



Open Data Licensing and Artificial Intelligence: Balancing open access with commons sustainability for AI practitioners

Workshop Report and Analysis of Key Questions plaguing the OS movement

Implemented by:



Introduction

It has been demonstrated that benefits can accrue to everyone when the barriers to learning, using, sharing and accessing technology systems are removed.

These benefits are realised when AI systems, products and tools are licensed in a manner that is open-source or under a similar open access, ethical use licence. Benefits arising from open AI models, tools and systems include models with explicit bias mitigation, scientific collaboration, reproducibility of content, corporate and government transparency, trust from society, collaboration across geographical boundaries, and a competitive, inclusive market that counters vendor lock-in.

The benefits and essential freedoms of openly licenced AI models, tools and systems contribute to the advancement of local AI ecosystems. The current status quo of proprietary systems, blackbox AI and models whose decisions cannot be explained poses major challenges to the environments in which it is developed and consequently deployed in, as well as ensuing hegemony due to asymmetry in power and socio-economic imbalance and inequality. Vulnerable and marginalised sectors are the most adversely impacted in this way.

The question therefore, which formed the foundation for convening this group of experts, was how might we create an informed society that understand the risks and opportunities of open AI data and craft potential solutions for advancing creation and use of open AI training datasets.

Everyone should have equitable access and inclusive participation in the design, deployment, use of, and representation in AI.

Open for Good



Open for Good is focused on improving access to localised training data and AI technologies to spur local open innovation.

Open for Good is an international alliance, focused on improving access to localised training data and AI technologies to spur local open innovation.

One of the goals of the Alliance of African, Asian and other institutions working on Open Artificial Intelligences is to explore practical solutions and share lessons learned on how to implement Openness in AI. The main focus of the alliance is to create inclusive AI commons with localised data. The question of licensing is a key area here, along with other work – to make datasets openly available, facilitate the coordination and exchange of good practice and ideas and increase the public awareness for the benefits of openly available, unbiased and localized training data.

In addition to this workshop, members of the open for good alliance have also worked on additional aspects of "open AI". This includes an opinion editorial on "**Open-Source AI Data Sharing: yes! Data Colonialism: no!**"

Fair Forward

FAIR Forward is a strong advocate for creating open-source products and services (at both local and international levels) that embody a 'democratized' AI future through advancing AI for development, and doing so in a risk-aware manner.

It is therefore against this background that the FAIR Forward project convened this roundtable and workshop whereby participants reflected on certain critical, contemporary questions:

- What practical ways can we sustain open data licences (and avoid the tragedy of the commons) and which licences should be used?
- How might we create a network to solve existing problems in the open data space – including emergent issues such as power (electricity) demands and the managing the high costs of compute power efficiently and effectively through 'commons'?
- Which are most impactful ways to raise awareness about data licences and their application to upcoming AI products and services around the following: (a) existing licences which may not be suitable for AI training data commons; (b) options that exist for current open AI licence uses and (c) what licences have what implications for the user?

FAIR FORWARD

Artificial Intelligence for all.

While it would have been prudent to delve into all the above topics, the roundtable and workshop remained user-driven on priorities and immediate concerns of the practitioners in the ecosystem. The outcome of this event has provided a foundation for further development of the proposed research cases, as well as informed access and use of open AI data and potential solutions for use and creation of open AI training data.

'FAIR Forward – Artificial intelligence for all'

What practical ways can we sustain open data licences (and avoid the tragedy of the commons) in AI systems, and which licences should be used?

How might we create a network to solve existing problems in the open data space of the AI ecosystems in African & Asian contexts – including emergent issues such as power (electricity) demands and the managing the high costs of compute power efficiently and effectively through 'commons'?

The initiative "FAIR Forward – Artificial Intelligence for All" contributes to democratizing artificial intelligence (AI) worldwide. It uses AI to fight poverty, reduce inequalities and achieve a Just Transition in the areas of climate change and agriculture. Globally and together with seven partner countries in Africa and Asia, FAIR Forward implements Germany's AI strategy internationally. The initiative promotes local AI innovation through access to open source AI training data, strengthened AI skills and policy frameworks for responsible AI.



FAIR Forward is one of the initiatives of the global project "Digital Transformation". For more information on FAIR forward **[click here](#)**, On the global project **[click here](#)**.

Contextual Background: Data governance in the African and Asian context



With the rise of the data revolution in Africa and the emergence of technological advancements like machine learning, artificial intelligence, and big data, there has been a growing focus on the importance of data governance and data responsibility¹.

This focus has spurred discussions about suitable models of data governance and avoidance of data extraction for global South / emerging market countries.

To take a step back, there are two important but broad aspects at play:

- 1 Intellectual Property Rights
- 2 Data Rights

Data rights encompass more than mere privacy and ownership of data. These rights enshrine the fundamental freedoms that allow individuals to protect themselves against unwarranted invasions of privacy and excessive control and surveillance by both state and non-state entities, in addition to self-determination of the use of their data.

Copyright licenses expand on these rights – it plays a crucial role to determine the restrictions and permissions associated with data usage, distribution, and access. It defines the terms under which data can be accessed, shared, reused, and/or modified, adding a legal lens and framework for protecting specifically intellectual property rights.

By selecting the appropriate licence for use whether copyleft, copyright, permissive or something in between (example a data use agreement), stakeholders can safeguard their data, ensure proper attribution, and enable the responsible and ethical use and re-use of information.

¹ Bennet, C., and Raab, C.D. (2018) "Revisiting 'The Governance of Privacy': Contemporary Policy Instruments in Global Perspective." Regulation & Governance, Vol 14:3. P 447–464. Wiley.
<https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12222>

Contextual Background: Data governance in the African and Asian context



An alternative framework for data governance is emerging that places the collective interests of communities and their aggregate data at the centre. Indeed, in many African communities, data is considered an extension of a community itself – embodying the language, culture and history of the people³.

A good example, is the concept of **‘Ubuntu’**, an African philosophy and value system that emphasizes the interconnectedness and interdependence of individuals and communities.

Ubuntu promotes the idea that one's humanity is realized and expressed through meaningful co-existence with others. The concept of Ubuntu (humanity, inter-connectedness, community) is not exclusive to African philosophy. In many cultures and contexts across emerging and developing nations this fundamental principle exists to a lesser or different degree but ultimately holding the same core principle – going beyond the self. Also, the CARE (Collective Benefit; Authority to Control; Responsibility; Ethics) Principles for Indigenous Data Governance, set up by the Global Indigenous Data Alliance, and inspired by the UN Declaration on the Rights of Indigenous Peoples reaffirm local communities' rights to self-governance and authority to control their cultural heritage embedded in their data⁴.

It becomes clear then that whereas a lot of African data privacy legislation is modelled after the U.S. federal laws⁵ or European centric laws⁶, an emerging school of thought and experimental frameworks for data governance seeks to offer a different model for data protection in other contexts⁷.

The question of intellectual property and how this ought to be treated for the collective is also another important area of consideration – and this extend to all citizens, globally but most notably the most vulnerable sectors and nations of society.

² Singh, P.J. (2019) Data and Digital Intelligence Commons. Making a Case for their community Ownership ITFC Working Paper 05

³ Milan, S., and van der Velden, L. (2016) The Alternative Epistemologies of Data Activism Digital Culture and Society. Vol. 2, Issue 2

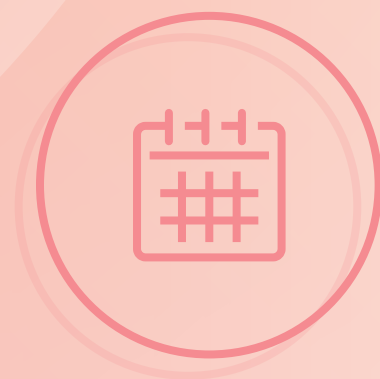
⁴ Global Indigenous Data Alliance (2019) "CARE Principles for Indigenous Data Governance" <https://www.gida-global.org/care>

⁵ California Consumer Privacy Act (CCPA) 2018 or California Privacy Rights Act (CPRA) 2023.

⁶ General Data Protection Regulation 2016/679 (GDPR).

⁷ Abraham, R., vom Brocke, J., Schneider, J. (2019), Data Governance: A conceptual framework, structured review, and research agenda, in: International Journal of Information Management (IJIM), forthcoming.

Workshop Context



2 workshops and experimentation
between December 2021 and May 2022

Open innovation is not without complexities and consequences. Making AI training data openly available for open-source use and development illuminates operational considerations like sustainability of the data pool, use of representative data sets, legalities of combining differently licensed data and use of synthetic data, amongst others concerns.

This workshop on Open Data Licensing aimed to tackle some of the above-mentioned issues with a blended approach. Specifically, the intent was to investigate what practical solutions are available that foster open data use and development of open AI training data, improve access to open AI datasets and highlight the risks and advantages of using open source AI from combined perspectives of practitioners.

The online workshop was structured over 2 sessions, with session one used to critically understand the problem, the ecosystem of who is impacted, what tactical or strategic methods they are deploying presently to bridge the gaps and what opportunities exist to solution for affected persons or institutes.

Session two picked up on the opportunities identified and framed them as 'How might we' (HMW) type questions. Taking inspiration from international and local actors on how similar problems were solved in different sectors, the plenary broke into groups to disseminate and unpack the key HMW question and form hypothesis on their practical solutions and experiments can be conducted to test these. The process and results of the workshop are presented here.

Participants

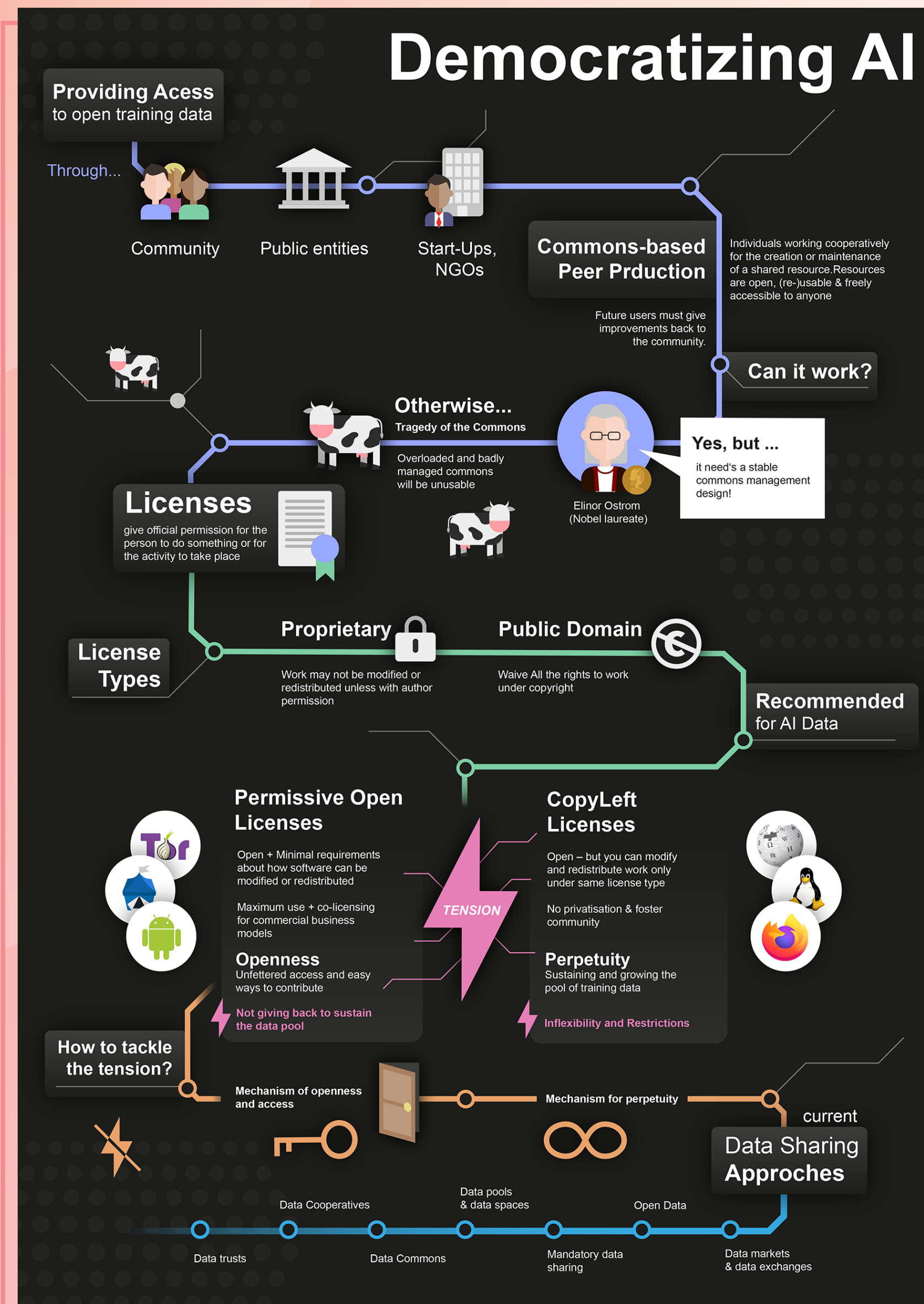
A mix of private sector, civil society, academia and international development representatives.



A Refresher on Open Source



[View the full infographic](#)



Workshop part

1

1 Dec 2021



What do we
want to achieve?



Let's find some
challenges.



Who's affected
by the problem?



How are they solving
their problem today?



What opportunities
can we focus on?

What analysis can we draw from the crazy ideas that people came up with?

Most ideas **centred around creating more defined licence structure for AI data** (improving or tweaking existing licences to be fit for purpose) e.g. A mash-up of copyleft and permissive licences with an ethical lens).

Other Ideas focussed on:

- standardised data formats for open data
- have governments open institutional data for sharing and collaboration
- big tech data commons.

○ Licence or framework needed to **ensure creators, curators and data subjects should receive benefit from the data that is collected and processed. Also use of data should be regulated to ensure data is used for good.**



Key idea:
pertinent to get a licence type created explicitly for AI data.



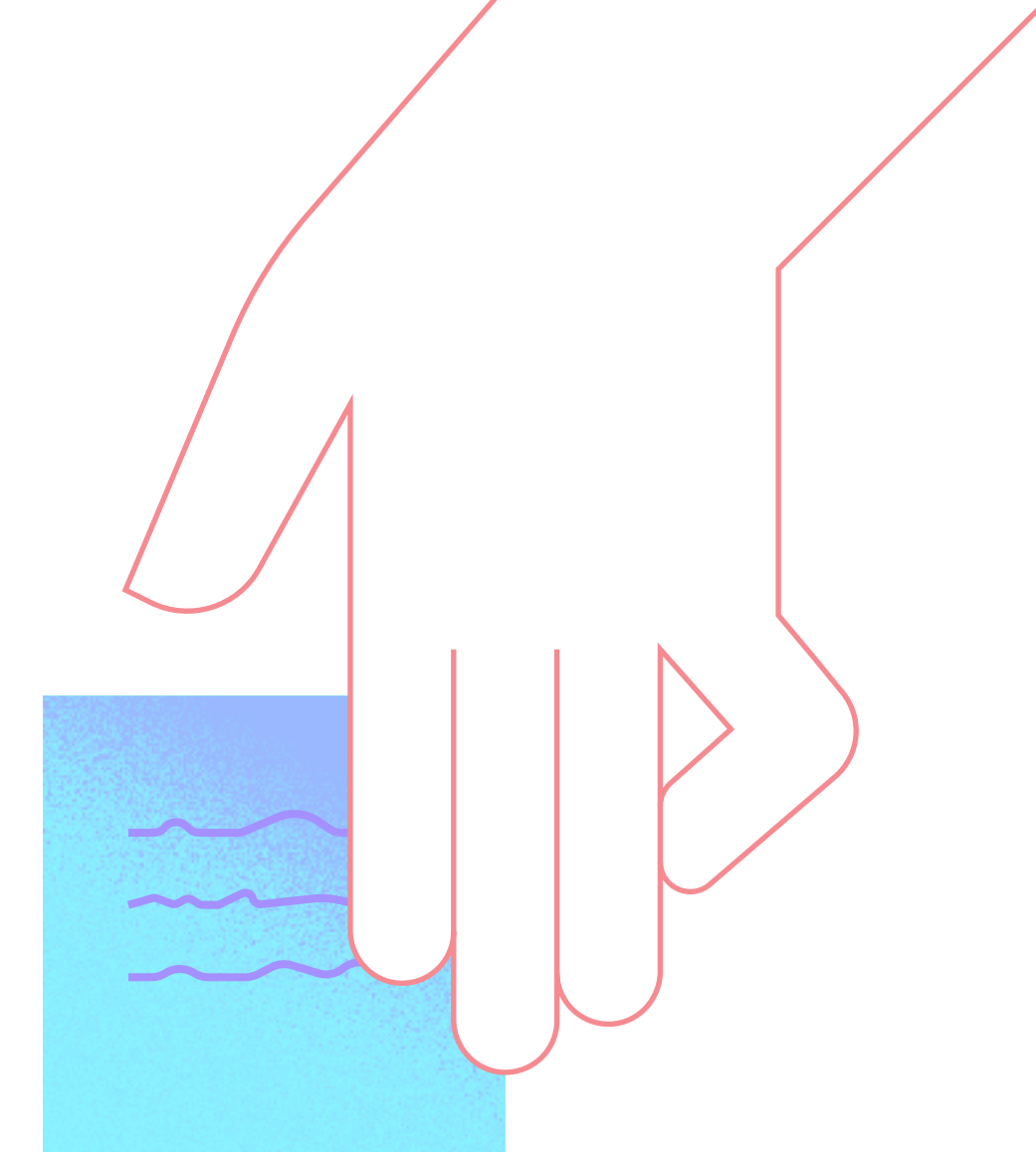
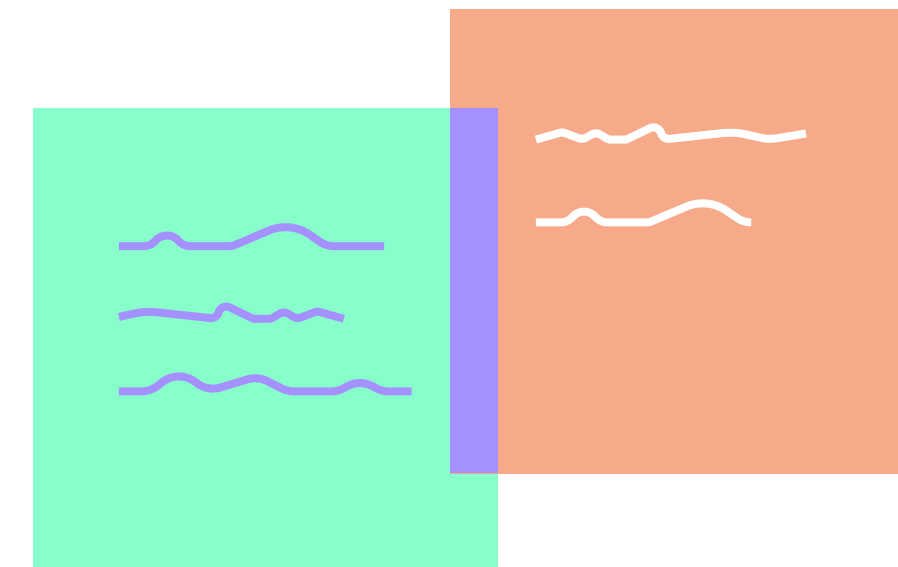
What do we want to achieve?

Key ideas

- A blueprint for an AI-first data license (and provide a justification for why collaborating on this will be important).
- Licence checklist for practitioners who want to open AI training data, to know which type of governance a licence enables, of the underlying AI commons.
- A set of principles or guidelines on how to best approach licensing for AI data.
- Tips, tools and tricks on how best to share training data.
- Developers share datasets - often using licensing - but that has the potential to propagate biases and harms. Can licenses be revised to share liability?
- To demystify (different types/options) licencing for devs working with AI training data they find out there.
- In licensing discussions, the importance of addressing rights pertaining to redistribution and reuse could also be as important as promoting data openness. Balancing accessibility with protection of creators' rights ensures responsible and ethical data use while fostering transparency and accessibility.



Let's find some challenges



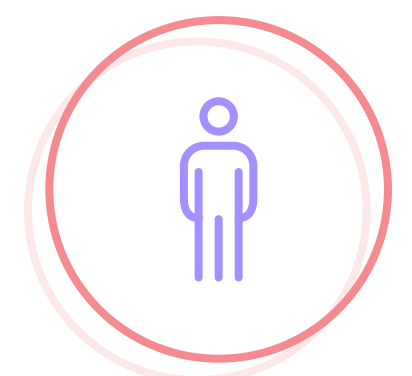
Current problems with Open Data and sharing of data for training open AI?

- Lack of communication between relevant fields (technical, legal and regulators).
- Inability to sustain useful training data sets from multiple sectors (who manages the commons?).
- Lack of clarity / readily available information on licensing regimes for Open AI.
- Obtaining high value data and monetising it without returning an equivalent value to the data originators and data curators.
- Private sector actors not collaborative to share anonymized data that could be used for social good.
- Existing licensing frameworks are not formulated with an ethical lens.



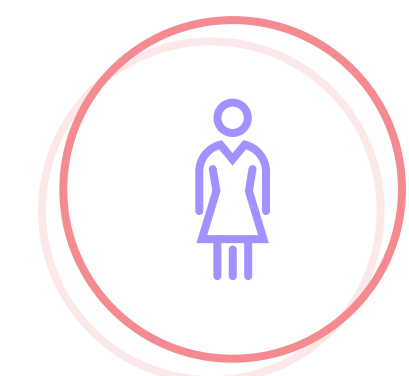
Who's affected by the problem

Who are the people that are experiencing pains or gains? What are their issues?



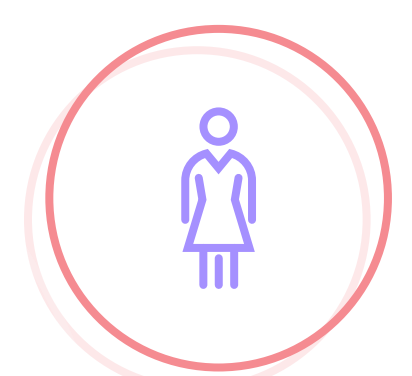
Founder:

This data would help me build a cool product: who can I ask permission to use it?



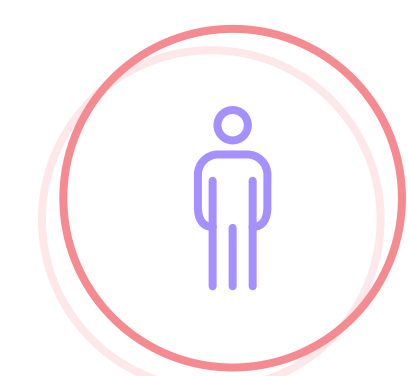
Researcher:

I only want the data I create or where a specific group is represented in to be used for a purpose, how can I specify and limit those?



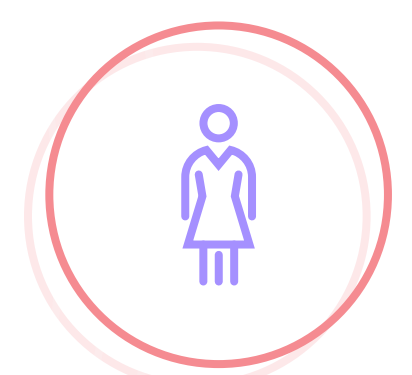
Government / Regulator:

if we open institutional data, it may lead to scrutiny of government practices and possible action from public interest groups.



Developer:

if I build a commercial application from this data, how can I share the benefits with the data creators?



Innovator:

Will using this data could lead to litigation in the future?



What opportunities can we focus on?

After analysis of the people, roles and general issues facing Open AI training data and licensing, the group found the following key opportunities areas to work on.

Framed as **How Might We** questions.

- Creation of guidelines that are human-rights centric, for not just for more training data for AI purposes, but also ethics, equity, cultural and some of the political contexts.
- Mapping the current ODL practices and impact thereof if possible.
- Develop legal frameworks (for civil suits) against biased AI.

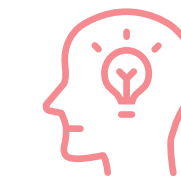
How Might We...

- ... ensure our data is being used as we intended? create equity not only in ensuring more training data for AI purposes, but also ethical guidelines that are considerate of cultural and political contexts?
- ... make data accessible and understandable to non-technical stakeholders?
- ... collaborate to find new ways of sharing AI training data for everyone's benefit?

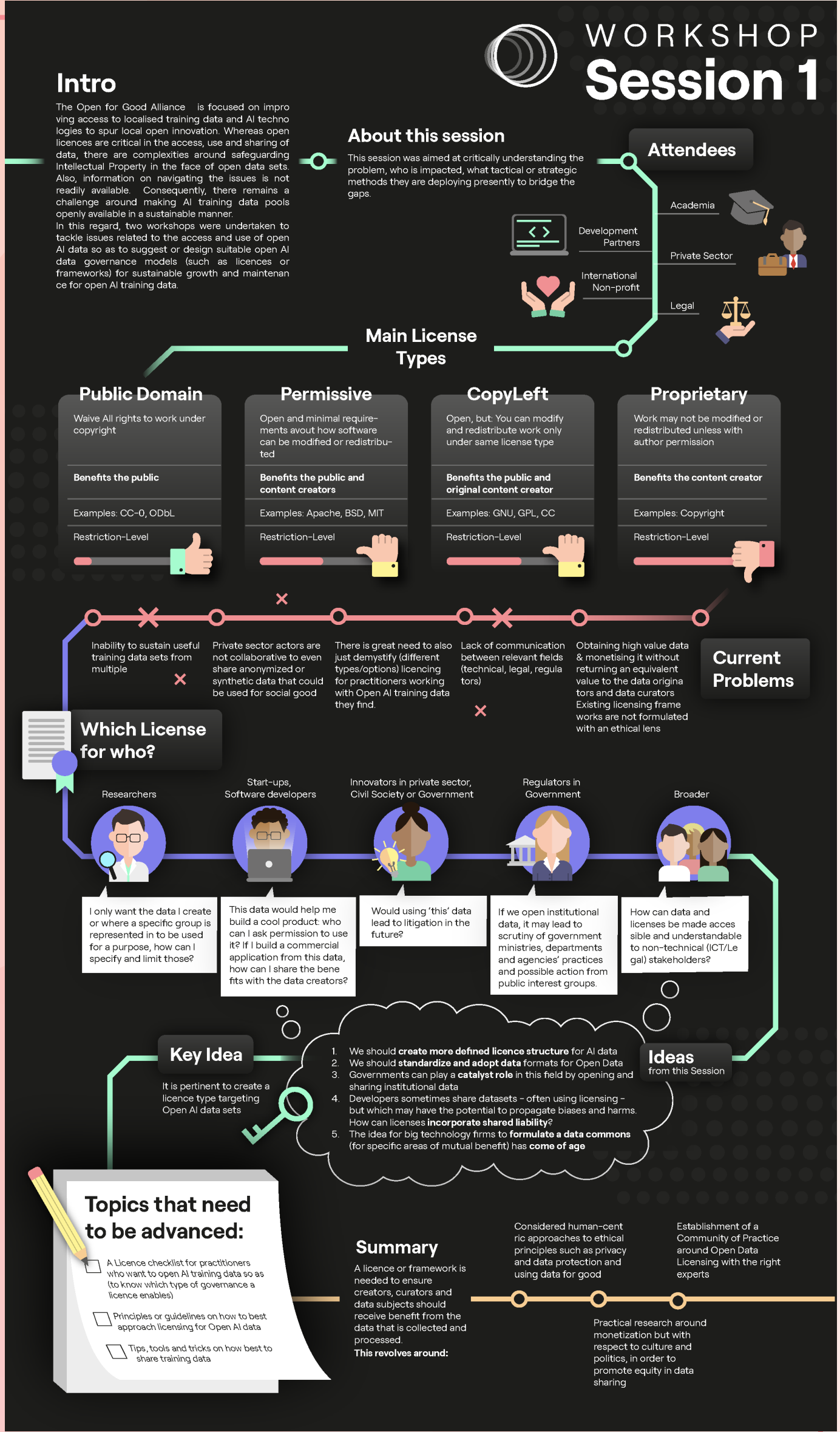
Session 1 feedback

At the end of day 1, with a critical analysis and in-depth discussion of the issues plaguing Open AI & Open Data Licences, the group agreed that a longer discussion is needed on how licensing frameworks can consider the ethical principles and social impacts of AI.

E.g. privacy concerns, concerns around monetization, and concerns around equity and data sharing!



Thoughts?
Establishment
of a Community
of Practice around
ODL with the
right experts.



Workshop part Round up



[View the full infographic](#)

Workshop part

2

26 Jan 2022



Flashback/Context setting by

Deshni Govender – Country
Focal Point: South Africa
& **Mark Irura** – Country Focal
Point: Kenya (FAIR Forward)



HMW Prioritizing



Inspiration!



Ideation



Hypothesis-building



Target audience



Critical assumptions



Experiment design



Flashback / Context setting



At start of session 2, the group re-looked at the HMW questions from the previous session and added some new ones.

Most participants had a strong interest in the following (ordered by priority).



How might we use ODL to:

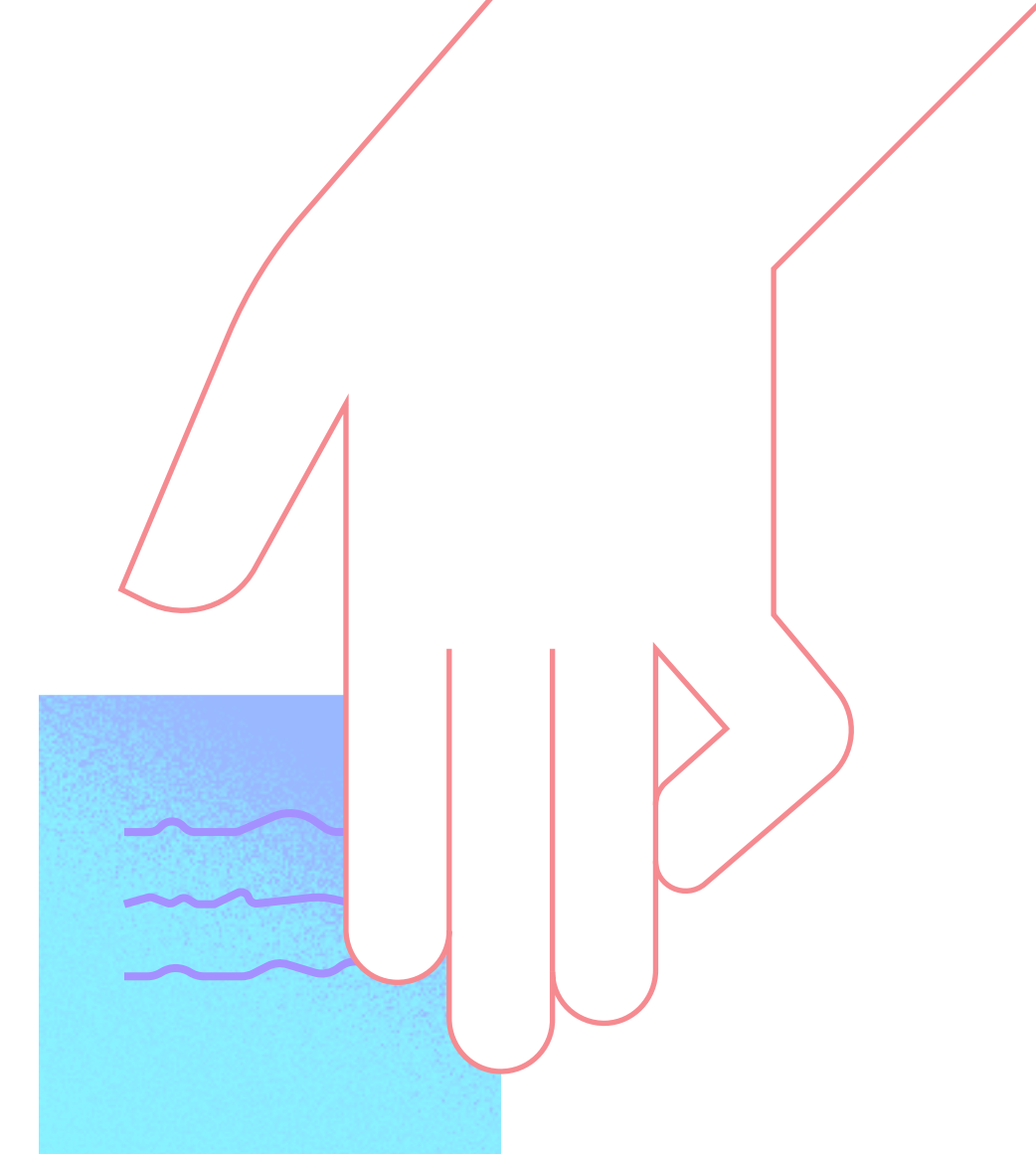
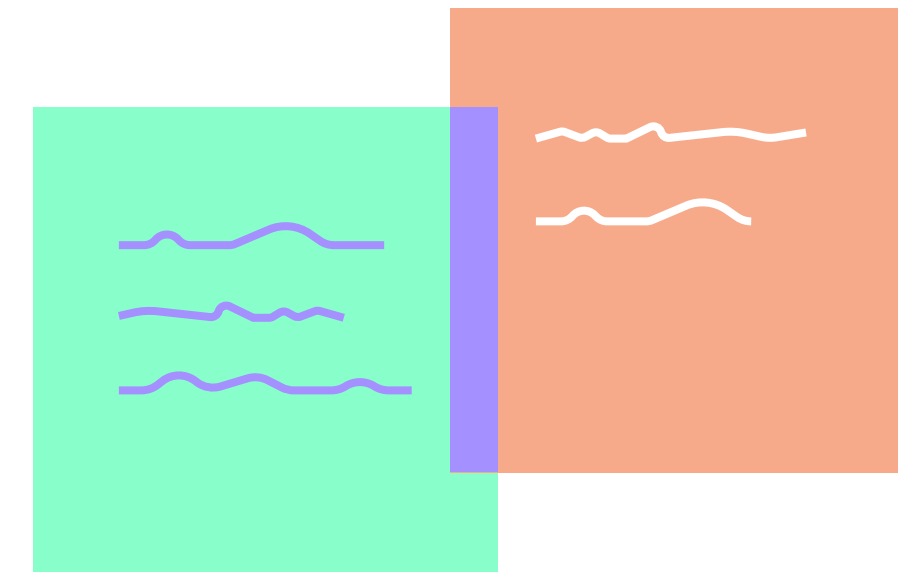


- ☐ ... ensure ethical, equitable and context-sensitive uses of open AI training data for public benefit?
- ☐ ... know the use of open AI training data is in line with its creators' intention?
- ☐ ... promote sustainable data sharing practices that are beneficial to all parties involved?



Inspiration!

Group shared ideas / concepts
from similar industry in which
the same problem was faced?



**TE MANA
RARAUNGA**
Māori Data Sovereignty Network



**Open
Conversational AI
Community**

**GLOBAL
.hEALTH** | a Data
Science
Initiative

Disclaimer: GLZ and Open for Good are not affiliated, associated, authorized, endorsed by, or in any way endorse or are officially connected with Te Mana Raraunga, Open Conversational AI Community or Global Health. These entities, as well as related names, marks, emblems and images are registered trademarks of their respective owners. trademarks of their respective holders – and their names, logos and/or images are used solely for illustration purposes in connection with workshop participant discussions.



First level ideation



The group then agreed on the key question that they would take into smaller group discussions:

HMW use ODL to ensure that open data has a fully positive impact on equity, avoiding issues such as helicopter research or extractive practices?

Put differently: what solutions, processes or safeguards can be created when using open data to ensure there is parity and benefit for contributors, while resourced entities or individuals do not colonise data or benefit from open data without sustaining the data pool.

How can ODL be leveraged to ensure sustainability as well as benefits to data creators, data stewards and the data subjects especially at the local level?



Groups



Group 1

Franziska Heine

Executive Director of Wikimedia Germany

Yasin Jernite

Researcher at Hugging Face



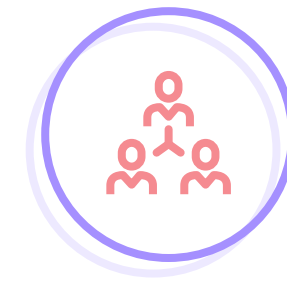
Group 4

Joshua Meyer

Co-founder of Coqui

Tobias Schonwetter

Associate Professor, University of Cape Town (South Africa)



Group 2

Alex Diaz

Google.com

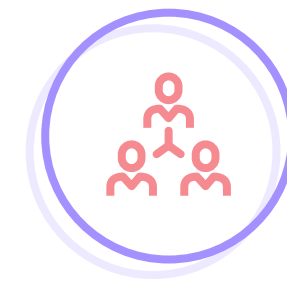
Deshni Govender

Country Focal Point: South Africa

FAIR Forward: Artificial Intelligence for All

Ruth Schmidt

Global Focal Point (FAIR Forward)



Group 5

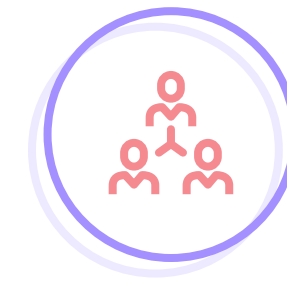
Mark Irura

Country Focal Point: Kenya

FAIR Forward: Artificial Intelligence for All

Herkulaas Combrink

Co-director Centre for Digital Futures, University of Free State (South Africa)



Group 3

Daniel Brumund

Global Focal Point (FAIR Forward)

Stephen Moore

Senior Lecturer at Kwame Nkrumah University of Science and Technology (Ghana)





Innovative ideas

Visualizing data licenses based on context of creation and use.

A tool that helps anyone to know what ODL to use depending/across jurisdictions.

Data Sharing frameworks for various actors (researchers, start-ups, govt).

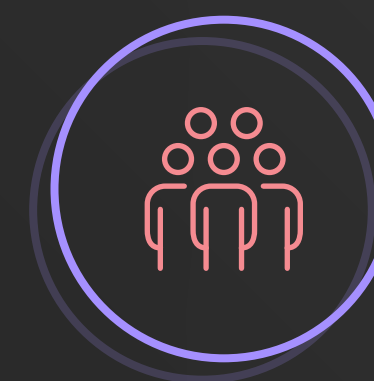


Behaviour change ideas

Data creators, data stewards and data subjects should benefit from being recognised and from knowing how their data is being used.

Currently data is not viewed as an asset in all contexts. We want to educate civil society, governments, Industry, and academic institutions on data as an asset.

Changing attitudes across sectors will require lobbying, training and a buy-in.



Quick wins

Promote community-centric Licensing.

Outline why CC licenses can create problems and how this can be remedied.

Group 2 analysis

Alex, Deshni and Ruth

Hypothesis: API access is of crucial importance when accessing data

We believe that a Data Sharing framework for sectors of researchers, start-ups and govt will facilitate the access and usability of data. By providing an API (application plug-in interface), data can be assessed in a controlled manner from commercial entities wishing to share data for public good – this may partly solve the issue of data not being accessible (see FAIR guidelines).

Potential pain points identified: researchers will not benefit from this type of data sharing as oftentimes not even API existent or data are not findable or are not being made available open access. So, the step here needs to start sooner: make data always accessible and usable for others via an open access platform such as NCBI, or provide clear documentation of where data are hosted and how to use them (see API).

The main users of APIs will therefore be smaller entities, private sectors, civil society organizations or even international development agencies wishing to create open data sets.

Thoughts / Questions:

Access barriers to personal data and info protected under GDPR, POPI, etc. (Google specific or other big tech companies in terms of accessing) – e.g. in health or financial sectors

Are there any **policies** around data access, e.g. COVID-19 and mobility (location) data are extremely hard to obtain, particularly from private sector or big tech and thus this becomes arduous and challenging to aggregate insights relevant for public health consumers – this would need to be assessed on a case by case.

At times, where private sector may be willing to share **washed data** once key elements are stripped (e.g. geo-location) then the data may not be valuable much or at all to third party users. It can however be helpful that a specified data set is washed for access only to a sub-set of data needed for research and/or to train a model.

Discrete problem statement is the most crucial point when requesting access to private sector data e.g. flood forecasting for a specific region, during rainy season. Categories of data needed would be relevant for private sector to easier determine what and how much can be shared.

What's benefit in API access? Scale, uptake and credit to entity. Data creators and data stewards (even in private sector / big tech) benefit from being recognised and from knowing how their data is being used. It builds their confidence, access opportunities for e.g. grants for future research. Conscious effort to improve and contribute to further data analysis or insights.

When working with different organisation and on global issues, e.g. vulnerability mapping, data and base line models should be shared for organisations that work in these areas, as such an API / federated learning mode would work to ensure benefit to all parties.

Group 3 experiment analysis

Daniel and Stephen

Resulting hypothesis:

We believe data creators and stewards – particularly if they are smaller organisations – would benefit from better recognition for their work (e.g. through citation) and from more transparency about the usage of the data (e.g. through contact information / reporting by data users).

It may help them with mobilising further data contributions, strengthening their credibility and ability to access funding as well as create accountability and transparency of usage.

Results from testing hypothesis:

How can locally beneficial uses of relevant datasets (e.g. local language data) be ensured and promoted?

Need for fitting license:

Need for an (adjusted) ODL to help ensure the use of the dataset is in line with its creators' intention. For instance, an ODL that ensures services, products or research on open data is also made open.

Need for skills development:

Creating open datasets needs to be complemented with skills development to ensure it can be used meaningfully by local AI ecosystems.

ODL Playbook:

How about creating a playbook that outlines suitable ODL arrangements in line with expected outcomes (e.g. "If you want products built on your open data to be open, then use license XY"; "If you want users of your open data to notify you about the product/service they built, then use license XY and ask for e-mail contact").

Group 5

Herkulaas and Mark

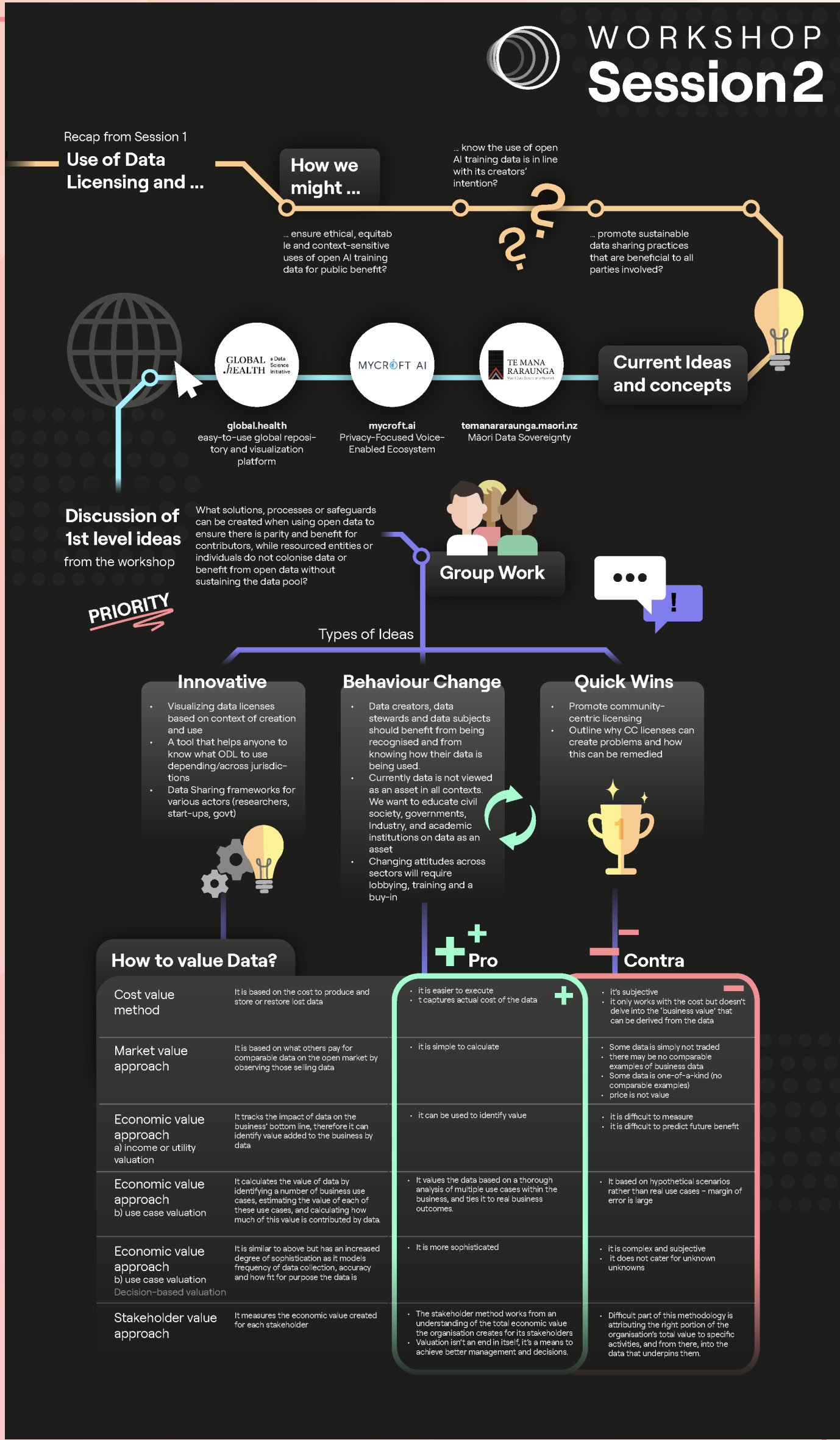
Hypothesis:

That the more organizations provide attention to data, the more valuable data assets in the data portfolio become (in creating value to these very organizations). The table below summarizes ideas on how to value data ¹.

Method	Description	Pros	Cons
1. Cost value method	It is based on the cost to produce and store or restore lost data.	<ul style="list-style-type: none">- It is easier to execute.- It captures actual cost of the data.	<ul style="list-style-type: none">- It's subjective.- It only works with the cost but doesn't delve into the 'business value' that can be derived from the data... ie. Price is not value.
2. market value approach	It is based on what others pay for comparable data on the open market by observing those selling data.	It is simple to calculate.	<ul style="list-style-type: none">- Some data is simply not traded – there may be no comparable examples of business data – either because others are not interested at this time, or because a company is keeping its data to preserve a competitive edge.- Some data is one-of-a-kind, so there will be no comparable examples.- Price is not value.

¹ <https://www.anmut.co.uk/an-introduction-to-data-valuation/>

Method	Description	Pros	Cons
3. Economic value approach Income or utility valuation.	It tracks the impact of data on the business' bottom line, therefore it can identify value added to the business by data.	It can be used to identify value.	<ul style="list-style-type: none"> - It is difficult to measure. - It is difficult to predict future benefit of data.
3. Economic value approach Use case valuation: Business model maturity index (Internet of Water).	It calculates the value of data by identifying a number of business use cases, estimating the value of each of these use cases, and calculating how much of this value is contributed by data.	It values the data based on a thorough analysis of multiple use cases within the business, and ties it to real business outcomes.	It based on hypothetical scenarios rather than real use cases – margin of error is large.
3. Economic value approach Use case valuation: Decision-based valuation.	It is similar to above but has an increased degree of sophistication as it models frequency of data collection, accuracy and how fit for purpose the data is.	It is more sophisticated.	<ul style="list-style-type: none"> - It is complex and subjective. - It does not cater for unknown unknowns.
4. Stakeholder value approach	It measures the economic value created for each stakeholder.	<ul style="list-style-type: none"> - The stakeholder method works from an understanding of the total economic value the organisation creates for its stakeholders. - Valuation isn't an end in itself, it's a means to achieve better management and decisions. 	Difficult part of this methodology is attributing the right portion of the organisation's total value to specific activities, and from there, into the data that underpins them.



Workshop part

Round up



[View the full infographic](#)

Thank you.



Deshni Govender

Country Focal Point: South Africa

FAIR Forward: Artificial Intelligence for All

+27 72 747 1951

deshni.govender@giz.de



Mark Irura

Country Focal Point: Kenya

FAIR Forward: Artificial Intelligence for All

+254 723 230593

mark.irura@giz.de



Follow us: FAIR Forward (@fair_forward)



Info on FAIR Forward: [BMZ Digital Global – FAIR Forward](#)



Learn more about Artificial Intelligence and how you can get involved [Atingi Online](#)

