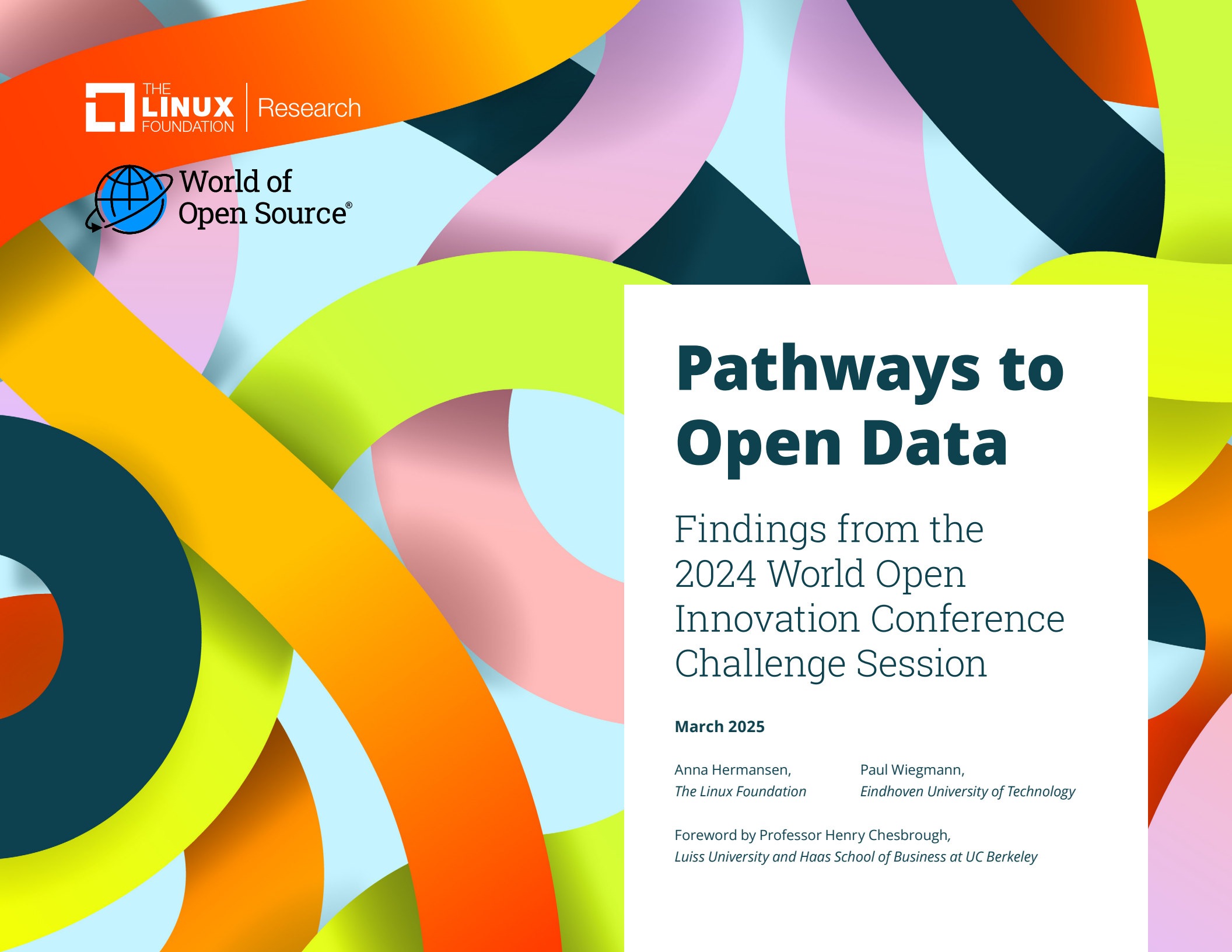


Pathways to Open Data

Findings from the 2024 World Open Innovation Conference Challenge Session

March 2025

Anna Hermansen,
The Linux Foundation

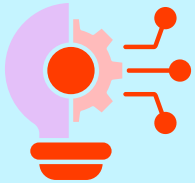
Paul Wiegmann,
Eindhoven University of Technology

Foreword by Professor Henry Chesbrough,
Luiss University and Haas School of Business at UC Berkeley

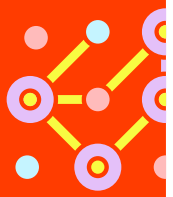
Pathways to Open Data

Data silos hamstring research & innovation, and have become increasingly onerous alongside growing data needs to train AI models.

Open data is freely accessible for universal use, leading to **new avenues for innovation, greater reliability, & increased trust.**



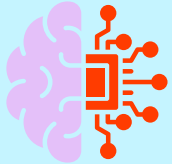
The unique qualities of data as compared to software — such as **maintenance, quality, privacy, & license diversity** — make its openness challenging.



Significant human resources are required for cleaning, standardizing, & maintaining a dataset.

The financial & resource **costs of dataset maintenance** engender a **tradeoff between** the **quality** of the data & the **cost** of accessing it.

While a lack of standardization makes datasets unusable, AI tools offer opportunities to better manage unstructured data.



Data privacy concerns stem from compliance with regulations such as GDPR, which create **a climate of risk aversion.**

Proprietary control over data gives companies greater certainty around compliance & quality while **reducing the fear of losing competitive advantage.**

"Semi-open" data platforms allow for collaborators to share best practices and other pre-competitive data while maintaining their competitive advantage.

Overture Maps Foundation has built an **open, agnostic, & standardized geospatial data platform** for data owners & service providers to leverage.

Building open data infrastructure requires a **reworking of current data collection & sharing processes.**

Open data requires **incentivizing collaboration** around a pre-competitive layer while **incorporating checks & balances** in the governance structure.

Contents

Foreword

In an increasingly digital world, we are all both users and producers of data. But we often ignore the possible ramifications of this, as we are hustling to make a purchase, read a story, make a post, or react to a photo. Companies that figured this out early on have become enormously valuable, and now mediate the access to the data they have accumulated. The concept of open data is a response to this state of affairs. This is the backdrop of this report, on pathways to open data, supported by the Linux Foundation.

The Linux Foundation chose to host a workshop on the topic of open data at the 11th meeting of the World Open Innovation Conference. Fueling interest and participation in the workshop was the rapid growth of artificial intelligence software (AI), which requires extensive amounts of data to train the algorithms that AI employs.

This report summarizes the takeaways from the workshop, so I will simply underscore a few insights here that I found helpful. One was that everyone wants to protect their own data, yet everyone seeks algorithms that can perform very well. To get algorithms to perform at a high level, a lot of data are required. So for all but the very, very largest organizations, it makes sense to open up access to the data so that better AI algorithms result.

It is also critical to recognize that data will grow and change over time. So there isn't a one-time set of actions and expenses to move to Open Data. Rather, this will be an ongoing journey, and identifying an economic model to support the cost and effort to keep up with the inevitable changes in the data must become part of management's commitment to support the project.

Where to from here? There were at least three ideas put forward about the ways in which Open Data might advance in the future. The first was data ownership, in which users might have the ability to provide their personal data under certain conditions, and might choose to restrict the use of their data in other conditions. A second idea was to create incentives for contributing data to a "pre-competitive" data set. This would protect contributed data from being used to identify specific people, for example, while allowing more general characteristics to be analyzed. Importantly, this pre-competitive dataset would be made widely available, democratizing access to data that previously was prohibitively expensive to access, or simply unavailable, to smaller firms and individuals.

A third important idea was that of governance. Repositories of large amounts of data must be stored somewhere. There are costs for hardware, for software, for security, and for maintenance. In order to sustain broad access to useful data repositories, there needs to be an economic model of some kind. And the decisions that are taken around access to data, and whatever expenses might be involved in that access, have to be taken within a governance mechanism that is credible to the stakeholders supporting the Open Data process.

Henry Chesbrough
LUISS University in Rome, UC Berkeley in USA

Introduction: The history and future of open data

Our data is everywhere and powering everything. From marketing, to healthcare, to government services, to the emerging phenomenon of programming AI agents, organizations leverage data to be as efficient and effective as possible. However, data is often siloed within entities and any third-party data access requires overcoming significant technical, legal, economic, operational, and cultural obstacles that are multifactorial and at times may seem intractable.[1] The increasing reliance on data calls for an assessment of these obstacles and how organizations can shift toward greater openness and sharing.

The concept of open data has its roots in open science, where non-personal and non-commercial data is freely published for the purpose of greater innovation, transparency, and collaboration.[2] This culture of openness is strongest in public institutions, where the data collected is considered a public good without profit-generating opportunities, and where transparency of government and public-sector information is encouraged.[3] The Open Government was popularized in the 2000s with the Obama administration's open data initiative (2009) and the Public Sector Information Directive in Europe (2003),[1] and soon, many governments developed open data portals for citizens to access and analyze public information about their municipality. According to the United States government's data.gov portal, its mission is to "unleash the power of government open data to inform decisions by the public and policymakers, drive innovation and economic activity, achieve agency missions, and strengthen the foundation of an open and transparent government."[4]

When entering commercial and/or personal data ecosystems, the notion of open data becomes much more complex. Without the open government mandate, organizations grapple with profit incentives, privacy concerns, and expectations of control that diminish the value of open data in the eyes of data owners. For some sectors, data sharing becomes an ethical imperative (e.g., in healthcare), while others may be incentivized by the value of triangulating with other third party datasets (e.g., in marketing).[5, 6] However, when data access involves personally identifiable information, abiding by the privacy regulations that protect this data becomes paramount, and opening up data becomes risky. Added to this privacy risk is the fact that data generation and collection has become a key component of the profit model, causing large corporations to build walled gardens around their data and controlling the flow of information.[7] The current data market model consists of commercial entities that take ownership of the "data commons."[9]

The walled garden concept is felt across industries and sectors. For example, in healthcare, data is siloed within different hospitals and clinics using their own electronic record systems that lack interoperability between the different systems. These silos reduce the value of the data and negatively impact the patient, the clinician, and the researcher. They also cause a lack of standardization, making the data messy, fractal, and even unusable.[8] The European Commission has worked on initiatives and programs for electronic patient file transfer across healthcare providers in different countries, but this interoperability is still nascent and is not the norm across

Defining open data

In this research, we defined open data as data infrastructure that has the technical and legal requirements in place to make the data freely accessible for universal use, reuse, and redistribution.[16] We also explored openness beyond data-centered definitions. This included dimensions of openness in the context of open standards, such as: access to, control over, and cost of the development of the artefact; access to, control over, and cost of use; the completeness of the artefact; sharing of the artefact; and collaboration with competing systems.[17]

many geographies.[9] Similarly, as the energy sector digitally transforms and electrifies, sharing the data collected between all the connected devices in a system is a challenge without standardization and interoperability. Without better access to data generated at different points in the system, operators and distributors lack insights needed to study demand and grid health.[10]

Although access to data is not a new problem, the explosion of generative AI tooling has introduced a heightened pressure for data needed to train the models — and in particular, data that is licensed in a way that makes this kind of use legal. Organizations are turning to their own proprietary data to train their models. As found by Lawson et al (2024), organizations are relying on a portion of their own data to train both their proprietary models and the open source models they are implementing.[13] The desire to build proprietary models is strong, as it gives organizations more control over their data.[11] However, complete reliance on proprietary data is not sustainable, and organizations are in need of quality training data from other sources to build effective, robust, and unbiased models.[10] In this regard, data governance becomes a top priority for open source AI projects, where data workflows are managed responsibly with attention to quality and compliance.[13]

In this next era, where generative AI becomes a key tool across industries, the future of open data becomes paramount. In November 2024, the authors of this report attended the World Open Innovation Conference (WOIC) in Berkeley, CA, and held a session asking participants: **What are the pathways to open and accessible data?** Focusing on the data ecosystem's obstacles, needs, and opportunities, we asked participants to discuss the following questions:

- What are some challenges you face in access to and use of data?
- How does your organization or project rely on data to innovate?
- How does your organization make its data open to internal and external access?
- How do you access relevant third-party data?
- How have you incorporated technology to address your data needs?
- What solutions outside of the technological realm have been implemented at your organization (e.g. cultural or policy change)?
- How do you believe we can make data more open? What is needed, given your experience?

The following report is structured around a thematic analysis of this 75-minute session. Under The Chatham House Rule, participants are free to use the information received, but none of the participants from the session may be identified. Session participants included academics and practitioners from a variety of industry sectors with expertise in the area of Open Innovation. They shared relevant insights about barriers and opportunities for open data, based both on their theoretical expertise and their practical experience of working with open and closed data in various settings.

Why open data matters

For the audience of entrepreneurs, academics, and innovators, access to data is a crucial part of business intelligence and innovation. According to one participant, the analysis of publicly-available data about investments is an important factor of business intelligence for their clients. Another participant discussed the value of individual-level internal employee data, and their client's desire for transparency of this level of data. They commented, "the team leader can look at the data and say, look, we can do that. It's empowerment." Being able to drive certain outcomes using public and internal data makes the case for data openness and accessibility. As studied by Ambiel (2024), triangulating third party data with internal proprietary data "is essential to train large AI models, validate research, or discover market opportunities."[7]

Session participants noted that the current state of data access does not necessarily allow for these opportunities and points of validation. For example, as one participant explained, "We are finding it hard to find out who is researching what." Similarly, another participant expressed the lack of collaboration among academics on one dataset: "If you get your hands by any chance on a good data source, it's a gold mine. Then, you cannot possibly analyze all facets of it, and you don't have the bandwidth to understand what this could mean for another science — for example, maybe it's good for engineering, maybe it's good for social science, maybe chemistry … but how would I even know?" This represents a missed opportunity, where that dataset may be useful to other research teams but is not discoverable by those groups.

These missed opportunities mean unfulfilled innovation. One participant in the sports industry commented, "Sports data wants to be free … where the teams compete is to find a commercially valuable product that is usually a value add on top of the data. In other words, predictive analytics." Sharing data, even among competitors, can help an organization innovate faster and build a product on top of that shared data. This collaboration produces big data that allows analysts to abstract away from the individuals who represent the datapoints, reducing privacy concerns. One participant gave the example of FootFall data, where an individual's location data can be very personal on its own, but when aggregated with other datapoints to demonstrate how many people show up at a location, "it [becomes] just the general histogram. And so you can abstract that data into something less personal." As another participant stated, "If you zoom in on the micro level of big data, it's worthless anyway. It's only the trends and the analysis on top … it only becomes useful the moment that it's big enough." Knowing that the bigger the dataset the more valuable it is, participants argued that this should be an incentive to contribute data — to make a more valid dataset for all.

Beyond the analytical value of a shared dataset, this activity also increases trust. One participant gave an example of a consortium around a data sharing and analytics platform and how "there's an implication of trust by the participants … that is rather compelling, when you're talking about building a partnership." This social contract of collaboration allows for shared activities that are dependent on trust, as described by another participant, "Maybe I cannot see [the data] myself … but I can rely on my partner to do that for me. And I have to trust them, but I trust them because we're part of the same group." The act of opening up data leads to new avenues for innovation, increased effectiveness and reliability of datasets, and greater team trust.

Current challenges of open data

As discussed above, open data platforms are hampered by a myriad of technical, regulatory, economic, and cultural challenges. In the first half of the challenge session, we introduced some of these barriers and asked participants to reflect on their own experiences confronting them in their work.

Uniqueness of data

When considering the challenges faced by open data, it is important to consider the characteristics of data that make it unique as compared to other content, such as software. In his blog post, Marc Prioleau, the executive director of Overture Maps Foundation, lists six characteristics that make open data different from open code:

- The proprietary origins of data;
- The patchwork of data licenses to navigate;
- The scale and cost of collecting, hosting, and maintaining data;
- The workflows required for the ongoing production of data;
- Assuring accuracy and quality of data, and;
- Protecting personally identifiable information.[12]

Bennet et al (2024) also point out the unique challenges faced by data-intensive applications — in their example, AI applications — in particular the potential violations of consent and managing the different open licenses of datasets.[13] Participants picked up on these and other challenges during the challenge session.

The cost / quality tradeoff

One theme that came up a number of times during the session was the idea that there exists a tradeoff between cost and quality. Some participants expressed that they pay for data because they consider it better quality than open data: "I would pay for private data … because it's better data, curated and so on," one participant said. However, this is an expensive option that is "not sustainable in our business model," and so they use a mixture of free and paid sources. This tradeoff was expressed by another participant, explaining that they bolster their free data with paid sources, but "if I had unlimited money, I would pay for private data … because it's better data, curated and so on."

Reflecting on the cost of data, participants brought up the expense of curating data — and the lack of incentive to do so without charging for access. Without an economic model, open datasets rely on volunteer contributions that are considered unreliable and that are not standardized. As one participant reflected, "It would be nice if the people creating the data did it in a standard way, right? … But the problem is, they don't really have much benefit … [and] unless they have a benefit to doing it, they're going to say, Why should I do that? … The person who actually has to do it either doesn't have an incentive or they're not forced to do it."

Another complication in maintaining a quality dataset is data mutability. The artifacts that data are collected on can change, which make datasets less reliable. One participant gave the example of mapping data: "The hard part about maps is, they reflect the physical world, and the physical world changes, so the map data has to change." The speed at which these artifacts can change in some sectors creates the need for a continuous feedback loop to make sure the data reflects reality.

Without continuous updating and maintenance of the dataset, concerns arise around the quality of the data. This includes whether or not the data is up to date, and how well the data represents different populations and geographies. A participant gave an example from their work, explaining how they want to provide their clients access to worldwide information, but this is rarely the case: "Worldwide means North America, maybe Brazil if we have data from Brazil,

Central and Western Europe, but maybe not Eastern Europe. So it comes with all sorts of different geographical limitations." Because of these limitations, skepticism arises around claims of a dataset being up to date and complete. This puts pressure on those relying on open datasets, as they become a "source of truth" to their customers, despite third-party data streams being outside of their control.

The labor-intensity of creating open datasets

As described above, data curation is expensive. This is in large part due to the labor required to manage the different workflows of data, from collection to maintenance to quality control. As one participant plainly stated, "I'm currently doing a lot of data cleaning. I think there's no shortcut for that. There's no workaround, it's part of research, of course." However, there are also important resource considerations for integrating third-party open datasets within an organization's intelligence. Participants explained the work that goes into conducting quality control on their use of open datasets: "I cannot even describe to you how difficult it is to find up-to-date information … you always end up having to call the person, and that is labor intensive, it doesn't work … These databases are a good starting point but they are rarely the single source of the truth or the end point of the search."

Because these datasets are often incomplete or are potentially unreliable, a decision must be made on how to use them. As one participant explained, "I need to import the data, good or bad, and then I need to do some manual work. And then the challenge is, do I put in the work, or do I just leave it one-quarter filled out and the rest is not filled out because I don't want to google it myself … and what this leads to, which is the ultimate challenge, is an incomplete data set, and then it's unusable."

Of course, once the data is cleaned, there is still significant human input involved in subsequent activities. "It can last years," one participant stated. "Maybe it should not just be, how many hours do you put into cleaning, but how many hours you put into analyzing, researching, publishing, reviewing."

Standardization

One important aspect of cleaning data is its standardization. If a dataset is not standardized, this impacts the reliability of the data and its usefulness in comparison with other data, as expressed by one participant: "If everyone can write whatever, you will never get standardized data, which can give us the overall picture — what are the competencies of that department or that group of researchers and stuff like this. So then, we would have a false positive that we found the right person for that problem … So this is the main problem for us right now." The process of standardization is more complex for some industries than others, impacting the ability for stakeholders to read and use third-party data: "For engineering data, you have a number and you have a unit, and that's it, and maybe a timestamp that comes with it," explained a participant. "But to interpret healthcare data, you also need the methods used, the conditions where it was measured, when it was measured, and so on." This makes cross-industry data sharing even more complex.

Although standardization came up as a significant concern, some expressed that it may be less of a problem in the future, in part because of AI. As one participant stated, "I would expect that standardization becomes easier over time, on two aspects. First of all, there's … more and more data, which is by default being annotated and more structured than it used to be, let's say, ten years ago … Secondly, … we see more and more algorithms capable of doing something structured with unstructured data. So for me, from a technology point of view, I'm very optimistic that that problem will solve itself." The benefits of transparency around the need for standardization, and the tools available to make unstructured data more usable, may potentially diminish the impact that non-standardization has on data sharing practices.

Data privacy

Beyond the quality of the data, protecting data for privacy and business reasons was considered another significant barrier to open data. The potential to expose sensitive data, such as personally iden-

tifiable information (PII), made some participants hesitant to open their data: "We have to develop ourselves, because our data are private, so we cannot use open source, open data. So we have all our AI and advanced analytics groups who develop our own on-prem tools to monitor everything." Concerns mainly came from compliance with regulations, such as regulations to only use on-prem storage and not cloud storage, as well as complying across different borders: "governance mechanisms and policies are quite sensitive when it comes to how open you can make the data, because it really depends on each country, right?" Participants also expressed concern about AI models, where "each country has their own AI act or not, and that makes it quite difficult when the data — because data is crossing borders — becomes global." The complexity around which regulations will apply to the activity, particularly when considering activities that flow across borders, cause understandable hesitation.

The General Data Protection Regulation (GDPR) became the archetype for a number of participants when discussing the impacts of regulation on data sharing. One participant shared an example from work they had completed for a client, where "[the] company wanted to map how the workers in the production process move, in order to avoid risk for them. But then it came out that there was a GDPR problem because they were collecting personal data." This hampered the participant's ability to provide meaningful results, and interestingly, was in opposition to what the client and the workers wanted. Despite GDPR regulations prohibiting the collection and sharing of datasets comprising fewer than five people, "people actually would like to do it [regardless], because for them, it's a great tool to work with their data and to see the data … from that perspective, the employees actually kind of fight with us against the Worker's Council, because they want their data to be seen."

According to one participant, GDPR's negative effect on data sharing is not uncommon, where "instead of using GDPR for its intended purpose, they just have made everybody scared for what it could do. The first five years, GDPR was only used to kill stuff, whereas, in fact, GDPR allows tons of stuff. There's no issue. But most people who don't know what you can do with it, they just go to this very safe side and say, you can't do anything anymore." This creates asymmetric risk, as referred to by another participant, where if a lawyer approves a certain data sharing activity and they're wrong, they risk losing their job, and so it becomes easier to deny where any uncertainty exists. "When you ask, can I share this data? That eventually goes to someone who is incentivized to tell you no … So, you have to solve the asymmetric risk." This creates strong resistance to opening up a dataset.

Control over data

A need for control, from a regulatory as well as business perspective, explained the resistance to open data. This manifested as internal scrutiny, where one participant explained how their "IT [people] don't want to provide open access to data." As discussed above, an organization's legal team can also impede data access: "Data sharing or data receiving is stopped by someone saying — It's almost like somebody's saying, Listen, we have to ask Mr. X, and this guy, with all respect, is kind of a legal guide … Even if we needed to open source code, whatever, open source software which is available, we need to use the IT department to say no I'm not allowed to. So sometimes you get into conflict with the regulation." Finally, another form of internal control came from one participant who explained, "instead of relying on [a third party's] blood pressure measurements, it's way more safe for me to just redo the measurement, because then I have everything under control." The safer and easier alternative is for organizations to keep their data closed.

Beyond the privacy and quality concerns that keep data closed, participants also expressed the possibility of losing a competitive advantage by sharing data. One academic participant gave an example in the context of publishing a paper as well as the dataset they collected, where they ask themselves, "should we publish [the dataset] beforehand for other researchers? … It's a question of strategy, and also, to be honest, being afraid of other researchers being much faster than us in using the data set afterwards in the

same research topic we are in." There was a fear that opening up an organization's data to the public would diminish its potential.

This leads to the question: Is some data not meant to be open? One participant expressed this sentiment, arguing: "Of course, not all data lends itself to being open." As discussed by Ambiel (2024), for some enterprises and organizations in industries such as financial services and healthcare, their data is too valuable or sensitive to introduce new risks, and as a result must limit the distribution and use of their data.[7] However, one participant made the case that there is some nuance to consider: "Is all your data highly proprietary? What is the data that actually is proprietary to you and your competitive advantage, and what data actually is not that proprietary?" It is important to consider what data that makes up your competitive advantage, or is too sensitive, compared to the data that would be more valuable when shared with others.

Open data successes & opportunities

The second half of the challenge session focused on the future of open data, and participants discussed case studies and ideas for the open data ecosystem. One participant shared an example of a fundraising platform where collaborators share best practices and outcomes from working with startups. In this system, "the only open data is all the characteristics of the startup, so who they worked with, and they had some data about the timeline, when the experiment took place, the scale … So you have to contribute enough [data] to say we did something or we didn't achieve something, but not enough for you to give away, for example, the donors' names. So, semi-open data." Building this pre-competitive layer provides an opportunity for different players in the ecosystem to share non-proprietary data in a way that benefits all groups, without reducing the advantage of the individual organization.

From an academic perspective, the concern around openly publishing datasets led to discussion around potential crediting and licensing strategies. One participant asked the group, "You'd like to make [the dataset] available before publication, but you fear that others might publish before you … so what if, just by collecting the data, you get the credit anyway? They are doing the research faster, but still, you get the credit … licensing the data. This is my data, if you use it for research, I get credit for that." This audit trail of data collection and use through licensing presents a potential solution to competitive advantage fears, in particular in an academic context.

OVERTURE MAPS

Overture Maps Foundation is transforming the mapping industry by creating reliable, interoperable open map data that is freely accessible for use in any map product. Through strategic collaboration, member organizations develop standardized schemas and datasets, combining data from community, government, and corporate sources. Overture ensures data quality through rigorous validation and standardization, ensuring its suitability for commercial applications while maintaining the benefits of open data.

Overture addresses a fundamental industry challenge: the increasing cost and complexity of processing and conflating geospatial data, which often exceeds licensing costs. By building shared infrastructure and standardized data pipelines, Overture eliminates redundant efforts across organizations. A key innovation is the Global Entity Reference System (GERS), which provides stable, unique identifiers to map features globally. GERS is distinctive for being global, open, and entity-based, enabling organizations to link external data directly to the base map and ensure interoperability across applications. This collaborative approach enables organizations to focus on value-added services while leveraging a standardized, continuously improving base layer that accelerates innovation across the industry.

Overture has made significant strides since it was founded in December 2022, with its data powering applications used by hundreds of millions of consumers through platforms like Facebook, Instagram, Bing/Azure maps, and Esri's ArcGIS Living Atlas. As of 2024, Overture has released production-ready datasets covering 2.3 billion building footprints, 54 million points of interest, divisions, and contextual layers including land and water data. The transportation dataset maps 86 million kilometers of roads worldwide, including detailed traffic rules and restrictions. From its four founding members, Overture has expanded to over 37 organizations across diverse sectors, establishing itself as an open foundational layer for the entire mapping ecosystem.[14]

Next steps: What is needed for open data structures

The discussion on open data case studies and opportunities led the analysis to some next steps to build more open data systems. Similar to what was investigated by Majer (2024), the entire walled garden approach needs to be dismantled with new governance mechanisms, decentralization, collaboration, and open source.[9] Analysis of the discussion revealed three important themes to help reshape the data sharing landscape and shift the ecosystem toward greater openness.

First, **data ownership**, as a significant public concern, is a useful avenue to rethink current data collection and sharing practices. One participant referred to the current power and control dynamics as "asymmetric," stating, "we come from an era where data was mainly gathered and used for tons of money, generating a business model, without me as a user ever getting feedback or ever getting a refund … And this is the reason why we kind of over-regulated a number of things." From their perspective, this asymmetric power dynamic — and the regulations that attempt to counter it — could be addressed through reconfiguring usage rights. This could look like, "I give you my data, or I don't give you my data. Or, I give you my data, but I only allow you to do certain things with it … I'll give you my data, but you can't use it for advertising or anything. You can only use it to cure cancer or something like that." These usage rights increase visibility and transparency of the use of data, reduce fears around data openness and privacy, and create an environment where sharing becomes more important than protecting data.

Participants shared technological ideas as a way to establish usage rights in practice. For example, one participant discussed the idea of guaranteeing that the data will be used a particular way: "how are you addressing that? Just thinking about, you know, blockchain or something, I think technology can play a role here in giving the assurance to the people who are willing to share their data under certain conditions or for certain purposes." Another participant agreed, saying, "you could add a layer of smart contract, or something with a blockchain." Beyond blockchain, another participant suggested the Solid standard as a potential way to regain control over one's own data: "For me, that's going to be a huge change … Because I as a user will be able to switch on or switch off sharing at my volition, so the moment you as a company are no longer doing what I like, I just turn it down, which is something even today is not possible."

Data ownership should also be managed through licenses such as the Community Data License Agreement (CDLA) which provides the legal framework to share data. Their latest release, CDLA 2.0, outlines the terms under which the data can be used, modified, and shared, protecting the owner of the data while allowing widespread sharing and use. When building a governance structure and collaboration model around an open dataset, this kind of license provides structure around dataset activities that increases confidence and streamlines the data sharing process.[15]

Second, **incentivizing collaboration around a pre-competitive dataset** could address some data sharing challenges and support a cultural shift toward greater openness. As described above, participants understood that there is a layer of data — or an abstraction of data — that becomes useful for collaboration purposes with others in the industry. Building a value proposition to contribute data to a collaborative dataset is key for this to happen. As one participant noted, "Very often the incentive to build the dataset in a certain way is dependent on one party, but benefits another party … The question is, how do you get the people who have the [data] to give that to you?" The incentive for collaboration then becomes the positive externalities that take place when data is shared. For some, building a dataset with others means adding datapoints that actually make the dataset workable: "In those early phases, sometimes it makes sense [to share data], just because there's not enough data. So that could be an incentive. There's not

enough. If I only have 100 observations, and maybe someone else has 200 and I find 100 more somewhere, maybe that makes for a more valid data set for all of us." This contribution makes the shared dataset more valuable, and leads to greater potential for innovation.

Altruism is another key incentive to consider, as hinted at in the discussion of user rights. This is particularly clear in the health sector, where sharing data to save lives is a strong incentive for most. However, encouraging altruistic behaviour is dependent on the way that the value proposition is stated, as one participant expressed: "If we ask the question … can we use your data for drug development, where then the net result is Pfizer, or X company … a number of studies have come to the conclusion people do not want to share. But if you make the value proposition different, more like: with your blood samples, we will be able to find the treatment for cancer, and put it open, and you kind of guarantee that it will not go to just one player on its own — all of a sudden, a lot of people become quite okay with sharing all their medical data." This is based on trust in the data request, and that the entity will be using it the way they say they will: "I'm going to share it … because I'm helping this or that … you're hoping that ends up somewhere, and you trust that party to do the right thing with it." Altruism is a key behavioral mechanism that can help incentivize greater collaboration around contributing to an open dataset.

Third, a stumbling block to building open datasets is outlining the right kind of **governance structure** that balances a culture of collaboration and neutrality while still managing for checks and balances. Current perspectives of open datasets, at least from some members of the challenge session, is that there is no ownership, which impacts the quality of the data. As one mentioned, "No one technically owns the data, so there's no incentive to keep it updated." When considering how to solve for this, one participant suggested, "what kind of government mechanism or policy would be needed that this could happen, right? … What would [an open database] need? Trust, and hierarchy in some way." According to another participant, encouraging a hierarchy would mean that one entity hosts it, pays for the servers, and manages access and security: "Who will pay for all the servers? And if I want to access it, I need a user name, and who will take care of that? Security? … You want logs of who accessed it when. Someone has to take care of that. Whose IT office should do that? Inevitably you need a hierarchy." Considering new forms of governance that support incentives to share data and bring the individual back into the process has the potential to transform the data landscape and encourage greater publishing of data for the benefit of all.

Conclusion

The WOIC Challenge session revealed insights into how academics and practitioners are considering the tradeoffs between open and closed data and identified some realistic concerns and expectations for open databases. Through the analysis and reporting of this session, we hope to shed light on the importance of open data and encourage those working with data to consider the ways that they can better collaborate on datasets, incentivize sharing, and reshape the culture of their organization to support greater openness. As policies and cultures shift with new technologies, new governments, and new economic concerns, it is crucial to establish an orientation of openness no matter the headwinds.

Methodology

The findings discussed in this study were developed from transcripts of a 75-minute session at the 2024 World Open Innovation Conference in Berkeley, California on November 6th. The authors hosted the session, introducing the topic before guiding the group discussion. The discussion was recorded and turned into transcripts using Otter.ai. The first author coded the transcripts, developed themes from patterns in the codes, and wrote the report using secondary literature to bolster the findings. The report underwent peer review by the second author and other stakeholders before production.

Endnotes

1 Attard, Judie, Fabrizio Orlandi, Simon Scerri, et al. “A systematic review of open government data initiatives.” Government Information Quarterly, no. 32 (October 2015): 399-418. https://doi.org/10.1016/j.giq.2015.07.006

2 Braunschweig, Katrin, Julian Eberius, Maik Thiele and Wolfgang Lehner. “The State of Open Data: Limits of Current Open Data Platforms.” (2012). https://api.semanticscholar.org/CorpusID:17298359

3 Zuiderwijk, Anneke and Marijn Janssen. “Open data policies, their implementation and impact: A framework for comparison.” Government Information Quarterly, no. 31 (January 2014): 17-29. https://doi.org/10.1016/j.giq.2013.04.003

4 “Data.gov Home.” Data.gov, accessed February 14, 2025. https://data.gov/

5 Ambiel, Suzanne. “The Case for Confidential Computing: Delivering Business Value Through Protected, Confidential Data Processing.” The Linux Foundation. July 2024. https://www.linuxfoundation.org/research/confidential-computing-use-case-study

6 Gaba, Jeanne Fabiola, Maximilian Siebert, Alain Dupuy, et al. “Funders’ data-sharing policies in therapeutic research: A survey of commercial and non-commercial funders.” PLoS ONE, 15(8). https://doi.org/10.1371/journal.pone.0237464

7 Majer, Alan. “Decentralization and AI: The Building Blocks of a Resilient and Open Digital Future.” The Linux Foundation. November 2024. https://www.linuxfoundation.org/research/decentralized-internet

8 Hermansen, Anna. “An Open Architecture for Health Data Interoperability: How Open Source Can Help the Healthcare Sector Overcome the ‘Information Dark Ages.’” The Linux Foundation. October 2024. https://www.linuxfoundation.org/research/health-data-interoperability

9 “Exchange of electronic health records across the EU.” European Commission, accessed February 25, 2025. https://digital-strategy.ec.europa.eu/en/policies/electronic-health-records

10 Dover, Mike. “Open Source and Energy Interoperability: Opportunities for Energy Stakeholders in Canada.” The Linux Foundation. August 2024. https://www.linuxfoundation.org/research/canadian-energy-interoperability

11 Lawson, Adrienn, Stephen Hendrick, Nancy Rausch, et al. “Shaping the Future of Generative AI: The Impact of Open Source Innovation.” The Linux Foundation. November 2024. https://www.linuxfoundation.org/research/gen-ai-2024

12 Prioleau, Marc. “The Unique Challenges of Open Data Projects: Lessons From Overture Maps Foundation.” The Linux Foundation. January 13, 2025. https://www.linuxfoundation.org/blog/the-unique-challenges-of-open-data-projects-lessons-from-overture-maps-foundation

13 Bennet, Karen, Gopi Krishnan Rajbahadur, Arthit Suriyawongkul, et al. “Implementing AI Bill of Materials (AI BOM) with SPDX 3.0: A Comprehensive Guide to Creative AI and Dataset Bill of Materials.” October 2024. https://www.linuxfoundation.org/research/ai-bom

14 “Overture provides free and open map data.” Overture Maps Foundation, accessed February 14, 2025. https://overturemaps.org/?utm_source=LF&utm_id=opendatareport

15 “Open Data Sharing.” Community Data License Agreement, accessed February 28, 2025. https://cdla.dev/

16 “What is open?” Open Knowledge Foundation, accessed February 25, 2025. https://okfn.org/en/library/what-is-open/

17 West, Joel. “The economic realities of open standards: black, white, and many shades of gray.” In: Greenstein S, Stango V, eds. Standards and Public Policy. Cambridge University Press; 2006: 87-122.

Acknowledgments

The authors would like to thank the organizers of the World Open Innovation Conference for hosting a seamless and immersive conference and for incorporating this challenge session into the agenda. The session participants were a diverse and highly engaged group that brought relevant, personal, and constructive insights which make up the foundation for this report. Thanks to Hilary Carter and Henry Chesbrough for their keen review of the manuscript and to the Linux Foundation Creative Services team and Christina Oliviero for producing this report and managing its publication.

About the authors

Anna Hermansen is a Researcher and the Ecosystem Manager for Linux Foundation Research where she supports end-to-end management of the Linux Foundation's research projects. She has conducted qualitative and systematic review research in health data infrastructure and the integration of new technologies to better support data sharing in healthcare, and has presented on this research work at conferences and working groups. Her interests lie at the intersection of health informatics, precision medicine, and data sharing. She is a generalist with experience in client services, program delivery, project management, and writing for academic, corporate, and web user audiences. Prior to the Linux Foundation, she worked for two different research programs, the Blockchain Research Institute and BC Cancer's Research Institute. She received her Master of Science in Public Health and a Bachelor of Arts in International Relations, both from the University of British Columbia.

Paul Wiegmann is an Assistant Professor at Eindhoven University of Technology (TU/e), where he researches and teaches about standards and standardisation in an innovation context. His work is at the intersection of management and policy, and investigates how various stakeholders in standardisation ecosystems can shape and implement standards to support innovation and positive societal change. Paul's work has been published in outlets, such as *Research Policy,* the *Academy of Management Annals, Environmental Innovation and Societal Transitions,* and a single-authored book. Paul is the president of the European Academy for Standardisation (EURAS), and has stayed as visiting scholar at the University of California, Davis, Yonsei University, and the Technical University of Berlin. Prior to joining TU/e, Paul received PhD and MSc degrees in Innovation Management from Erasmus University Rotterdam, and a BSc in Management from the University of Warwick in the UK.

To reference this work, please cite as follows: Anna Hermansen and Paul Wiegmann, "Pathways to Open Data: Findings from the 2024 World Open Innovation Conference Challenge Session," foreword by Henry Chesbrough, The Linux Foundation, March 2025.

x.com/linuxfoundation
facebook.com/TheLinuxFoundation
linkedin.com/company/the-linux-foundation
youtube.com/user/TheLinuxFoundation
github.com/LF-Engineering

THE LINUX FOUNDATION | Research

Founded in 2021, Linux Foundation Research explores the growing scale of open source collaboration, providing insight into emerging technology trends, best practices, and the global impact of open source projects. Through leveraging project databases and networks, and a commitment to best practices in quantitative and qualitative methodologies, Linux Foundation Research is creating the go-to library for open source insights for the benefit of organizations the world over.