to the specific needs of the communities of the Amazon region. However, as a participatory mapping software, it requires a comprehension of social and environmental issues within the context of Indigenous communities. This understanding is crucial for users supporting partner communities to adopt the appropriate approach, utilising inclusive and participatory methodologies.

The platform offers granular control mechanisms over data sharing (private, restricted and publicly viewable settings). Communities can then decide which stories to keep private and which to share publicly with the world. This flexibility seems critical for accommodating diverse needs and preferences. The communities may implicitly license the data based on these access settings.

Redressing structural inequality

Public benefit

Reduce harm

Credit: Photo by Jaume Galofré on Unsplash

Women in Data Women's Health Initiative

United Kingdom

Women in Data (WiD) is a membership organisation with a mission to achieve gender parity in the data, tech and analytics sector. The WiD Women's Health initiative is led by a voluntary team of experts with experience across private, public and third sector data. The WiD initiative aims to demonstrate the power of data to reduce gender inequality in healthcare in the UK by exploring the potential to link a range of clinical and non-clinical datasets to provide insights to women's health and healthcare.

An initial pilot study is linking pharmaceutical and retail transaction data with Office for National Statistics (ONS) data to gain insights relating to fertility, menopause and contraception. The pilot study is combining anonymised pharmaceutical data from Boots, retail 'shopping basket' data from NCR, and ONS population data. For the purpose of the study, the data is being held in the Boots secure research environment and the analysis is being conducted by the NCR data science team, with advisory support from the consumer health company Haleon and the University of Leeds. They are also conducting workshops across diverse community groups to crowdsource and document women's experiences of health services in the UK.

Credit: Photo by Christina @ wocintechchat.com on Unsplash

Practices that demonstrate the responsible stewardship of data

While this initiative is in the early discovery phase, it is considering practices in responsible data stewardship from the start, both for its initial experimental work and its potential future data stewardship model. There is documented gender inequality in healthcare. Men have historically been treated as the default patient in clinical practice and medical research, and this has informed the design of the UK healthcare system, leading to wide gaps in research and treatment for areas unique to women.

Although in its early stages, there are a number of practices that demonstrate the initiative's commitment to the responsible stewardship of data. With the early commitment of anonymised datasets from private sector companies NCR and Boots, as well as the provision of advisory expertise from Haleon, this is a strong example of private sector data being used and shared for public benefit, rather than only for the benefit of the organisation that

And finally, the initiative recognises the potential harms that can come from collecting, using and linking this data. They are developing a data governance and ethics framework structured around the ODI's trustworthy data practices to ensure it upholds high legal compliance and ethical standards with full accountability.

Redressing structural inequality

Public benefit

Systemic

Credit: Photo by Clay Banks on Unsplash

ABALOBI

South Africa

ABALOBI is a social enterprise that strives to use community-led data and technology to foster the growth of sustainable small-scale fishing communities worldwide through inclusion, social entrepreneurship and a knowledge-driven approach to fisheries restoration. ABALOBI provides fishermen with useful tools to collect, visualise, analyse and promote their catches to potential buyers. It has built a complex technological ecosystem that offers a range of services from data collection (ABALOBI Fisher, which allows fishers to log various information about their catches) to data monitoring (ABALOBI Monitor), data visualisation and data sharing.

ABALOBI collects community-generated data on catch details, including methods, timing, locations, weather forecasts, sea conditions and more. Standardised, scalable baseline and longitudinal surveys are used to complement the data, monitoring the progress of users throughout the programme implementation. Prior to data collection, ABALOBI provides capacity-building training, run by an internal community development team, that empowers small-scale fishers as ICT users and data owners. This proactive approach ensures meaningful engagement with the technology and encourages fishers to utilise their data through ongoing training for improved knowledge co-production. This helps ensure the quality of the data collected, used and shared. Ultimately, the mobile apps and website dashboards offer valuable insights for fishers, enabling them to demonstrate the provenance of their catch.

Credit: Photo by Killari Hotaru on Unsplash

Practices that demonstrate the responsible stewardship of data

ABALOBI is a good example of stewardship that tries to address the existing structural inequalities ongoing in data ecosystems. The initiative empowers communities and local fishermen that typically lack access to data to record critical information about their catches. An important part of ABALOBI's strategy is that fishers remain the owners of the information collected on the platform.

The data is never disclosed to external entities without prior consent and on the individual's own terms. Nevertheless, many individuals choose to voluntarily share their data to convey trends and highlight potential areas of concern. For instance, sharing can be granted to local Departments of Agriculture to enable them to monitor fish populations.

The insights of coastal communities derived from this data can be used to shape food behaviours or inform policy actions regarding sustainable fishing practices. In 2015, ABALOBI was officially endorsed as the catch management system for implementing the new Small-scale Fisheries Policy.

Redressing structural inequality

Public benefit

Credit: Photo by Benjamin L. Jones on Unsplash

Sightsavers

Global (UK)

Sightsavers is a global organisation operating in more than 30 countries that focuses on preventing avoidable blindness and advocates for the rights of people with disabilities. It provides eye care services, promotes inclusive data practices, and works with local communities and governments to build sustainable solutions to improve methods of data collection, analysis and use, specifically focused on improving the collection of data about marginalised groups.

An active part of Sightsavers' mission is to provide accurate data on disabilities, which is crucial for understanding and supporting disability in-country.

For Sightsavers, inclusive data means collecting, analysing and using data that is broken down (or 'disaggregated') by disability, sex, location and age. To ensure the high-quality and accuracy of the data it collects, analyses and uses within projects and programmes and to improve service delivery, Sightsavers works with local development partners and community-based organisations from available sources (both official and non-official).

Credit: Sightsavers | David Gnaha

Practices that demonstrate the responsible stewardship of data

Sightsavers sees stewarding data responsibly as a reflective and iterative process. As an example, one objective of its inclusive data strategy is to continue to improve its intersectional approach to the collection, analysis and use of data, including age, sex and disability, to improve programme delivery and outcomes for those at the greatest risk of marginalisation. To advance this agenda, Sightsavers put together an Inclusive Data Charter, alongside learning modules, accountability mechanisms and frameworks such as the Strategy Implementation Monitoring (SIM) card to measure progress.

The organisation has developed resources and guidance to assist data collectors in effectively managing inclusive data collection. For example, Sightsavers has incorporated the Washington Group questions on disability to its internal strategies for collecting and leveraging disability-disaggregated data. Unlike conventional methods, which typically yield a self-reported disability rate of 1%, Washington group questions revealed a more accurate figure of 15%, aligning closely with estimates provided by the United Nations. This has the potential to enhance the depiction of disability in various data collection approaches, including census surveys, thereby offering a clearer and more comprehensive understanding of the prevalence of disability. This data is used by different stakeholders like governments or development organisations to improve their programmes and policies regarding the prevention of avoidable blindness and to fight for the rights of people with disabilities. Access to inclusive data empowers development organisations to pinpoint barriers that marginalise certain groups, enabling them to address these obstacles and enhance the effectiveness and impact of their programmes.

Redressing structural inequality

Public benefit | Iterative

Credit: Kyan | Sightsavers

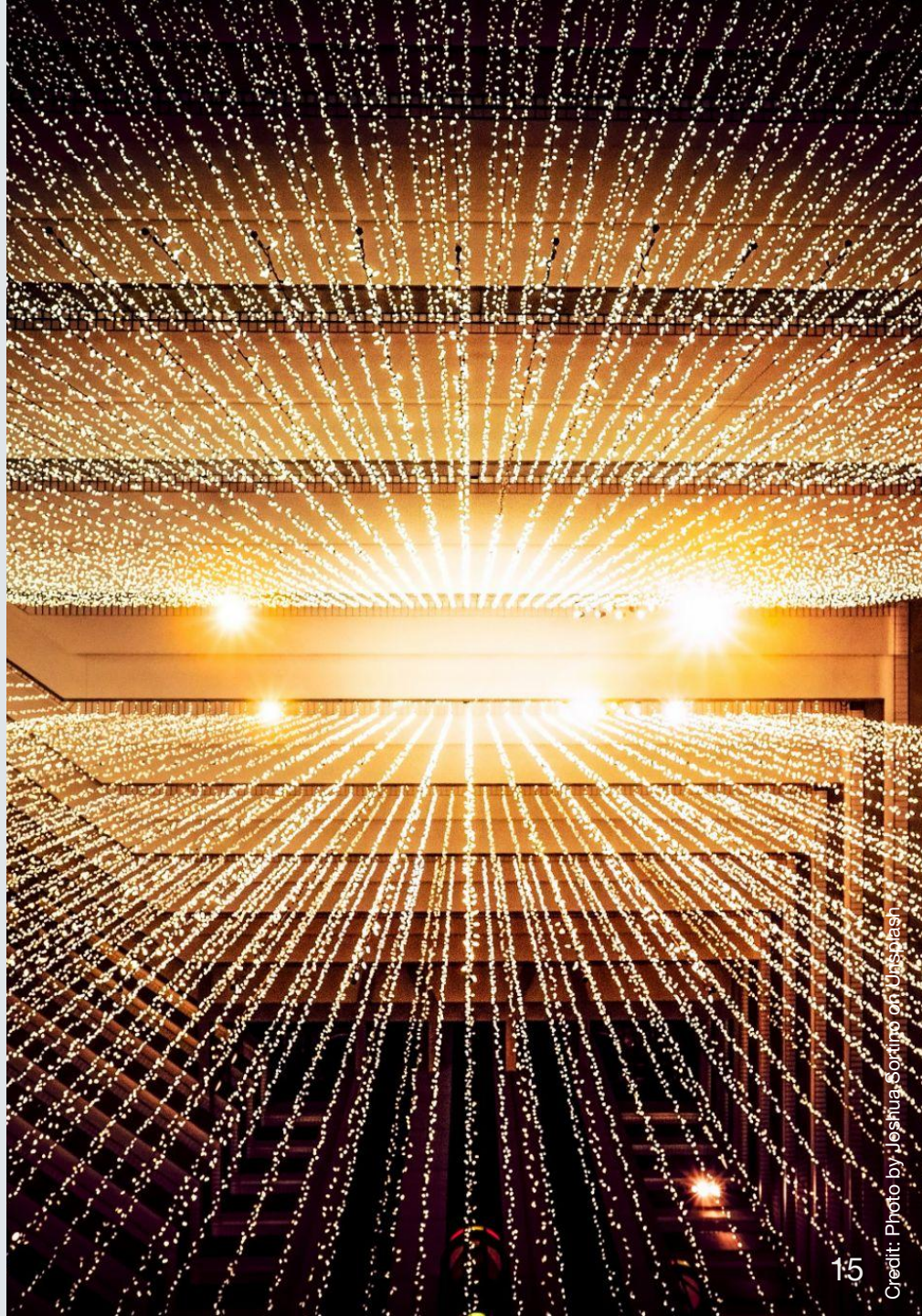
Masakhane

South Africa

Masakhane is a pan-African open source project aiming to use artificial intelligence (AI) to translate the internet into African languages. This initiative aims to build neural machine translation (NMT) systems to bring African languages to the technological forefront, alongside fostering connections among the diverse linguistic communities in Africa. The initiative was developed in collaboration with organisations such as RAIL Lab (University of Witwatersrand, South Africa) and Translators Without Borders.

African languages are 'low-resourced', meaning that there is a lack of quality and publicly available documents that are essential to NMT systems. To address this challenge, Masakhane works with stakeholders to collect their own high-quality 'corpora' (a set of texts that are equivalent, sentence-by-sentence, in multiple languages) from publicly available datasets, including governmental documents, religious texts, literature, news and more.

Masakhane Web also includes a feature that allows users to contribute new data by offering feedback on translations, which is then used to retrain the models for continuous improvement. The project is fully open by default, and the datasets and the translation models are freely available on Github. To improve the quality and availability of datasets that are critical to train NMT, Masakhane has led several projects [here, here or here] aiming at ensuring the availability and quality of datasets that will be used to train NMT models.



Credit: Photo by Joshua Sortino on Unsplash

Practices that demonstrate the responsible stewardship of data

Many of the data, digital tools, and services available worldwide are primarily offered in English or other major Western languages. This situation creates profound structural inequalities concerning access, power and influence within digital ecosystems. Therefore, language emerges as a crucial tool for specific communities to contribute to shaping data ecosystems. This is an important space for many communities and languages around the world: for example, Lelapa.ai and Te Hiku aim to redress the structural linguistic inequalities at stake within digital ecosystems, and Mozilla's 'Common Voice'" enables people to donate their voice to improve the breadth of languages online.

This project also showcases the importance of stewarding data for public benefit. Only 16% of the world speak English, yet the language dominates due to its use in more than half of online content. By providing a translation tool, Masakhane broadens the possibilities online for various communities and languages, especially those largely marginalised by digital ecosystems. The initiative also serves as a starting point for fostering further African technology research: Masakhane's members have produced more than 200 academic papers, including the 2021 Wikimedia Foundation Research Award of the Year, facilitating knowledge transfers and equity within the continent. Masakhane supports these researchers by providing opportunities to pursue academic research in a context where educational enrollment is often limited.

Redressing structural inequality

Public benefit

Reduce harm

Credit: Photo by Reset - Digital for Good

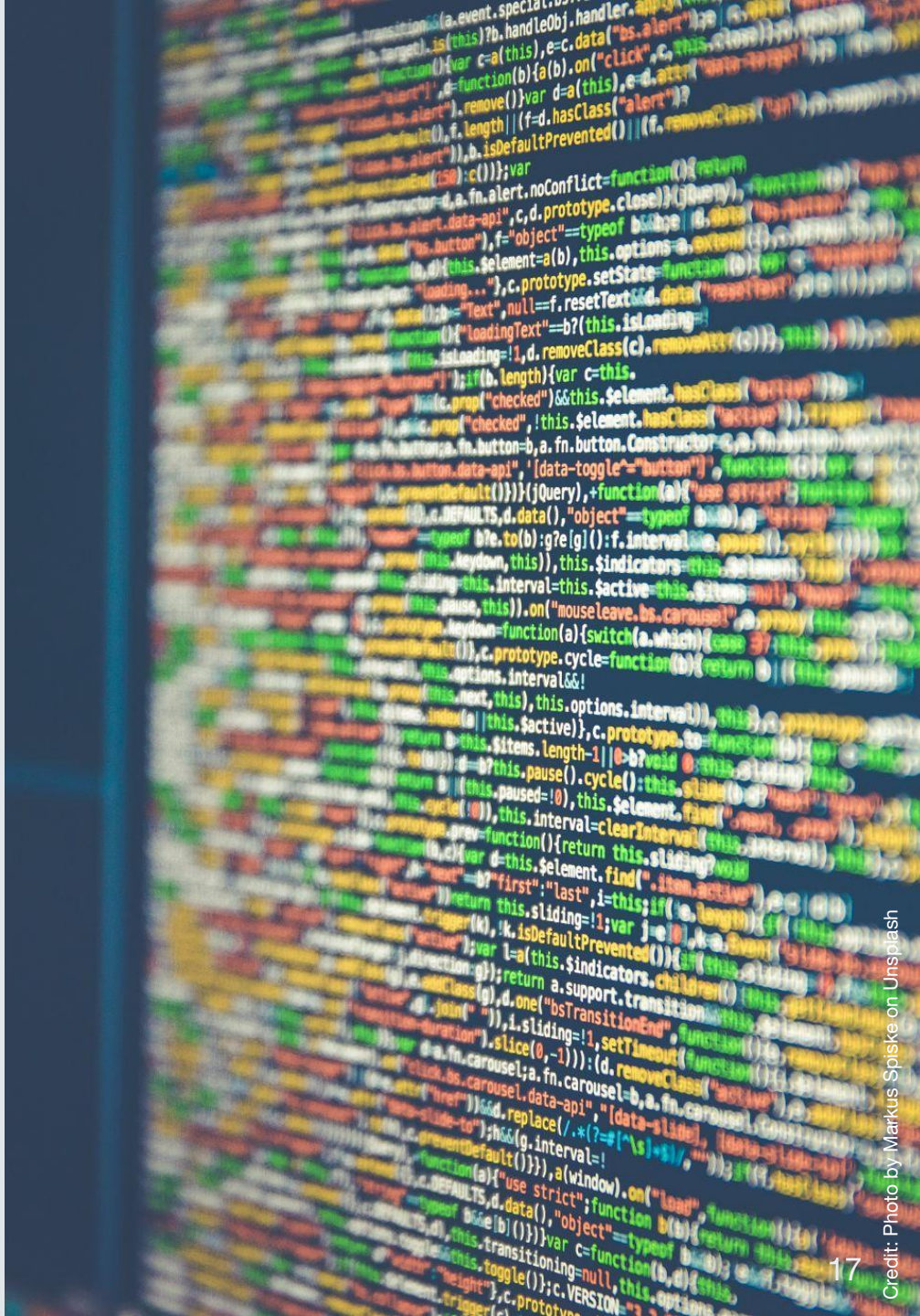
Wikirate

Germany

Wikirate is an open data platform powered by a global community that gathers and organises open data about companies' performance, making sustainability data more usable and enabling meaningful engagements among stakeholders to hold companies accountable. Wikirate's mission is to spur corporations to be transparent and responsive by making data about their social and environmental impacts accessible, comparable, and free. The platform serves as a database for partner research organisations like the Fashion Transparency Index and Beyond Compliance project, including data on supply chain relationships, wages and working conditions.

The data stewarded by Wikirate comes from a wide range of different sources such as civil society organisations, open data platforms, annual reports, news articles, academic research, company reports, brand supplier lists, corporate sustainability reports and more. It includes data related to areas such as working conditions, supply chain relationships, labour violation, facility compliance. The data is licensed under Creative Commons BY Attribution 4.0 International. Evolving reporting standards and varying levels of disclosure can make it hard to maintain up-to-date data. While imperfect in terms of data completeness, Wikirate provides timestamps and calendar year associations for all data, with structured datasets updated regularly by the team.

Users are encouraged to contribute and update data themselves to ensure accuracy. Wikirate ensures data quality through rigorous referencing, data volume and outlier identification, and community-driven verification mechanisms. All data must cite public sources, allowing transparency and review. Community members can flag and correct dubious data, while various verification layers indicate the reliability of information.



Credit: Photo by Markus Spiske on Unsplash

Practices that demonstrate the responsible stewardship of data

The Wikirate platform provides the data for research projects that seek to identify and mitigate harmful impacts to individuals and communities through companies' misuse of their data. For example, the Beyond Compliance living data dashboard assesses modern slavery reporting across business sectors, using 16 standardised metrics to identify shortcomings in corporate reporting and ensuring non-compliance is held to account. The ambition is to drive systemic change by mobilising companies at scale and transforming whole industries, for example, through building an open and integrated data ecosystem for labour rights in supply chains.

It is also iterative: the flexibility of the Wikirate platform allows anyone to set up new research projects that feed into, and scale, collective objectives. It is an ongoing, reflective process that responds to the evolving needs of its data users and contributors.

Its community-driven approach enables Wikirate to encourage continuous dialogue with data contributors and users to gather ideas, share new projects and methods, and get feedback on improvements to the platform.

Systemic

Iterative

Reduce harm

Credit: Photo by Reset - Digital for Good

Karya

India

Karya was established in 2018 as ‘the world's first ethical data company’, with the aim of enabling economic opportunities for Indians living in rural areas. Karya creates, labels and annotates data for machine learning (ML) models, offering both bespoke datasets for clients and datasets for some of the underserved languages of India. The platform also provides work for rural Indians, paying them a minimum of $5 per hour for their contributions - significantly more than many other data labelling services.

Karya stewards speech data and data labels that have been created by their contributors. There are currently 22 datasets listed on the Karya marketplace, including ones in the Kannada and Marathi languages. However, they also steward data for the services provided to clients including Microsoft and Google, which cover high-quality data annotation services for AI/ML models across various media.

Karya has created an app for its contributors, which is fully open source and available to be repurposed anywhere in the world for those who sign up to the Ethical Data Pledge. The platform enables people from rural India to earn money by annotating images and videos and reading out passages of text in native languages via a mobile application to create audio datasets. Where possible, this data is community-owned, and those who have contributed to these datasets will continue to profit from their work as their datasets are used in the future.

Practices that demonstrate the responsible stewardship of data

Two main practices contribute to the responsible stewardship of data. The first is the public benefit created by creating datasets of languages that are not yet annotated nor converted into speech. By translating text, Karya helps technologists to create services for local contexts. This enables small communities that speak languages not usually covered on these platforms to access online services, or create new services to better meet their needs.

Secondly, Karya seeks to be an ethical alternative to more extractive established data labelling services in the way that it collects data and shares the benefits of data stewardship. Karya pays its contributors significantly more than many other services, offering a minimum hourly wage that is more than 20 times the Indian minimum wage and much higher than other data labelling services. Karya prioritises the wellbeing of its contributors, and caps the number of hours per week that they are able to work on the platform. Karya workers, where possible, own 'in principle' the datasets to which they contribute. These datasets are covered by a Public Data License that allows workers to continue to profit from repeated sale of a dataset to which they have contributed over the long term.

Redressing structural inequality

Public benefit

Reduce harm

Magic Box

Global

The United Nations Children's Fund (UNICEF) has a mission to protect the rights of children. Magic Box is a collaborative data sharing platform developed by UNICEF in partnership with private companies such as Telefonica, Google, Amadeus and IBM. UNICEF states: 'By harnessing real-time data generated by the private sector, UNICEF can gain critical insights into the needs of vulnerable populations, which inform decisions about how to invest its resources to respond to disaster, epidemics and other challenges.' For example, in Iraq, UNICEF is collaborating with the government and the leading Kuwaiti internet and network services provider Zain to explore the use of cell phone data and satellite imagery as predictors of child poverty.

Partner companies share their anonymised and aggregated datasets to MagicBox via dedicated APIs. The UNICEF expert team combines this with public data sets, for example relating to climate, location, socioeconomic and epidemiological data, and provides insights in dashboard form. The platform is composed of multiple GitHub repositories and includes real-time and historical mobility, satellite imagery, expenditure and social network data, as well as data about the consumption of content, goods or technology.

Credit: Photo by NASA on Unsplash

Practices that demonstrate the responsible stewardship of data

Magic Box is a strong example of data being used and shared for public benefit. It is a collaborative platform made possible by the contributions of multiple partners that provide their data and expertise for public good. For example, global travel technology provider Amadeus provides real time mobility data that, when combined with UNICEF's public source data, provides insight into the movement of people, which helps to better understand patterns of the spread of diseases and how communities are formed. UNICEF is also using high-resolution satellite imagery from the Magic Box platform in a project to map schools, to help identify gaps and information needs, to serve as evidence when advocating for connectivity and to help governments optimise their education systems.

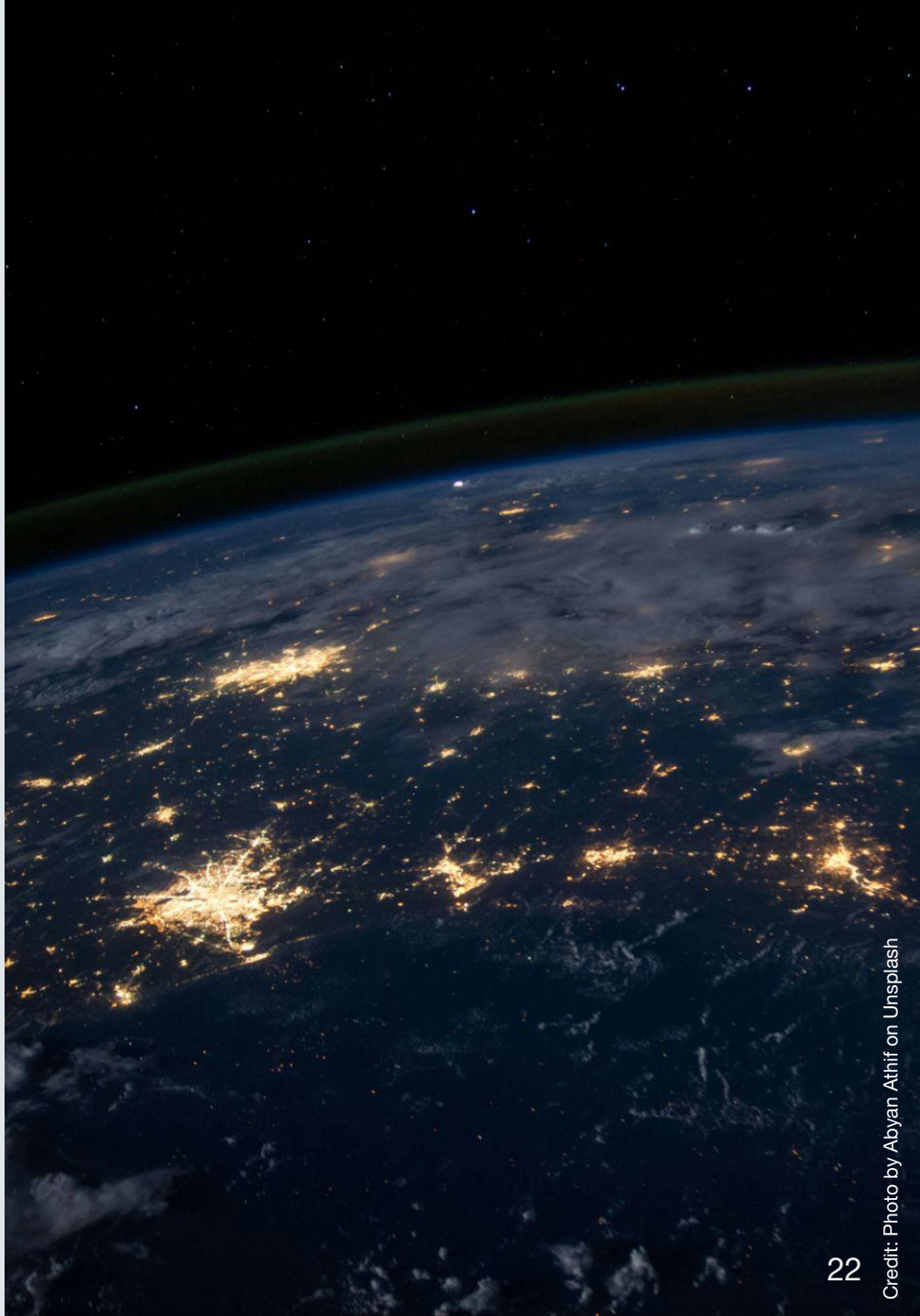
Magic Box also illustrates the iterative nature of responsible data stewardship, both in the development of the technology and the datasets available.

The first version of the platform was developed during the 2014 Ebola crisis in West Africa, and a second was developed in response to the Zika outbreak in 2015. Since then, it has been adapted to multiple applications and made available to open-source collaborators. The use of real-time data to provide insights enables UNICEF to address realities as they change. For example, Magic Box was used to monitor the ongoing effects of physical distancing in 10 countries during the Covid-19 pandemic. It invites the open data community to collaborate to help identify the need for new features, optimising the technical tools involved to accommodate data challenges.

Redressing structural inequality

Public benefit

Reduce harm



Credit: Photo by Abyan Athif on Unsplash

Methodology

This project was conducted from September 2023 to January 2024. It began by creating a long list of examples of responsible data practices. This process involved comprehensive desk research encompassing various sources, including the data institutions register, academic literature, case studies and grey literature focusing on the application of responsible principles to data stewardship. We also collected examples referenced in the prior research project. To find the relevant literature and examples, we conducted queries using keywords related to responsible data stewardship to identify potential responsible stewardship practices, and explored specific organisations resources who work on data stewardship. The long list featured more than 40 organisations from more than 20 countries. The long list was reduced to eight case studies for in-depth exploration.

This process was driven by a set of criteria:

- A balanced geographical representation of organisations, with more than 40% of cases from the Global South.
- Diversity in organisational size and sectoral representation.
- A variety of examples of the five dimensions of responsible data stewardship – iterative, systemic, public benefit, reducing harm and redressing structural inequality. While organisations were not required to meet all of the dimensions simultaneously, their practices needed to address at least two directly.

The assessment process was conducted collectively by members of the project team. The final longlist of eight case studies represents diverse sectors, including workers' rights, environment, sustainability, open data, health, indigenous knowledge and biodiversity.

Limitations

This research focuses on specific responsible data stewardship practices adopted by these organisations, according to the ODI's definition. It is important to clarify that while we acknowledge these responsible ways of stewarding data, it does not necessarily imply that the organisation as a whole is entirely responsible in all of its actions. Desk research relies upon the information freely available online, and where some organisations have limited publicly available information, we may have excluded certain information, and missed specific examples of responsible data stewardship due to a lack of information.

The methodology employed for developing these case studies solely relied on desk research, which inherently carried limitations in fully grasping the nuanced complexities of stewardship practices.

This study was confined to resources and materials available in English, thereby representing only a portion of the examples of responsible data stewardship. We specifically aimed to achieve a balanced representation of organisations from Global North and Global South. Four out of the eight case studies were based in the Global South, including organisations based in Asia, Africa and South America, reflecting our dedication to ensuring a fair representation of diverse geographical and cultural perspectives in our research methods and overall work. Moving forward, we remain committed to enhance this balanced approach to capture a comprehensive and inclusive understanding of the global landscape of data stewardship.

About this research

This slide deck was researched and produced by the Open Data Institute and published in May 2024.

Its lead authors were Sasha Moriniere, Joe Massey and Jane Crowe, supported by Elena Simperl and Josh D'Addario.

If you want to share feedback by email or want to get in touch, contact the ODI Research team at research@theodi.org.

Get in touch

For more information about how the ODI can support your data needs:

- Research
- Consulting services
- Policy
- Membership
- Learning

Open Data Institute
Kings Place
4th Floor
90 York Way
London N1 9AG
United Kingdom

info@theodi.org
+44 (0)20 3598 9395