Introduction to critical data studies

A handbook for educators

Authored by Javiera Atenas with an introduction by Caroline Kuhn

How to cite this handbook:

Atenas, J. (2021). Introduction to critical data studies. A Handbook for educators. Editors C. Kühn, J. Atenas & L. Havemann. Understanding Data: Praxis and politics. HDI - Data, Praxis and Politics. http://doi.org/10.5281/zenodo.4912463

A funded project by the **EPSRC** EP/R045178/1

Introduction to the context of data and farming

by Caroline Kuhn

Information and Communication Technologies (ICT) are being used across the world to generate efficiency gains for farmers. This has led to an information and data explosion with an associated boom in new applications, tools, actors, business models, and entire industries. Agri-food systems are being transformed.

Beyond the technological developments, data for and from farmers has become a growth area, driving expectations and investments in big data (but also small data), blockchain technology, precision agriculture, farmer profiling and e-extension. Investing in data-driven agriculture is expected to increase agricultural production and productivity, help adapt to or mitigate the effects of climate change, bring about more economic and efficient use of natural resources, reduce risk and improve resilience in farming, and make agri-food market chains much more efficient. Ultimately, it will contribute to worldwide food and nutrition security.

Smallholders in particular have much to gain from data – small improvements in their operations are likely to provide larger gains at household level, proportionally, and, if the improvements are widely adopted, the whole agricultural sector in many countries that depend on smallholder agri-food systems can be transformed.

However, for smallholders to benefit from data-driven agriculture, tools and applications need to be designed for their specific situations and capacities; they – and the organizations that support them – need to grow their capacities to **become smart data users and managers**; measures are needed to ensure that farmer-generated data is not exploited or misused; and smallholders, usually the least powerful parts of a value chain, must grasp every opportunity to be included in the collective data flows within agri-food systems.

Based on data there are questions that can be answered much quicker. Such question can be:

- Where does our food come from?

A funded project by the **EPSRC** EP/R045178/1

- Can we manage risks in our farm and take control measures against droughts or pests?
- Are we able to predict problems such as floods or low yields?
- Can we make informed decisions on what to grow, what treatment to apply, when to plant, treat or harvest?

Technologies today allow us to build services to answer these questions but data only offers these opportunities when it is usable. For this reason we are going to devote the first unit to understanding what is data and what does it mean ‘to be usable’, namely to open.

In what follows we are going to present the different units that compose this booklet. We start with a glossary of terms that can be edited by you, adding new terms that you learn while you undergo this learning experience.

In order to understand the state of the art in the world of smart-farming the core text to read is

> Digital and Data-Driven Agriculture: Harnessing the Power of Data for Smallholders (2018). Ajit Maru, Dan Berne, Jeremy De Beer, Peter Ballantyne, Valeria Pesce, Stephen Kalyesubula, Nicolene Fourie, Chris Addison, Anneliza Collett, Juanita Chaves. Published 01 MAY 2018, available from: https://doi.org/10.7490/f1000research.1115402.1, the material is CC, BY, SA

A funded project by the **EPSRC** EP/R045178/1

🅭 🅯 🄏

• Learning outcomes •

1. Understanding the basics of open data
2. Understanding the key principles of data
3. Understanding the basics of Open Science
4. Understanding data ethics
5. Understanding the concepts of data agency and sovereignty
6. Developing ideas to innovate using open data

• Introductory multimedia [video - podcast] •

The big wins of Open Data for agriculture and nutrition. (10 min)
https://youtu.be/o43P8SO82qU, here is a shorter version
https://youtu.be/IZKSFX4Dwb8 (2 min)

Open data is data that anyone can access, use and share. It sounds simple, but not everyone is aware of the open data available to them. In this video, we explain what open data are and explain some of the benefits these can bring to the key players in agriculture and nutrition: from farmers to researchers, and from government officials to journalists.

A funded project by the **EPSRC** EP/R045178/1

Table of content

A funded project by the EPSRC EP/R045178/1

• Glossary of terms • (wikipedia when possible)

Agency is the means to manage “our” data and access to it, agency enables us to act effectively in these systems, as and when we see fit.

Artificial intelligence (AI) is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals.

Data: is a symbolic representation that describes facts, conditions, values or situations

Information: is an organised set of processed and related data in a way that allows us to communicate or acquire knowledge

Dataset: is a collection of organised data records where each element has the same structure, ordered for processing by a computer

Database: is an information management system that aims to be a single point of reference for those who want to search for and access data. It is made up of a management system for datasets and their metadata, which provides users with tools to speed up the publication, access, search and navigation of data.

Data ecosystem: as the community of actors, stakeholders,and entities who engage with data, the data assets (datasets, data products, platforms, tools, technologies) with which they interact,and the rules, norms, and structures that govern those interactions (policies, cultures, organizational structures, etc.). (can be find in p.3 of the publication

Data ethics: refers to systemising, defending, and recommending concepts of right and wrong conduct in relation to data, in particular, personal data.

Data sovereignty: is the idea that data must be subject to the laws and governance structures within the nation in which it is collected. The concept of data sovereignty is closely linked with data security, cloud computing and technological sovereignty. Also, it can be understood as the relation between data and groups of vulnerable or

A funded project by the **EPSRC** EP/R045178/1

minority groups, which must have agency and voice-over how their data is collected, shared and portrayed.

Data protection: is the relationship between the collection and dissemination of data, technology, the public expectation of privacy, and the legal and political issues surrounding them. It is also known as data privacy..

Data agency: is the individual's ability to influence and shape his/her life trajectory as determined by his/her cultural and social contexts. Agency in the digital arena enables an individual to make informed decisions, where his/her own terms and conditions can be recognised and acknowledged at an algorithmic level. This not only includes the ability to opt-in or opt-out of data collection and processing but also the broader ability to engage with data collection, storage and use, and to understand and modify data and the inferences drawn from it.

Data infrastructure: Data infrastructure consists of data assets supported by people, processes and technology. (Available at the ODI website)

Co-creation: those processes or activities where at least two actors (for example, public, private, governmental or civic) collaborate in the realisation of a project to achieve a certain result

GDPR: The General Data Protection Regulation is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA). It also addresses the transfer of personal data outside the EU and EEA areas. The GDPR's primary aim is to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU.

Hackathon: a marathon of ideas, design and software prototyping, which brings together specialists, technical programmers, designers and entrepreneurs to work collaboratively. The days tend to last between 24 and 48 hours and usually have specific challenges to guide the event

A funded project by the **EPSRC** EP/R045178/1

Interoperability: denotes the ability of diverse systems and organizations to work together (inter-operate). In this case, it is the ability to interoperate - or intermix - different datasets. It allows for different components to work together (Open Knowledge Foundation)

Legibility is concerned with making data and analytic algorithms both transparent and comprehensible to users. (Encyclopedia of Human Data Interactions).

Negotiability is the means to navigate the social aspects of data, which supports interaction between other data subjects and their policies. This enables the ongoing engagement of users so that they can withdraw from data processing either completely or in part and can derive value from data harvesting for themselves. (Encyclopedia of Human Data Interactions).

Open data: are data that can be freely used, reused and redistributed by anyone, subject only, at most, to attribution and distribution requirements with the same licence

Open Database License: is the authorisation to use the data issued by the source that owns the copyright of the data

Open Innovation: offers functionalities on which independent companies, developers or innovators can build complementary products, services or technologies

Personal data is any information relating to an identifiable person.

A funded project by the EPSRC EP/R045178/1

• Recommended reading •

1. Johnson, J. A. (2014) ‘From open data to information justice’, Ethics and Information Technology, vol. 16, no. 4, pp. 263–274. doi: 10.1007/s10676-014-9351-8
2. Baack, S. (2015) ‘Datafication and empowerment: how the open data movement re-articulates notions of democracy, participation, and journalism’, Big Data and Society, vol. 2, no. 2, pp. 1–11. doi: 10.1177/2053951715594634
3. Arzberger, P., et al., (2004) ‘Promoting access to public research data for scientific, economic, and social development’, Data Science Journal, vol. 3 (November), pp. 135–152. doi: 10.2481/dsj.3.135
4. Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P. L., Görögh, E., Helbig, K., ... & Ross-Hellauer, T. (2018). The open science training handbook https://www.fosteropenscience.eu/content/open-science-training-handbook
5. Davies, T., Walker, S. B., Rubinstein, M., & Perini, F. (2019). The state of open data: Histories and horizons (p. 592). African Minds. https://stateofopendata.od4d.net/. In particular the section of Agriculture is relevant for our pilot group in Nairobi, link here (https://stateofopendata.od4d.net/chapters/sectors/agriculture.html).
6. Gurin, J., Bonina, C., & Verhulst, S. (2019) Open Data Stakeholders - Private Sector. In T. Davies, S. Walker, M. Rubinstein, & F. Perini (Eds.), The State of Open Data: Histories and Horizons. Cape Town and Ottawa: African Minds and International Development Research Centre. Print version DOI: 10.5281/zenodo.2677777

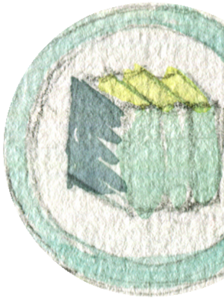
• Complementary resources •

1. Open Standards for Data https://standards.theodi.org/#:~:text=Open%20standards%20for%20data%20are,adopt%20open%20standards%20for%20data.

2. Tennant, J. (2020). A value proposition for Open Science. https://osf.io/preprints/socarxiv/k9qhv/

3. Train-the-trainer card game for Open Science training | FOSTER https://www.fosteropenscience.eu/content/train-trainer-card-game-open-science-training [teaching resource].

4. Open Data Innovation Week tools https://labs.webfoundation.org/projects-2/open-data-innovation-week-2/
5. Nyeleny declaration of the Forum for Food Sovereignty (2007)



6. Disruptive technologies in agricultural value chains. Insights of East Africa. Available from here.

   https://www.odi.org/sites/odi.org.uk/files/resource-documents/disruptive_agritech_-_5_mar_2020_-_final_draft.pdf

7. What does data-driven farming mean? https://blog.heatspring.com/what-does-data-driven-farming-mean/

8. Producing, using, innovating: How 50x2030 is closing the agricultural data gap. The 50x2030 Initiative to Close the Agricultural Data Gap aims to empower and support fifty low and lower middle income countries (L/LMICs) to build strong national data systems that produce and use high-quality, timely agricultural survey data. Link here.

1 • Introduction to open data •

By Javiera Atenas with contributions from Juan Ignacio Belbis and Juan Pane

> We need to look at the whole society and think, "Are we actually thinking about what we're doing as we go forward, and are we preserving the really important values that we have in society? Are we keeping it democratic, and open, and so on?"
>
> *Tim Berners Lee (xxxx)*

• Introduction •

Data are characteristics or information, usually numerical, that are collected through observation. In a more technical sense, they are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum is a single value of a single variable. Data are transformed into information when they are created, extracted, elaborated and used with pre-established objectives. The information system, often made up of data of the same or different type (the data set is defined as a “dataset”), is transformed into knowledge when it is interpreted thanks to tools, applications, methods, indicators, etc.

Data can be small or big, private, personal, governmental, military, scientific, public, confidential, commercial, financial or open, and normally pertain to information delivered in machine-readable file formats (machine-readable) in a format known as raw data. The most common formats of data are integer, floating-point number, character, string and Boolean.

With the constant evolution of technology, the informative content and the data held by public administrations represent excellent opportunities to promote transparency in the actions of governments and administrations. Moreover, they can offer more efficient services and, since they facilitate reuse by other public and private subjects, they also can be used in areas other than those for which they have been produced or collected. Knowledge, in practice, acquires the value of awareness - in the case of open

A funded project by the EPSRC EP/R045178/1

data these can be defined as "collective", understood as being for the "common good" - when used for change and the improvement of reality (the facts).

Whilst data are features of information that are collected through observation, information is understood as a symbolic representation that describes facts, conditions, values or situations, collected and arranged in an appropriate way to fulfil the objective of the institution that manages it. On their own, these values lack a semantic value, that is, they do not have a meaning for someone, so they do not add value to the recipient of the message. For these data to make sense, they must be processed, associated or grouped within the same context to form information. Thus, we can conclude that information is an organised set of processed and related data in a way that allows us to communicate or acquire knowledge.

An important point to keep in mind is that there are **different formats of data.** For example, if the data is tabular, that is, it is contained in a table, one of the most used formats is CSV (comma separated value. is CSV is in a plain text **format** information with is a series of values contained in a sheet and is separated by commas.). On the other hand, if the data indicates geo-referencing there are other specialised formats to represent this information. Below are some of the most commonly used data types and formats.

Kind of data	Description	Common formats
Generic Data	Data that does not have specialised applications, and that normally corresponds to data from databases and reports, such as spreadsheets and information tables.	The characteristics of the formats to use are: 1. The use of standard, open (non-proprietary) formats should be considered. In certain justified cases (historical data, expensive transformations) proprietary formats can be used. 2. For traditional database and spreadsheet data, consider using CSV and JSON formats.
Images	Data that comes	1. You should use image file types based

A funded project by the **EPSRC** EP/R045178/1

🅭🅯🄏

	essentially from photographs.	on the JPEG method and PNG files. 2. For vector and web display formats that require rendering, SVG (scalable vector graphics) is recommended.
Statistical data	Data normally used by specialised users of the statistical area. They are in particular formats for reuse and exploitation of specialised statistical applications.	As a data format, the Statistical Data and Metadata Exchange (SDMX) (OECD standard) is recommended. Acceptable for its widespread use are SPSS (PSPP) or STATA. In the latter case, you should also consider publishing a distribution of the data in CSV format.
Georeferenced data	Geographic data designed to capture, store, manipulate, analyse and display geographically referenced information in all its forms.	Among the most common formats we can find are: ➔ .kml, ➔ .geojson ➔ .topojson.

1.1 • Understanding Open Data •

According to the International Open Data Charter, “*Open Data is digital data that is made available with the technical and legal characteristics necessary so that it can be freely used, reused and redistributed by anyone, at any time and anywhere.*" The Charter has arisen from a conversation between governments and civil society, which has resulted in the promotion of the adoption of the six principles described below. Moreover, Open Data has been defined by the Open Knowledge Foundation as that which can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share-alike. Open Data core technical principles can be understood as follows.

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and Redistribution:** the data must be provided under terms that permit reuse and redistribution, including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, reuse and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

1.2 • Principles and value of Open Data •

The six principles of open data developed by the Open Data Charter are a globally-agreed set of aspirational norms for how to publish data, which can be summarised as follows.

1. Open by Default: This represents a real shift in how government operates and how it interacts with citizens. At the moment, we often have to ask officials for the specific information we want. Open by default turns this on its head under the perspective that there should be a presumption of publication for all. Governments need to justify data that is kept closed, for example, for security or data protection reasons. To make this work, citizens must also feel confident that open data will not compromise their right to privacy.

2. Timely and Comprehensive: Open data is only valuable if it is still relevant. Getting information published quickly and in a comprehensive way is central to its potential for success. As much as possible, governments should provide data in its original, unmodified form.

3. Accessible and Usable: Ensuring that data is machine-readable and facilitates its dissemination, with portals being one way of achieving this. It is also important to consider the user experience of those accessing data, including such matters as the file formats in which information is provided. Data should be free of charge and under an open licence, as demonstrated by Creative Commons.

4. Comparable and Interoperable: Data has a multiplier effect. The more quality datasets you have access to, and the easier it is for them to talk to each other and hence, the more the potential value that can be acquired from them. Commonly agreed data standards play a crucial role in making this happen.

5. CImproved Governance and Citizen Engagement: Open data has the capacity to let citizens (and others in government) have a better idea of what officials and politicians are doing. Transparency can improve public services and help hold governments to account.

6. For Inclusive Development and Innovation: Finally, open data can help spur on inclusive economic development. For example, greater access to data can make farming more efficient, or it can be used to tackle climate change. Finally, we often think of open data as just about improving government performance, but there is a whole universe out there of entrepreneurs making money off the back of open data.

A funded project by the **EPSRC** EP/R045178/1

The Government of Canada summarises the Benefits of Open Data for the society as follows.

> Support for innovation - Access to knowledge resources in the form of data supports innovation in the private sector by reducing duplication and promoting the reuse of existing resources.
>
> Advancing the government's accountability and democratic reform – Increased access to government data and information provides the public with greater insight into government activities, service delivery, and use of tax dollars.
>
> Leveraging public sector information to develop consumer and commercial products - Open and unrestricted access to scientific data for public interest purposes, particularly statistical, scientific, geographical, and environmental information, maximises its use and value, whilst the reuse of existing data in commercial applications improves time-to-market for businesses.
>
> Better use of existing investment in broadband and community information infrastructure - Canada has invested in information and communications networks in the form of technical infrastructure and community services, such as libraries and social service agencies.
>
> Support for research - Access to federal research data supports evidence-based primary research in the Canadian and international academic, public sector, and industry-based research communities. Access to collections of data, reports, publications, and artifacts held in federal institutions allows for the use of these collections by researchers.
>
> Support informed decisions for consumers - Providing access to public sector service information to support informed decision-making, for example, real-time air travel statistics, can help travellers to choose an airline and understand the factors that can lead to flight delays.
>
> Proactive Disclosure – proactively providing data that is relevant to Canadians reduces the amount of access to information requests, email campaigns and

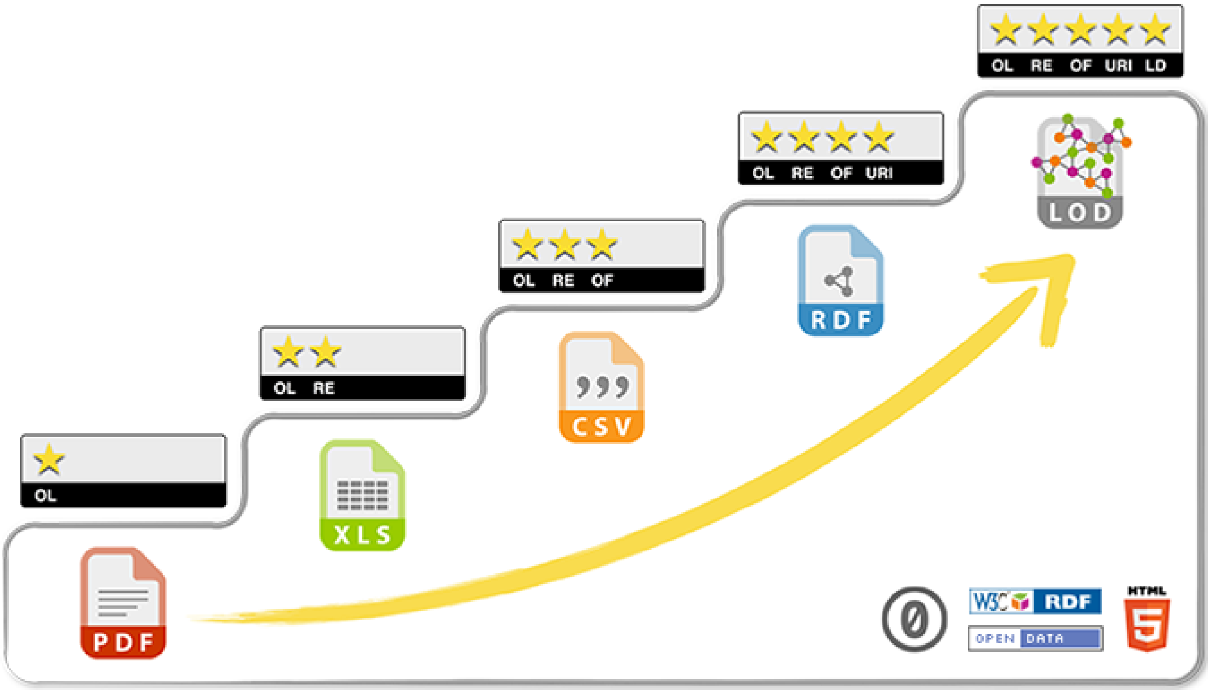
A funded project by the **EPSRC** EP/R045178/1

media inquiries. This greatly reduces the administrative cost and burden associated with responding to such inquiries.

1.3 • Quality of open data •

The technical approach to data opening is based on the five-star data opening scheme defined by Tim Berners-Lee, a summary of which can be seen in the five star figure. This scheme proposes an incremental scale of data openness levels, where each level implies progress in terms of the objectives of open data: freedom of use, reuse and redistribution.

5-star Open Data XX



Five-star data opening scheme: Tim Berners-Lee (2012)

The great leap to the third star: the third star implies that the data is in a non-proprietary format, that is, it can be consumed and reused by anyone. To this end, the open data organisations are championing the standardisation of the open formats

A funded project by the **EPSRC** EP/R045178/1

to be used in order to facilitate the work of data consumers. These formats are summarised in the following table.

Format	Description	Proposed Standard Scheme
CSV	Tabular data format where columns are delimited with a comma, although other separators such as semicolons are commonly accepted. Whilst it is not yet standardised, there are efforts to define good practices, such as RFC 4810.	JSON Table Schema
JSON	Data exchange format based on the key-value schema inspired by the Javascript object model. The main difference from the CSV format is the ability to define nested structures.	JSON Schema

An important point to keep in mind is that, depending on the types of data to be published, there are different formats to be used. For example, if the data is tabular, that is, it is contained in a table, one of the most used formats is CSV. On the other hand, if the data indicates geo-referencing there are other specialised formats to represent this information. Below are some of the most commonly used data types and formats.

A funded project by the **EPSRC** EP/R045178/1

2 • Introduction to open science •

By Javiera Atenas

Open Science is just science done right
Jon Tennant (2018)

• Introduction •

Open science is the movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society: amateur or professional. Open science pertains to transparent and accessible knowledge that is shared and developed through collaborative networks. It encompasses practices, such as publishing open research, campaigning for open access, encouraging scientists to practise open-notebook science, and generally making it easier to publish and communicate scientific knowledge.

According to FOSTER, Open Science represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools. For the EU Open Science, *it represents a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools. The idea captures a systemic change to the way science and research have been carried out for the last fifty years: shifting from the standard practices of publishing research results in scientific publications towards sharing and using all available knowledge at an earlier stage in the research process.*

The OECD defines Open Science as being *to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction.* For UNESCO, *the idea behind Open Science is to allow scientific information, data and outputs to be more widely accessible (Open Access) and more reliably harnessed (Open Data) with the active engagement of all the stakeholders (Open to Society).*

A funded project by the **EPSRC** EP/R045178/1

By encouraging science to be more connected to societal needs and by promoting equal opportunities for all (scientists, policy-makers and citizens), Open Science can be a true game changer in bridging the science, technology and innovation gaps between and within countries and fulfilling the human right to science.

In the Open Science MOOC, *Open research data refers to the publishing of the data underpinning scientific research results so that they have no restrictions on their access. Openly sharing data opens it up to inspection and re-use, forms the basis for research verification and reproducibility, and opens up a path to broader collaboration.*

2.1 • Open science principles •

In this section, materials from the Open Science Training Handbook and the Open Science MOOC are presented.

According to FOSTER, Open Science is about increased transparency, re-use, participation, cooperation, accountability and reproducibility for research. It is aimed at improving the quality and reliability of research through the principles of inclusion, fairness, equity, and sharing. Open Science can be viewed as research simply done properly, extending across the life and physical sciences, engineering, mathematics, social sciences, and humanities (Open Science MOOC). In practice, Open Science includes changes to the way science is done - including opening access to research publications, data sharing, open notebooks, transparency in research evaluation, ensuring the reproducibility of research (where possible), transparency in research methods, open source coding, software and infrastructure, citizen science and open educational resources.

One of the key elements of Open Science is reproducibility, which pertains to research data and codes being made available to others, who are able to obtain the same results as ascertained in scientific outputs. Closely related is the concept of replicability; the act of repeating a scientific methodology to reach similar conclusions. These concepts are core elements of empirical research.

A funded project by the **EPSRC** EP/R045178/1

Improving reproducibility leads to increased rigour and quality of scientific outputs and thus, to greater trust in science. There has been a growing need and willingness to expose research workflows from initiation of a project and data collection right through to the interpretation and reporting of results. These developments have come with their own sets of challenges, including designing integrated research workflows that can be adopted by collaborators, while maintaining high standards of integrity.

The concept of reproducibility means being able to do or apply a method again in another piece of research. It is directly applied to the scientific method by providing clear and open documentation, thus making the study transparent and reproducible.

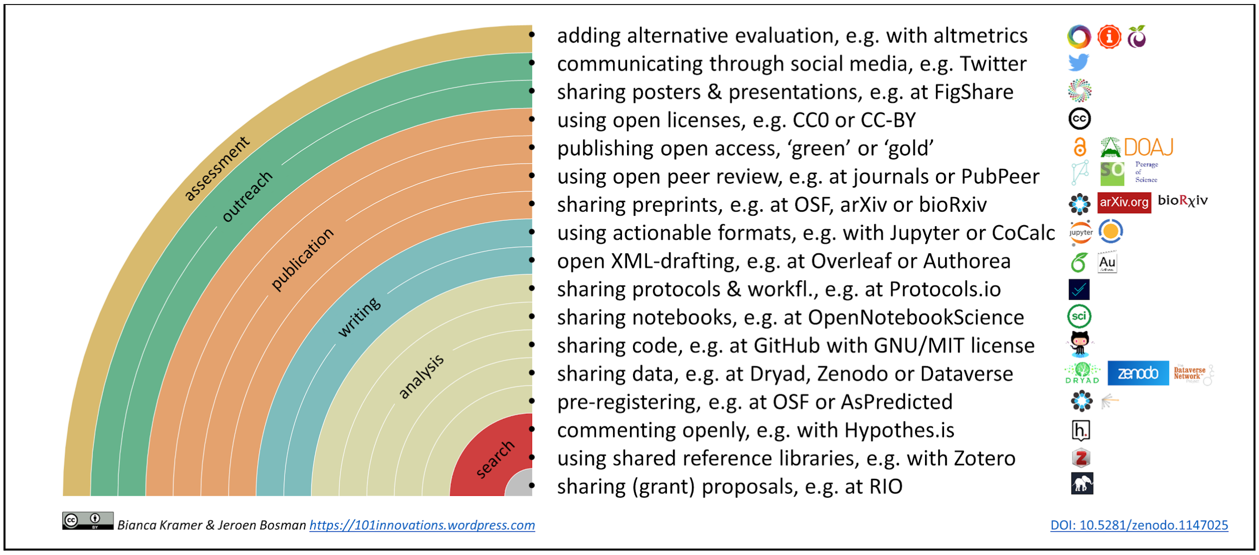
Goodman, Fanelli, & Ioannidis (2016) note that, in epidemiology, computational biology, economics, and clinical trials, reproducibility is often defined as *the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.*

This is distinct from replicability, *which refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected*. A simpler way of thinking about this might be that reproducibility is methods-oriented, whereas replicability is results-oriented.

Reproducibility can be assessed at several different levels: an individual project (e.g. a paper, an experiment, a method or a dataset), an individual researcher, a lab or research group, an institution, or even a research field. Slightly different kinds of criteria and points of assessment might apply to these different levels. For example, an institution upholds reproducibility practices, if it institutes policies that reward researchers who conduct reproducible research. Further, a research field might be considered to have a higher level of reproducibility, if it develops community-maintained resources that promote and enable reproducible research practices, such as data repositories, or common data-sharing standards

A funded project by the **EPSRC** EP/R045178/1

You can make your workflow more open by...



- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, 'green' or 'gold'
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO

Bianca Kramer & Jeroen Bosman https://101innovations.wordpress.com DOI: 10.5281/zenodo.1147025

It is key to critically enabling open science practices in teaching and learning because it fosters research-driven learning opportunities through open data, which is a condition *sine qua non* for reproducibility and scientific progress, facilitating reuse. For Ioannidis and Khoury (2011), *Opening up data enables to detect false claims and inaccuracies and allows for replicability tests.* In sum, opening up research data can have a wide societal impact.

A funded project by the **EPSRC** EP/R045178/1

3 • Understanding data ethics •

By Javiera Atenas

Data ethics refers to a series of principles or guidelines to which any data research-led project or activity must adhere, with the main focus being on human rights and personal data protection laws.

Data ethics are the principles governing what is right and wrong in the data cycle, from collection and production, to its use. In this unit, the different facets of data ethics and data protection concerning commercial, educational and public data are explored, starting with the premise that not all data, public or private, is publishable, and that not all uses are harmless. The discussion will centre on the different debates around data.

We need to consider that data is framed by regulations, which serve people, governments, organisations and industries to control and balance the potential uses of the data so that they can benefit the society without harming people. In the context of human-generated data, we will discuss two kinds to understand how they can and should be published as well as how to protect people, including vulnerable communities, from pervasive and intrusive uses of data.

We live in a 'datafied' society, where almost everything is continuously transcribed into data, quantified and analysed (Van Es and Schäfer, 2017), where decisions taken by corporations and governments are increasingly data- and algorithm-driven. Data have an impact that ranges from the economy to education and policy, to what we watch and connect with. It can be said that data permeate almost every single element of modern life and therefore, it is crucial to understand risks of their present and future uses. In addition, understanding the ethical conundrums we face when dealing with data will inform how data could be used in the future and how it can be interweaved to create new datasets that can be used to predict all kinds of behaviours and try to influence them (Hand, 2018).

A funded project by the **EPSRC** EP/R045178/1

The emergence of new technologies, the main aim of which is to process data to gain knowledge about human activities, is generating social asymmetries between those who own the tools and have the expertise to collect and analyse data and those whose data are subject to these applications (Belbis & Fumega, 2019). Artificial intelligence (AI) and other practices that are designed to exploit large volumes of data, which emerge as a product of the digitisation of the vast majority of information services, create the need to discuss ethical limits.

Data ethics can be understood as the responsible and sustainable use of data. It is key that we learn how to collect, select, analyse and use such data under the premise of 'do no harm', thus ensuring that data-led research projects are beneficial for people and society. Data ethics need to be understood as a social contract between the public and data users (Buenadicha et al., 2019 -article in Spanish). Data ethics refers to a series of principles or guidelines to which any data research-led project or activity must adhere, with the main focus being on human rights and personal data protection laws. Thus Data ethics principles must lead to actively design fair and unbiased research and motivate students to learn, from the very beginning, the value of data protection and data agency by raising awareness of the role of an ethical common ground when conducting research with data, by treating others' data as you wish your own is treated.

A funded project by the **EPSRC** EP/R045178/1

4 • Understanding personal data and individual agency •

By Javiera Atenas

> *Humans cannot respond on an individual basis to every algorithm tracking their behaviour without technological tools supported by policy allowing them to do so.*
>
> IEEE (Global Initiative on Ethics of Autonomous and Intelligent System)

Individuals may provide consent without fully understanding specific terms and conditions agreements. They might also not be equipped with the knowledge and skills to know how the nuanced use of their data to inform personalised algorithms affects their choices at the risk of eroding their agency. Here we take agency as the individual's ability to influence and shape their life trajectory as determined by their cultural and social contexts. Agency in the digital arena enables individuals to make informed decisions where their terms and conditions can be recognised and honoured at an algorithmic level.

Fostering agency requires enabling individuals, especially the more vulnerable and minority groups, to have the skills necessary to challenge unfair decisions and the uneven power dynamics that are enabled by data and data-driven technologies. Students should be aware of how data collection, processing, and use give power to some but not others. For Kennedy, Poell and van Dijck (2015), agency is critical when thinking about the distribution of data power. Yet, in the context of datafication, questions about agency have been overshadowed by oppressive techno-commercial strategies like data mining (p.2).

The IEEE recommends governments and organisations to provide mechanisms to strengthen individual agency through policies that let individuals create, curate, and

A funded project by the EPSRC EP/R045178/1

control the data associated with their identity. Specifically, they recommend the following:

- Create: Provide every individual with the means to create and project their own terms and conditions regarding their personal data that can be read and agreed to at a machine-readable level.
- Curate: Provide every individual with a personal data or algorithmic agent, which they can curate to represent their terms and conditions in any real, digital, or virtual environment.
- Control: Provide every individual access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data.

This is important, as people need to be able to see how their data is being collected by different actors, who are depicting a portrait that can render us subject to all sorts of uses of misuses of automated decisions, leading to what is known as the principal–agent problem. According to political science and economics (also known as agency dilemma or the agency problem) scholarship, this occurs when one person or entity (the "agent") is able to make decisions and/or take actions on behalf of, or that impact upon another person or entity.

Also, the IEEE argues that one of the key challenges is defining how certain uses of data can affect the individual directly. For example, an individual tube user's travel card can track their movements, so it should be protected from uses that identify or profile that individual to make inferences about his/her likes or location generally. Under current business models it is common for people to consent to the sharing of discrete data, like credit card transaction data, answers to test questions, or how many steps they walk. Once aggregated, these data and the associated insights may lead to complex and sensitive conclusions being drawn about individuals.

The Linking Artificial Intelligence Principles (LAIP) proposes the concept of contestability, which can be understood as being when an AI system significantly impacts on a person, community, group or environment, there should be a timely

A funded project by the **EPSRC** EP/R045178/1

process to allow people to challenge the use or output of the AI system. This principle is aimed at ensuring the provision of efficient, accessible mechanisms that allow people to challenge the use or output of an AI system when it significantly impacts upon a person, community, group or environment. The definition of the threshold for ‘significant impact’ will depend on the context, impact and application of the AI system in question.

Knowing that redress for harm is possible when things go wrong, is key to ensuring public trust in AI. Particular attention should be paid to vulnerable persons or groups. There should be sufficient access to the information available to an algorithm, and inferences drawn, to make contestability effective. In the case of decisions significantly affecting rights, there should be an effective system of oversight, which makes appropriate use of human judgment.

Thus, the European Council AI Guidelines propose a framework for Human agency and oversight in which AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should act as enablers to a democratic, flourishing and equitable society by supporting the user’s agency and fostering fundamental rights, whilst also allowing for human oversight.

In his article Why personal agency matters more than personal data, Searls (2018) states that “The first reason we have far too little agency in the networked world is that we settled, way back in 1995, on a model for websites called client-server, which should have been called calf-cow or slave-master, because we’re always the weaker party: dependent, subordinate, secondary. In defaulted regulatory terms, we clients are mere “data subjects,” and only server operators are privileged to be “data controllers,” “data processors,” or both”, and “The second reason agency matters more than data is that nearly the entire market for personal data today is adtech, and adtech is too dysfunctional, too corrupt, too drunk on the data it already has, and absolutely awful at doing what they’ve harvested that data for, which is so machines can guess at what we might want before they shoot “relevant” and “interest-based” ads at our tracked eyeballs”.

A funded project by the **EPSRC** EP/R045178/1

🅭🅯🄏

Privacy, as a right needs to be exercised, thus agency are skills needed to exercise such right and include two layers of abilities, a legal one which means being capable of understanding how data is collected, the terms and conditions and laws that protect the people in order to challenge data protection, and a technical ability, to understand how different platforms and devices capture data and how to prevent and protect one's data, this includes comprehending how encrypted data that has been pseudonymised works and that is reversible.

Some of the personal agency data skills that a person needs to develop, can be understood as,

- The ability of understanding and giving consent for one or more specific purposes;
- Understanding which and how data processing is necessary to enter a contract;
- To understand the processes needed for fulfilling a legal obligation for protecting the vital interests of the user or of another person;
- To understand the necessary process to perform a task carried out in the interest of the public or as contained under the official authority given to the data controller;
- To understand the legitimate interests of the data controller and, rights and freedoms of the user, in particular children.

A funded project by the **EPSRC** EP/R045178/1

5 • Understanding data sovereignty •

By Javiera Atenas

Data sovereignty, according to Kukutai & Taylor (2016) is "right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as their right to maintain, control, protect and develop their intellectual property over these"

This concept Indigenous Data Sovereignty [ID-SOV] was coined by In 2015 when First Nations scholars and leaders from Australia, Aotearoa/New Zealand, Canada and the United States defined data sovereignty to set up guidelines for the publication of indigenous data (e.g. personal, cultural, historical, land), thus, non-Indigenous people must seek consent in research before publishing any data about indigenous people.

According to IWGIA [in Spanish], the sovereignty of indigenous data is defined as the right of indigenous peoples to own, control, access and possess data that comes from them and that refers to their members, knowledge systems, customs or territories. The sovereignty of indigenous data is grounded in the inherent rights to self-determination and governance over their peoples, territories and resources, as stipulated in the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP), thus, recognising that indigenous data is a strategic resource thus, the concept of data sovereignty provides a framework for the ethical use of indigenous information in order to advance the self-determination of the indigenous communities, granting them the right to be decision-makers about how their data is used.

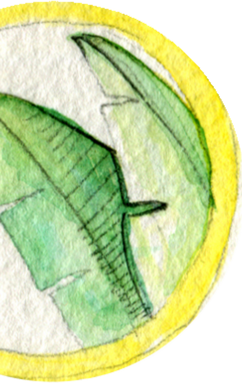
The international approach to the protection of personal data and privacy rights is inadequate for Indigenous Peoples, thus countries need to develop and implement laws, regulations and standards related to the privacy and rights of Indigenous Peoples through legal and regulatory approaches co-designed by themselves based on the principles of ID-SOV.

ID-SOV can be seen as a driving force in giving indigenous communities the right of self data governance using indigenous peoples' values, rights and interests to guide

A funded project by the **EPSRC** EP/R045178/1

decision-making about how their data is collected, consulted, stored and used, giving communities control of their data through data governance policies and practices and through mechanisms and frameworks that reflect indigenous values.

5.1 • Principles of indigenous data governance •



The Global Indigenous Data Alliance (GIDA) aims at providing guidance to co-develop frameworks and guidelines for ID-SOV and to disseminate its implementation internationally, through strategic relationships with global bodies and mechanisms. The United Nations Special Rapporteur on the right to privacy has acknowledged the importance of ID-SOV for the UN Permanent Forum on Indigenous Issues has published some recommendations on Data and Indicators in data disaggregation for indigenous people self-determination and development purposes.

GIDA has published a series of principles for the governance of indigenous data which includes the right to create value from Indigenous data in ways that are grounded in Indigenous worldviews and realise opportunities within the knowledge economy. The four principles can be understood as follows:

1. Collective benefit: Data environments should be designed and function in a way that enables indigenous peoples to benefit from the data.

2. Control authority: The rights and interests of indigenous peoples over indigenous data must be recognized and their authority to control such data must be empowered.

3. Responsibility: Those who work with indigenous data have a responsibility to publicize how that data is used to support self-determination and the collective benefit of indigenous peoples. Accountability requires substantial and openly available evidence of such activities and of the benefits that may be conferred on indigenous peoples.

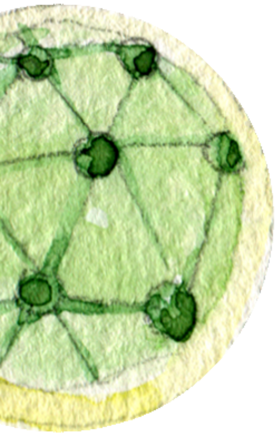
A funded project by the **EPSRC** EP/R045178/1

4. Ethics: The rights and well-being of indigenous peoples must be the primary consideration in all phases of the data life cycle and throughout the data environment.

In Australia, the Maiam nayri Wingara Indigneous Data Sovereignty Collective and the Australian Indigenous Governance Institute have developed protocols and principles for Indigenous Data Sovereignty and Indigenous Data Governance which can be understood as that indigenous peoples have the right to:

- Exercise control of the data ecosystem including creation, development, stewardship, analysis, dissemination and infrastructure.
- Data that is contextual and disaggregated (available and accessible at individual, community and First Nations levels).
- Data that is relevant and empowers sustainable self-determination and effective self-governance.
- Data structures that are accountable to Indigenous peoples and First Nations.
- Data that is protective and respects our individual and collective interests.

A funded project by the **EPSRC** EP/R045178/1



6 • Open data and social innovation •

By Javiera Atenas & Carla Bonina

> In this model of social change, innovators co-create solutions for a wide variety of social problems. These solutions are openly available to other innovators and users, who can then act to further improve and advance them.
> Bonina, López-Berzosa, Scarlata 2020

• Introduction •

The value of open data is realised through its use. However, capturing attention and driving innovation across the spectrum of data users, intermediaries and consumers is challenging. How to encourage participation in the demand for open data? How to generate and grow the community of data users? This module is aimed at explaining what kind of value can be created with open data, who are the main groups of users of the data and how to plan strategies to encourage both the participation of those groups and the generation of new initiatives with open data.

6.1 • The value of open data •

The value of open data lies in its use. This first unit aims to cover the following aspects:

- What kinds of benefits can open data and civic technology generate?
- What examples can be found in the world?
- How are these benefits generated?
- What is the role of participation and collaboration for the generation of value?

In a world with increasing poverty, inequality, environmental degradation and injustice, digital innovation with a social focus appears as a ray of hope to solve such problems. Civic technology is a type of social digital innovation, in which digital technologies are implemented to improve the relationship between citizens and the government in

A funded project by the **EPSRC** EP/R045178/1

order to include more people in public decision-making, citizen empowerment, or to improve access and delivery of public services. In many cases, in particular at a regional level, open data is an input for the development of innovations based on civic technology.

Open data opens up new possibilities to generate value, both in the economic, social, environmental, and democratic fields. There are several open data projects that have a civic or social objective that at the same time may be generating economic benefits. While the types of benefits that open data can bring are not necessarily exclusive, we present them separately below for simplicity.

> **Economic Benefits:** On the economic side, as companies, governments and citizens use or reuse open data, new products, processes and business opportunities emerge. In 2011, a study commissioned by the European Commission estimated that the economic value of opening up and reusing public sector information was approximately € 40 million per year for the European Union alone. On a global scale, McKinsey estimated in 2013 that open data could contribute to the generation of between $ 3 and $ 5 trillion per year in the global economy. Whilst these figures are still speculative, they open up a range of possibilities when it comes to generating economic value.
>
> **Social and democratic benefits of civic technology and open data**: Equally important, open data is seen as particularly beneficial in increasing transparency, fighting corruption, and promoting social inclusion, but the value of these benefits is much more difficult to assess in monetary terms. This perspective of value has been particularly important in Latin America, where the lack of transparency and accountability, as well as social needs, have been at the centre of the agenda of the open data movement in the region. Concern about transparency and accountability is not, however, a problem exclusively for developing countries. Open data policies in pioneer countries, for example, the UK, were partly the result of a lack of trust in the political system and its accountability.

A funded project by the **EPSRC** EP/R045178/1

There are also many cases in which open data is built on the basis of volunteer work, cooperation and civic engagement. There are open data initiatives of collaborative peer production based on the public domain, in which participants collaborate, contribute their time and exchange knowledge and experience to achieve a social or public interest good. A pioneering initiative in the field of geographic location is the OpenStreetMap, a collaborative project to create a free editable world map. Similar to Wikipedia, Open Street Map is produced by a community of local knowledgeable contributors, who update and maintain data on routes, train stations, cafes and much more around the world, all in an open format. This project has inherent utilitarian value - anyone with a GPS can download and use it for free, and it has proven particularly useful in remote areas. However, this common task has a higher value. As in other cases of open data, OpenStreetMap can be used as a platform to create other services based on mapping data, such as transport, humanitarian aid or monitoring.

Participation, collaboration and co-creation with civic technology and open data: Both, civic technology and open data, promise a variety of benefits economically as well as socially, politically and democratically. The potential of open data not only at a regional level but also globally, is particularly facilitated by greater access and affordability to the Internet, computers and mobile telephony, among other digital technologies. The generation and appropriation of its benefits result from the participation, collaboration or co-creation of services by a diversity of actors, two fundamental pillars in the framework of Open Government.

The concepts of participation, collaboration and co-creation are undoubtedly related. For, public or civic participation can lead to cases of collaboration or co-creation of services. For example, an open session between the government, civil society organisations and citizens can give ideas to prioritise a series of data to be opened up, which, over time, can also lead to concrete initiatives between a civil society organisation and the government. In other cases, an

A funded project by the **EPSRC** EP/R045178/1

open session can be a good input to modernise a government service, which is not necessarily co-created in collaboration with an outsider.

In simple terms, we understand co-creation as pertaining to those processes or activities where at least two actors (e.g, public, private, governmental or civic) collaborate in the realisation of a project to achieve a certain result. In other words, it refers to where more than one part is required for value creation.

6.2 • Open data for social innovation •

Despite expectations, realising the potential benefits of open data is complex. Part of the problem governments face is the need to cultivate an ecosystem of reusers. On the demand side, the authority responsible for the open data platform needs to cultivate a base of independent reusers, either to contribute new services or applications based on open data, as well as end users who consume these services. A key challenge that authorities of open data platforms face is governance, that is, how to attract and keep active those who have the capacity to generate new services and applications for the economy. This unit introduces a simple model to foster innovation with open data, based on rules and tools.

The logic of the contribution of open data to the generation of economic benefits can be summarised as pertaining to two types as follows.

- Direct benefits: include the generation of income from new products or services, the creation of jobs and/or the reduction of economic costs. An example would be a new company that earns its revenue from visualising or transforming open data for the financial industry.

- Indirect benefits: the reduction of waiting times for users of applications with open data, the increase in search options for decision making, the increase in the efficiency of the public sector and the growth of related industries, among others. An example of indirect benefits would be the reduction of hours lost in traffic by the inhabitants of a region, based on the use of open data regarding transport.

A funded project by the EPSRC EP/R045178/1

🅭 🅯 🄏

To cultivate an ecosystem of entrepreneurs and innovators around open data, the innovation platform approach is useful. It is key to having tools for fostering innovation with open data and working with a series of priorities or public problems to be solved. In order to cultivate the demand of entrepreneurs and innovators with open data, among the various social tools that can be used, it is possible to distinguish two types: i) information and software, and ii) social.

Information tools and software: These are the most obvious resources on an open data platform, and comprise data sets as well as other information tools, such as manuals and related videos for use. They also include software tools, such as APIs and web portals, that allow developers and innovators to access data sets programmatically.

Social tools: These tools complement those of information and software, being those that involve the participation of demand, through the deployment of events, financing or other types of incentives, of which three stand out.

- Hackathons: These are marathons of ideas, design and software prototyping, bringing together specialists, technical programmers, designers and entrepreneurs to work collaboratively. The sessions usually last between 24 and 48 hours, having specific challenges to guide the event, and in many cases they also have parallel workshops for capacity building. Hackathons are often one of the first strategies to promote the reuse of open data — and at the same time they create or strengthen the community, especially of entrepreneur

- App challenges: These competitions are similar to hackathons, but typically highlight the competitive element with prizes or recognition awarded by a panel of experts or judges, and can last longer than a hackathon. There are competition rules, which specify the products or applications considered "eligible" for the awards.

- Incubation processes and funds: These tools are used to incubate well-formulated projects that have potential for the development of an application, product or service based on open data. For this purpose, seed financing is offered to build and apply scalable models as these programmes may be more successful in generating impact with open

A funded project by the **EPSRC** EP/R045178/1

data. In reality, there are very few examples in Latin America regarding such programmes.

To ensure that Open Data properly benefits the society, it is key to promote more channels of collaboration between citizens, business and government. Opening up data without a community using it will not generate value per se. It is common to find that public officials report difficulties in prioritising those data that may be of greatest interest to the groups that make up the demand for data. Facilitating spaces for collaboration between the government and these actors is a strategy that has proven to be effective when creating and promoting the use of open data. Also, it is important to invest in understanding open data business models. Whilst technology entrepreneurs, startups and private sector companies are very relevant actors in the open data ecosystem, they are still disconnected or even absent in open data discussions. This is because it is not easy to generate business with open resources and hence, emphasis also needs to be placed on acquiring better understanding as to what kinds of business models are viable or sustainable in the ecosystem.

Back home

Disclaimer: unless otherwise stated, all the content produced by the authors for this project is licensed as CC-BY-NC, however, any content produced by third parties such as some of the activities, videos, and other resources may have a different license, so please make sure that you understand each individual licence before using or reusing the content.

Illustrations, editorial and graphic design by

A funded project by the **EPSRC** EP/R045178/1