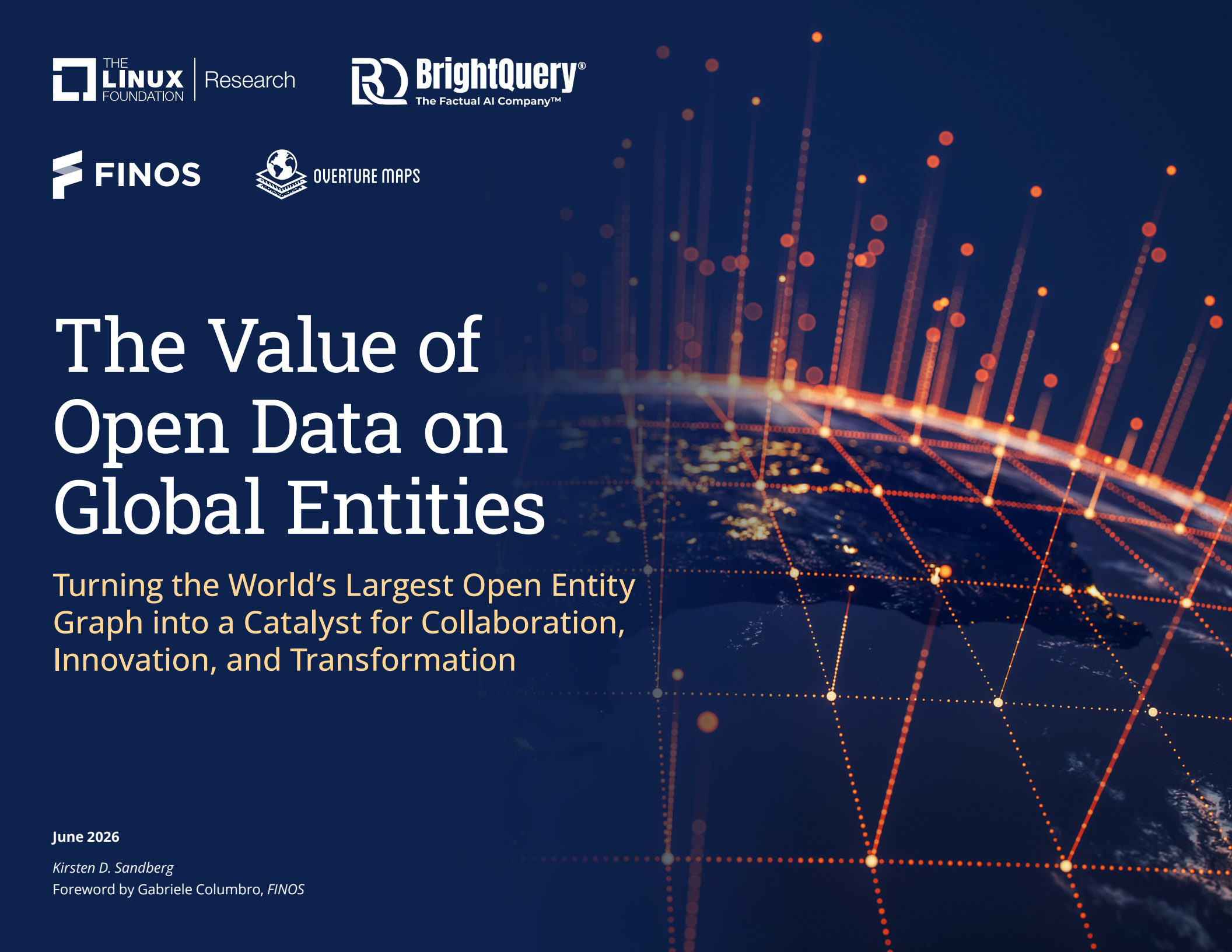
THE LINUX FOUNDATION | Research

BrightQuery® The Factual AI Company™

FINOS

OVERTURE MAPS



The Value of Open Data on Global Entities

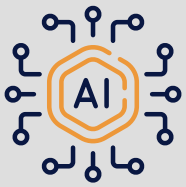
Turning the World's Largest Open Entity Graph into a Catalyst for Collaboration, Innovation, and Transformation

June 2026

Kirsten D. Sandberg
Foreword by Gabriele Columbro, *FINOS*

The Value of Open Data on Global Entities

"Open data serves as the essential **bedrock for the next generation of agentic AI**."
– Jim Zemlin, CEO, Linux Foundation.



OpenData.Org dataset covers **324 million organizations**, 1.2 billion people, and 512 million locations across 222 countries, all from 100,000+ verified public sources.

Shared open identifiers connect entities across silos to eliminate duplicate records, align data systems, and give enterprises **a single source of truth** (Gartner).

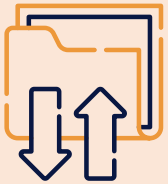
OpenData.Org's structured identifiers and metadata **ground LLMs in verified facts**, so that AI agents can reason without hallucinations and misattributions.

Financial institutions deploying agentic AI across entity resolution and due diligence pipelines could **unlock 2000% step-change productivity gains** (McKinsey).

Open entity data closes the loop on **supply chain and ESG reporting** by linking legal entities to their physical locations, supplier networks, and materials.

Organizations collectively spend billions cleaning the same core entity data. Pooling resources through open data **reduces operational costs** and improves data accuracy (McKinsey/GLEIF).

Authorized push payment fraud surpassed $1 trillion in 2023. **Shared entity resolution data** helps financial institutions verify payees, detect mule accounts, and identify anomalies (Experian).



Privacy-first by design, OpenData.Org aligns fully with GDPR, CCPA, and global privacy regulations; and governance processes honor requests to be forgotten.

The **era of proprietary, siloed entity data is ending**. OpenData.Org offers an open alternative to credit bureau identifiers and commercial firmographic vendors.

"Know your agent" is the emerging compliance frontier where canonical open entity identifiers become the **trust layer for authenticating agent identity** in cross-firms workflows.

"The opportunity for **openly licensed, high-quality, mutually maintained datasets** is simply incredible for every AI technology provider and adopter."
– Gabriele Columbro, Executive Director, FINOS

The Value of Open Data on Global Entities: The Top Use Cases

Organizations lose millions to poor data quality. **Entity resolution and master data management** link records, eliminate duplicates, and create a unified view for better decision-making.

Open data and AI streamline **trade compliance**, reduce onboarding costs, improve cross-firm identity checks, and accelerate fraud detection through shared entity graphs and automated workflows.

Open entity graphs cut bank costs, fine-tune **capital markets analysis**, and organize company relationships for risk detection, portfolio screening, and regulatory insight across open datasets.

Grounding AI agents in standardized entity data, knowledge graphs, and provenance tracking helps to decrease hallucinations, enhance accuracy and retrieval, and augment cross-domain intelligence.

Retrieval-augmented generation combines structured and unstructured information with LLMs to improve factual context-rich responses, support real-time entity resolution, and strengthen analysis across risk domains.

Multiagent infrastructures integrate entity graphs, agent-identity frameworks, zero-trust identity and access management architectures, and interoperability protocols to buttress workflows and governance models for **know-your-agent applications.**

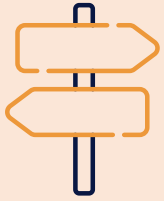
Graph-based **credit models** link corporate entities and financial data to hone risk assessments, detect distress, inform lending decisions, and strengthen credit monitoring across markets.

Single-entity views limit **customer lifetime value**. Integrated entity graphs can portray customers more fully, surface growth potential, prioritize opportunities, and deepen long-term, profitable customer relationships.

Environmental, social, and governance reporting depends on location-aware entity graphs linking suppliers, distributors, workers, materials, and funding flows to evaluate **supply chain performance**.

Enterprise intelligence draws from public web signals and alternative data to detect mergers, product launches, and strategic shifts earlier than users of traditional structured datasets.

Choosing sites for retail stores and warehouses requires demographic, mobility, and competitor location data. Open datasets reduce clustering and sharpen **infrastructure and logistics decisions**.



Geospatial and economic data guide **disaster planning, recovery, and resource allocation**. Open datasets and AI models facilitate cross-border coordination, governance, and resilient development after crises.

The Value of Open Data on Global Entities: Insights into Standing up New Open Data Projects

Open access to **foundational data accelerates innovation** and compounds the value generated through multistakeholder open data initiatives.

Successful open data projects **align incentives**, reward contributions, balance cost and access, engage vendors, and favor service-based models that sustain quality, growth, and broad participation.

Clear mission, defined scope, and strong leadership guide governance, attract contributors, and balance public benefit with private risk in open data and software projects.

Adaptable infrastructure, engaged stewards, and user feedback loops respond to evolving data needs, balance access levels, elevate quality, and sustain trusted, scalable data ecosystems.

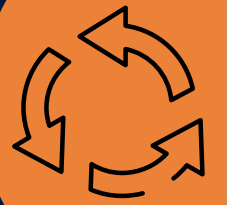
Visible governance, **active user participation**, flexible contribution models, and evolving legal safeguards drive data quality, relevance, and responsible growth in large-scale open data projects.

Traceable data origins, continual validation, active oversight, and collaborative curation maintain accuracy, as organizations **prioritize data quality over volume** and share resources to reinforce data reliability.

Successful **project teams address risks early**, define access tiers, track completeness and accuracy, apply rigorous checks, and keep contingency plans to build trust and resilience.

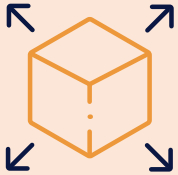
Open licensing supports global data sharing and AI training, lessens legal friction, clarifies provenance, and lets users combine datasets freely across organizations and use cases.

New data project founders must think critically about how they want the data to be used. Licensing decisions demand **scrutiny of potential downstream reuse**.



Active **community engagement drives ideas**, speeds inventiveness, builds consensus, and shapes sustainable models, as open dialogue and feedback guide decisions and strengthen collaboration.

Successful **minimum viable products** highlight unique data, fill gaps across silos, minimize technical debt, attract collaborators, and balance openness with proprietary data for scalable adoption.



Global equity in the benefits of open data requires **expanding compute access and capacity**, forming partnerships that develop skills and talent, and improving data stewardship.

Contents

Foreword

The era of proprietary, siloed data is giving way to a new paradigm of "open mutualization." As the executive director of the Fintech Open Source Foundation (FINOS), I have witnessed how open source collaboration can transform the financial services industry. However, the report before you, *The Value of Open Data on Global Entities*, illustrates that we are at the precipice of a much larger shift—one that extends far beyond software to the very foundations of the modern data-fueled, AI-supercharged global economy.

In financial services, the logic for open data is undeniable. The industry spends billions collectively to clean and maintain the exact same core entity data. By pooling resources to maintain a common "base layer" of high-quality data, institutions can decrease onboarding and trade processing costs by an estimated 10 percent. But the potential for this open global entity graph is not merely operational; it is transformational.

But, as we move toward a future where agentic cross-firm workflows redefine how business is conducted, this goes way beyond efficiency. For AI agents to act autonomously and reliably, they must be grounded in facts. Without high-quality, canonical reference data, large language models (LLMs) are prone to hallucinations and misattributions.

By providing a structured, verified directory of over 324 million organizations and 1.2 billion people, initiatives like OpenData.Org provide the "Rosetta stone" needed for AI to reason accurately across global markets. Beyond that, in an era where the very foundations of the "open web" are challenged by the dynamics of training LLMs, the opportunity for openly licensed, high-quality, mutually maintained datasets is simply incredible for every AI technology provider and adopter.

This is an opportunity that radiates outward to every data-driven industry. In retail and economic development, precisely geocoded spatial data allows entrepreneurs to optimize locations and governments to close underserved economic corridors. In cybersecurity, explicit mapping of domain and infrastructure ownership allows operators to secure network structures against upstream failures and hijacked assets. Even in the realm of environmental, social, and governance (ESG) reporting and sustainability aligned investing, this open dataset provides the missing link between entities and their physical supply chains, enabling the transparency required for true sustainability.

I encourage you to explore the use cases and incentive models detailed in this report. Whether you are training the next generation of AI models or seeking to eliminate friction in global trade, the path forward is open. Together, we can turn the world's largest open entity graph into a catalyst for global innovation.

Gabriele Columbro
Executive Director, Fintech Open Source Foundation

Executive summary

The time is ripe for breaking down data silos across organizations, industries, and supply chains and pooling entity data. Multistakeholder collaborations can provide the missing reference keyset for global entity resolution.

Open entity data project. By law, organizations must know whom they're transacting with. To help resolve the identities of entities quickly, open data projects like OpenData.Org's—a global entity graph—show the relationships among millions of entities with standardized identifiers and data transparency. This dataset draws from verified public sources, updates monthly, and offers flexible access and usage to fuel data-driven innovation.

Resources needed for project success. To sustain data quality, accessibility, coverage, fitness for purpose, and efficacy over time, such data projects require structural, legal, and financial support for governance at scale. Here BrightQuery has compiled an initial set of global entity data as a foundation for multistakeholder cooperation, with Senzing as a resolution partner to link records to real-world entities across systems.

An urgent need for open data. Open entity data helps to ground AI in verified facts. Datasets like this one address steadily growing compliance costs and give organizations a flexible and transparent alternative to proprietary solutions. As a shared resource, open datasets reduce system lock-in, improve interoperability, and steadily evolve through shared contributions across a growing data ecosystem.

Strong use cases across verticals. Among the use cases for open entity datasets are entity resolution and master data management, trade compliance, know your customer and anti-money laundering checks, and anchoring AI agents and AI models in facts for greater accuracy and speed in retrieval and response. Others cover capital markets, business credit, customer relationship management, supply chains and ESG reporting.

Cooperating on maintenance, competing on innovation. Entrepreneurs and organizations can share the costs of maintaining nondifferentiating infrastructure, which small and mid-sized businesses cannot build alone. Everyone gains stronger capabilities, better compliance tools, and improved AI quality with minimal duplicative efforts. Such "coopetition" pools resources, lowers data cleaning costs, and fosters talent, invention, and broader economic growth.

Strategic insights and market demand signals. The organizations with the most to gain from such open entity datasets are those seeking to lower the costs and complexity of trade compliance, streamline client integrations and reduce onboarding friction, build AI models on reliable reference data, and assess third-party risk and preserve supply chain integrity.

Mapping the world of entities

As early as the sixth century BCE, leaders have recognized the value of resolving the identities of people in economic life. From the Roman Census and the Book of Han to the Domesday Book and parish registries under the Tudors, civic administrations have sought to record who's who and who owns what in society.[1] The need for fraud-proof scalable solutions to entity resolution has only increased as local industrial economies have transitioned into globally digital ones. The cost of detecting and deterring fraudulent activity has risen for nearly all financial institutions, with an aggregate spend of $206.1 billion globally in 2023.[2]

The deeper problem is architectural. Early web pioneers and search engine architects organized information around documents and their uniform resource locators, and they ranked web pages largely by popularity. For three decades, this design choice—particularly its treating frequently linked pages as authoritative—has dictated information retrieval. Sir Tim Berners-Lee has warned that this design rewards clickbait and viral spread over genuine user value.[3] LLMs are more sophisticated expressions of the same mechanism. When a document refers to "ABC Inc." or "John Smith," LLMs resolve the reference by a probabilistic guess from surrounding text, usually right for well-known entities and brittle for most of the rest.

The building blocks of information are not documents but entities: organizations, people, places, governments, products, and the AI agents now acting on their behalf. Just as atoms constitute matter, entities compose information, and so a reliable foundational layer must begin there (Table 1).

Table 1: The foundational layer of information

	THE DOCUMENT WEB (LOCATORS PLUS RANKINGS)	THE ENTITY WEB (KNOWLEDGE GRAPHS)
Its unit of truth	Document container (webpage, PDF, post, etc.)	Fact/relationship (subject-predicate-object)
How users find it	Uniform resource locator (URL)	Uniform resource identifier (URI)
What indicates its value	Popularity (number of pages linking to it)	Consensus and provenance (verifiable lineage)
How AI reads it	Probabilistic guessing (AI predicts words)	Deterministic logic (AI follows exact paths)

Collaborating globally on an open entity dataset is a ripe opportunity for those developing open source solutions to such business challenges. Building out the open entity graph available at OpenData.Org is a prime example. It serves as a worldwide directory of organizations, locations, addresses, and people—the main pillars of the economy (Figure 1)—each with a unique identifier and transparent metadata, including schema, field descriptions, structural information, and data definitions.[4]

The graph covers 324 million entities in 222 countries and territories around the globe, not just in the United States (Figure 2, next page).[5] The project started as an open data initiative of BrightQuery, a Delaware corporation founded by its chief executive officer, Dr. Jose M. Plehn, who envisioned OpenData.Org as "an open source data alternative to proprietary ID systems from credit bureaus and similar vendors of firmographic and business entity data and offering the legal view and the business view of an entity."[6]

Figure 1: Main pillars of the economy in open entity graphs

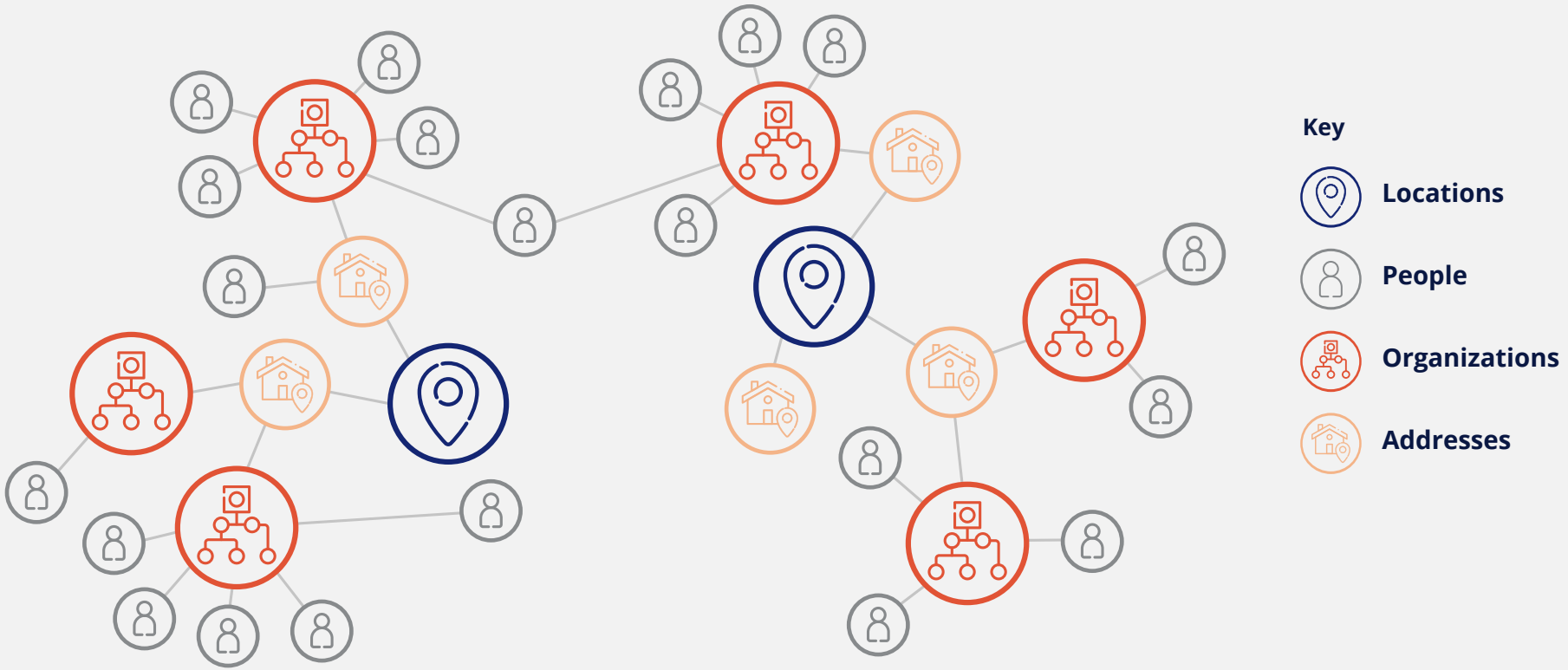
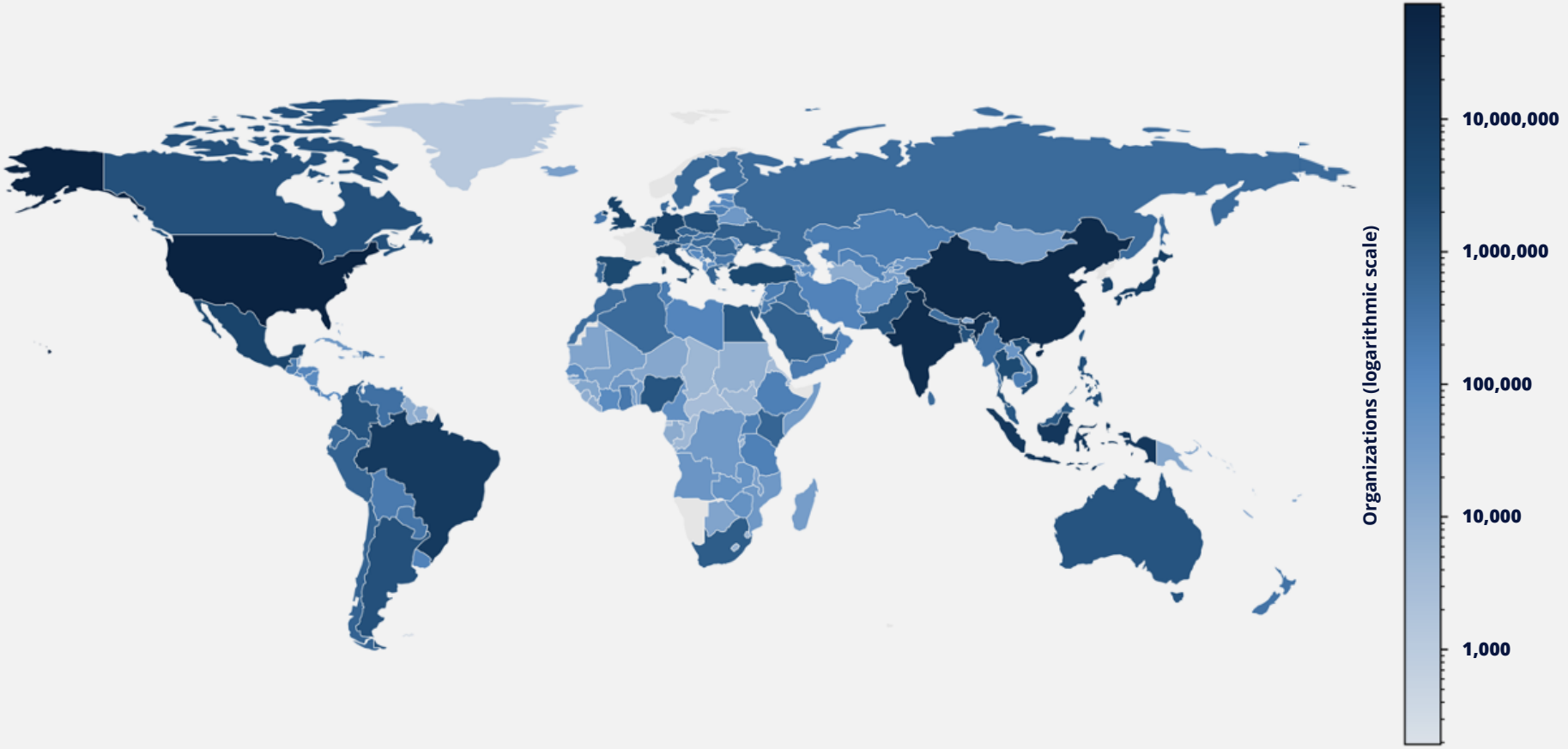


Figure 2: Global distribution of organizations by country

The open entity graph covers 324 million entities in 222 countries and territories around the world. Logarithmic scaling improves visibility across highly uneven distributions.



Source of data: Shay Hershkovitz, email to Kirsten Sandberg, 31 Mar. 2026. Map created by and used with permission from Brian Malone, who used a navy gradient to shade countries by organization count.

An innovation-ready open entity dataset

Open data assets that help to identify entities—such as Global Legal Entity Identifiers, OpenData.Org, and OpenSanctions—all help individuals and organizations to determine whether an entity is who or what it claims to be.

For example, the OpenData.Org dataset complements securities master and pricing services as well as commercially available sets of reference data and data on privately held companies so that it includes the vast majority of the core entities that financial institutions care about. Under the leadership of Dr. Plehn, whom gold members of FINOS elected to serve as their representative on the governing board, BrightQuery has curated an initial entity dataset to high standards with the following features.

100,000+ verified public sources over time: Thus far, the data comes from over 83,800 US government offices, over 32,300 US local jurisdictions, 23 US federal agencies, 51 US Secretary of State offices, and 28 Canadian agencies and registries, not from scraping private or proprietary data.[7] For greater data accuracy, the project sources entity data primarily from open government repositories.

Continuous coverage, monthly updates: The dataset covers 2017 to the present, with most data extending back to 2010, and with most data updated monthly compared with the US Census Bureau's Economic Census every five years.[8]

Privacy compliance: The dataset, data gathering methods, and data management practices fully align with General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and global privacy regulations; and governance processes honor requests to be forgotten under such "right to erasure" and "right to delete" provisions.[9] Since the data comes primarily from government sources, companies rarely reach out to delete or erase their data.

Optimal balance of privacy and utility: To protect the privacy and safety of persons and still provide organizations with a useful geographic context, the dataset includes only their country and state or province information, not their zip codes nor street addresses.

Privacy by reference, not exposure: A shared entity layer can shrink the circulation of sensitive personal data rather than expand it. Today, organizations store and exchange personally identifiable information largely to verify identity. If each entity carried a persistent, nonreversible public identifier, functioning much like a cryptographic hash, then a bank, hospital, employer, or counterparty could confirm whom they were dealing with by reference to that identifier, without storing or transmitting the underlying data that could leak in every breach. Layered with tokenization of data and selective disclosure, such a system would be a privacy contraction, not an expansion.[10]

AI-native and developer-first product: OpenData.Org set the data up to supply AI workflows, applications, and chatbots with context-rich data. "The original impetus of this project was the hallucinations, the misattributions, we were observing in conversational LLMs," said Dr. Plehn.[11] OpenData.Org's structured identifiers and metadata help AI systems to distinguish between entities with similar names. Each identifier is built to three standards:

- **uniqueness**, so that no two entities share one;
- **persistence**, so that it survives renaming, restructuring, or relocation; and
- **resolvability**, so that it retrieves the associated record.

Such identifiers create reliable links among people, organizations, and works, so that conversational LLMs like ChatGPT by OpenAI, Gemini by Google, and Claude by Anthropic can resolve entities more accurately in their responses. "Helping to eliminate hallucinations is a big motivation and a big benefit of this initiative," Dr. Plehn said. "We want AI grounded in facts."

Flexible data delivery with quarterly updates: Users can choose bulk file transfer protocol (FTP) downloads available in JavaScript object notation (JSON), Senzing, and other formats; model context protocol (MCP) servers; or application programming interface (API) integration; and eventually software development kits (SDKs).[12] With Senzing's collection of relatable data (CORD) structure, users can easily match records; and data contributors will have access to real-time APIs and MCP servers to search programmatically using natural language.[13]

Open and premium data: The project is an "open core data play," in that it has an open data core, and then BrightQuery and other contributing members can sell premium data separately on top of it. (See the appendix for details.)

BrightQuery has made its data available through Google Cloud Platform and Amazon Web Services. It looks to keep the dataset production ready, and so it has adapted its data architecture carefully to maintain flexibility without introducing disruptive, breaking changes. It also looks to mimic the templates, formats, processes, and other aspects of user engagement within the Linux Foundation ecosystem so that those familiar with those artifacts will feel right at home.

Figure 3: OpenData.Org coverage of US entities

OpenData.Org's open entity dataset, for example, covers these five pillars of the US economy.

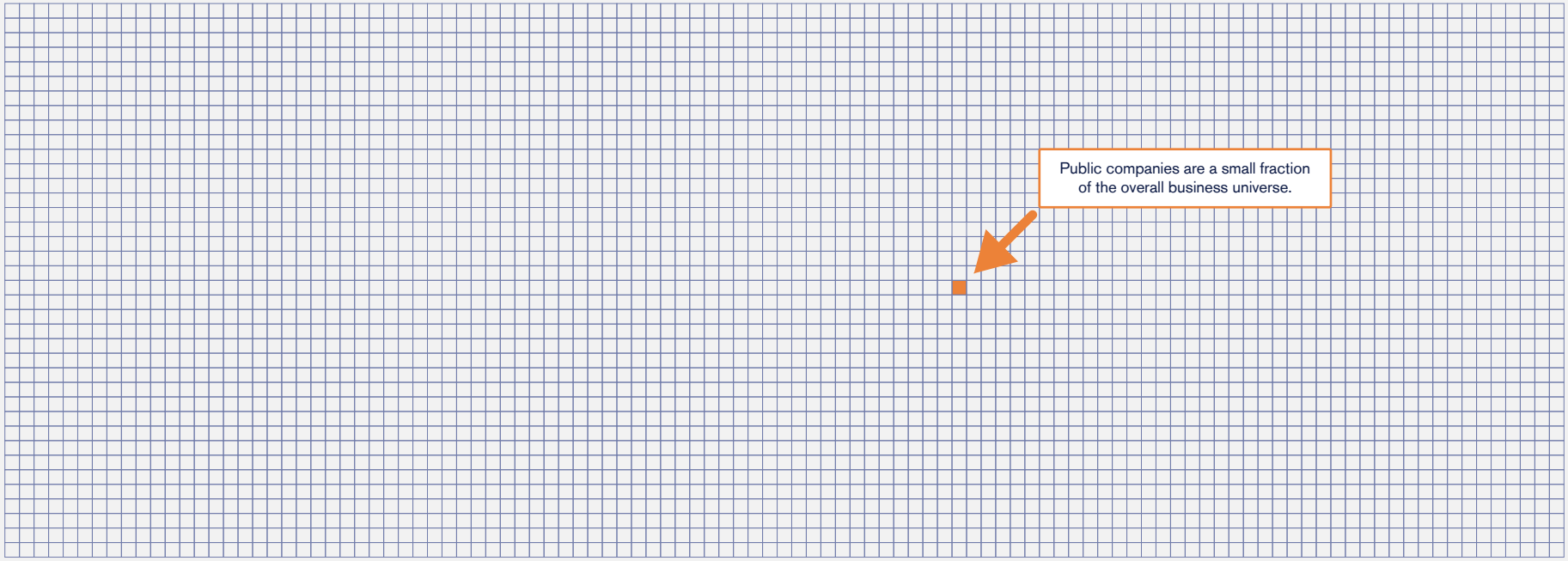
Source of data: "US coverage," BrightQuery documentation, as of 1 May 2026, https://docs.brightquery.com/dashboard.

An urgent need for open entity data

The record-breaking adoption of artificial intelligence, combined with the rising costs of complying with global financial regulations and the high costs of switching proprietary compliance solutions, underscores the urgency around standing up an open entity data project along the lines of other open data projects in the Linux Foundation's constellation of open assets.[14] The data available at OpenData.Org, for example, addresses these three issues head-on.

Figure 4: OpenData.Org coverage of US companies

The OpenData.Org dataset covers approximately 12,000 times more private companies (represented by the grid) than are listed publicly with tickers on US exchanges, represented by the orange square below.



Source of data: "Record Counts," BrightQuery Documentation, as of 1 May 2026, **https://docs.brightquery.com/record_counts.**

AI grounding

According to Linux Foundation CEO Jim Zemlin, "Open data serves as the essential bedrock for the next generation of agentic AI, providing the high-quality grounding necessary to train LLMs and eliminate hallucinations during inference. Open entity data will enable precise, real-time business entity queries across standardized datasets, accelerating the development of trustworthy and autonomous cross-firm workflows."[15]

In the World Bank's view, data aggregators, redistributors, and AI developers must adopt and apply common metadata standards, "prioritize authoritative sources, respect licenses and interpretive guidance, and anchor all AI outputs in canonical data with strong provenance safeguards" and traceable, auditable controls for risk management, regulatory compliance, and responsible AI.[16]

With such a layer of authoritative entity data, AI agents can retrieve accurate information quickly across systems within and outside an organization. The rigorous application of entity data standards increases the interoperability of agentic AI applications and autonomous agent workflows. BrightQuery believes that the network effects of data contributions from individuals and organizations across the ecosystem will continuously elevate data quality. Dr. Plehn expects the dataset to keep pace with the global economy as its catalog of reference identifiers and data sources grows.

> "Open assets compress the capital requirements of developing AI tools, but they do not eliminate execution asymmetries."

Open assets compress the capital requirements of developing AI tools, but they do not eliminate execution asymmetries. Prem Ramaswami, head of Data Commons, an open data initiative of Google that aggregates and makes public statistical macro data easily accessible, believes that an LLM plus a knowledge graph is greater than either of them individually. "If we can ground an LLM with the information in a knowledge graph—even if it's only one percent of the information in the world—then the LLM can infer the gaps to help us make better decisions," he said.[17]

Compliance costs

Complying with know-your-customer (KYC), know-your-bank or business (KYB), and anti-money laundering (AML) regulations is critical. According to LexisNexis Risk Solutions, the costs of doing so have risen for nearly all financial institutions: in Europe, the Middle East, and Africa, organizations have spent $85 billion, compared to $61 billion spent in Canada and the United States, $45 billion in the Asia-Pacific, and $15 billion in Latin America.[18]

Opaque boxes and system lock-in

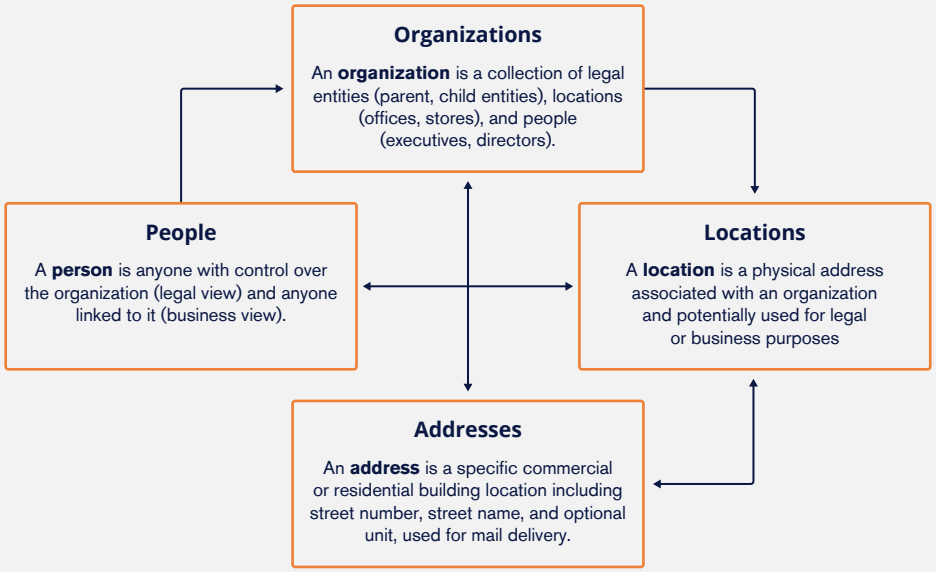
To get this degree of global entity resolution, organizations have typically subscribed to proprietary data feeds or accessed them via analytics platforms under the proprietor's terms of use or service level agreement (SLA), which tend to limit what organizations can do.[19] Some companies need a dataset that's more extensible, more modular, and more usable across their businesses without having to negotiate with the proprietor every time they use it elsewhere. Open licenses for data such as the Community Data License Agreement Permissive 2.0 (CDLA Permissive 2.0) give data adopters greater freedom to use the data and assurance that the data will remain open.[20]

Those subscription services also tend to limit what organizations can see, usually through tools or APIs. Data is in an opaque box. Users cannot see the changes in the database. Rather, they see only what the vendor is delivering at any point in time, usually based on an API query. With OpenData.Org data, for example, users can actually do a differential and see the changes in the dataset over time. Equally important, such open projects have clear participatory processes for making and implementing such changes.

Finally, these companies combine unique datasets such as firmographics, credit histories, and compliance records with persistent identifiers and then embed their workflows into their subscribers' customer relationship management (CRM) and enterprise resource planning (ERP) systems. But, unlike OpenData.Org's, their data schemas are not portable. (See Figure 5 for OpenData.Org's schema). Once integrated, the data becomes part of operations, and so switching data services disrupts operations. The contractual commitments are hard to exit. SLAs stabilize dependency and raise migration risk, and the network effects of many users turn proprietary data schema into de facto standards.[21]

"It's a really, really ripe time to look at alternate methods of collecting and publishing economic data," said Joel Gurin, president and founder of the Center for Open Data Exchange (CODE) and author of the seminal text, *Open Data Now*. "For obvious reasons, people across the spectrum are asking, 'Can we find alternatives to conventional federal data collections?' The degree to which open data projects can do that to a high standard would be very valuable."[22]

Figure 5: Open data schema and linkages



Source: "Open data schema and linkages," OpenData.Org, as of 7 May 2026, https://opendata.org/; "Universe data relationship," BrightQuery, as of 7 May 2026, https://docs.brightquery.com/universe_data_relationship_schema. Used with permission.

The forces behind the open data project

No open data project forms in a vacuum. For OpenData.Org, BrightQuery is the initial steward and provider of the dataset, which reflects BrightQuery's efforts to cover private companies as well as publicly held ones around the world. BrightQuery is also the lead government partner on the US National Secure Data Service Demonstration (NSDS-D) under the National Center for Science and Engineering Statistics and a primary supplier of government data for leading AI companies.[23] The company serves on the board of the AI Alliance and is a member of the Enterprise Data Management Council, AI Infrastructure Alliance Foundation, and Overture Maps.

BrightQuery and its collaborators propose to house this work in a neutral, independently governed body. To preserve that neutrality, BrightQuery plans to contribute its directories of organizations, people, and places, along with the validated relationships among them, as fully open source under permissive licensing and without any special governance rights for itself, to a neutral entity.

Senzing is the project's entity resolution partner. Senzing helps data platforms to determine when different records refer to the same real-world entity (i.e., a person, company, address, or other object) and then to link or merge those records correctly.[24]

To meet the pressing demand for entity data in grounding AI and controlling compliance costs with greater transparency and flexibility, the project needs a neutral home, open governance, world-class participants, and a 501(c) designation and structure. Dr. Plehn put it best in his open data manifesto: "The foundational layer of human and machine knowledge should not be owned. It should be open, governed in the public interest, and built by the institutions whose decisions depend on it."[25]

Use cases for open entity data

According to Dr. Plehn, the strongest initial use cases span four verticals—entity resolution and identification, capital markets, business credit, and sales and marketing—and AI model training cuts across them.[26]

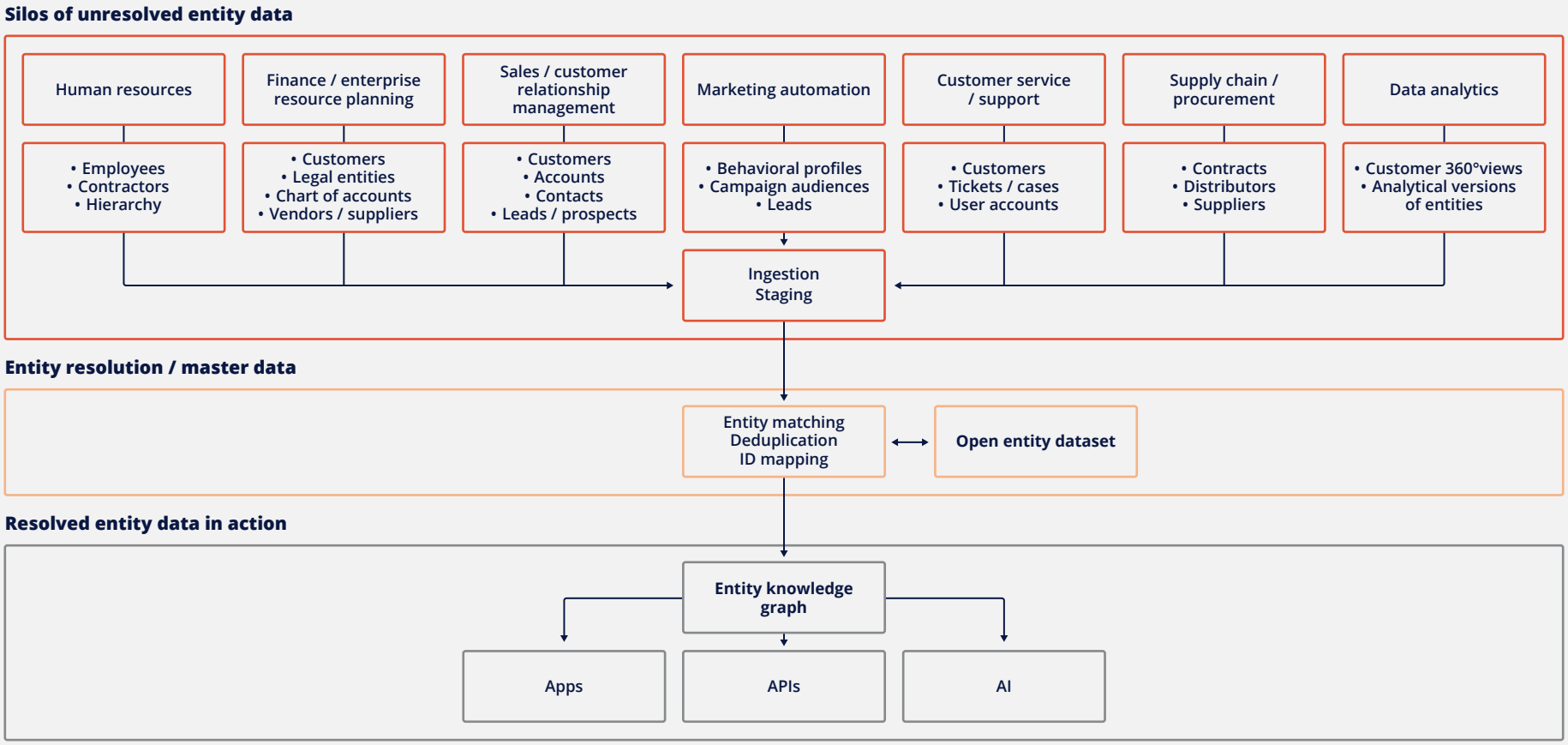
Entity resolution and master data management

When it comes to money and other valuables like data and intellectual property, entities must know which other entities they are dealing with. Gartner estimated that organizations lose $12.9 million per year because of poor data quality, including duplicate records, inaccuracies, and inconsistencies that compliance teams must remediate.[27] "Not accurately knowing who is who in your decision-making comes at a high cost," said Jeff Jonas, founder, CEO, and chief scientist of Senzing.[28] He pointed to the Paycheck Protection Program (PPP): evidence showed that some people received multiple payouts totaling $4 billion because of incomplete controls for detecting duplicates and mismatches between PPP application data and federal databases.[29]

Open entity graphs help to mitigate these issues by modeling entities and relationships explicitly and linking records from multiple sources through entity resolution workflows (Figure 6).[30] Such an approach reduces manual cleanup, aligns systems, and maintains a "single source of truth" for people and organizations, and that's what master data management (MDM) aims to achieve.[31]

"This project is essentially creating a Rosetta stone for entity resolution," said Isio Nelson, managing director of research, fraud, and thought leadership at the ProSight Financial Association. "If an institution has data silos, then it can bring in an entity identifier from one of its datasets and find the entity in its other datasets, link them together, and see everything it knows about each organization and individual. From there, leaders can see more clearly what they need to increase their productivity and to innovate."[32]

Figure 6: From silos to insights



Trade compliance, know your customer, and anti-money laundering

Complying with KYC, KYB, and AML regulations is critical as well. The McKinsey Global Institute estimated that resolving identity digitally while registering customers could decrease onboarding costs by up to 90 percent and payroll fraud by $1.6 trillion worldwide.[33]

Open data can help. For example, Companies House, the government agency responsible for maintaining a public registry of companies in the United Kingdom, launched its open data portal in 2015 "to improve corporate transparency."[34] Research has shown that the estimated total value of the registry to businesses under AML regulations was between £170 million to £460 million per year; and to law enforcement users, it generated about £2,600 per user each year.[35] In 2024, users accessed the registry 16.5 billion times, and the industry value of the open data ranged between £1 billion to £3 billion.[36]

By deploying agentic AI to aid human case handlers and investigators of potential financial crimes, financial institutions have already realized significant gains (15–20%) in productivity, according to McKinsey & Company. If they deployed a "digital factory," where AI agents collaborated across due diligence pipelines including entity resolution, and human supervisors handled exceptions, then they could unlock step-change productivity gains (200–2000%).[37]

The next frontier will be cross-firm workflows, where AI agents will assist in buying and selling securities or negotiating on a request for quote (RFQ) or a request for information (RFI). The identity of agents would be a critical element for regulated workflows, recorded conversations, reports to regulators, and so forth.[38] "In a digital economy increasingly shaped by automation and AI, reliable and traceable entity data is essential," wrote Zornitsa Manolova, head of Data Quality Management and Data Science at the Global Legal Entity Identifier Foundation (GLEIF).[39]

Table 2: AI agent driven workflows

PROCESS	HUMAN WORKFLOW (TODAY)	AI-DRIVEN WORKFLOW (TOMORROW)
Payment across firms	Manual initiation, compliance checks, routing via automated clearing house/ real-time gross settlement, and reconciliation	Smart contract and unified ledger processing, AI compliance and routing, atomic settlement
Financial trade across firms	Order placement, affirmation, clearing, settlement, reconciliation	Automated pre-trade risk, smart routing, and instant post-trade workflows with distributed ledger technology

The faster financial institutions and other global corporations can identify fraud together, the faster they can stop it and prevent losses. For example, the consumer reporting agency Experian estimated that, to reduce authorized push payment fraud, which surpassed $1 trillion in 2023, the global finance industry could leverage transactional and open entity resolution data to expedite how their financial systems verified payee identities and detected mule accounts and behavioral anomalies across institutions.[40]

Commercial implementations like the AWS Entity Resolution with Amazon Neptune Analytics show how investigators can transform standardized and matched customer and transaction data into graph structures so that they can traverse relationships and spot patterns indicative of fraud like "card not present" more efficiently than relational methods.[41] Platforms like the Poland-based company DataWalk SA likewise illustrate how unified knowledge graphs merge otherwise siloed internal and external structured and unstructured data into 360-degree views of entities that facilitate AML, fraud triage, and regulatory reporting.[42]

In the open domain, the Enterprise Data Management Council used global legal entity reference data to develop a prototype of an AML knowledge graph that improved the effectiveness of its AML and KYC processes, decreased the cost of manual work, and increased detection accuracy—all by connecting entities across datasets.[43] For example, with OpenData.Org data, users can customize knowledge graphs for their compliance needs. Innovators are also using transformers to sharpen entity resolution further.[44]

Capital markets

An open global entity knowledge graph could help banks cut operational costs. For example, according to McKinsey & Company, if banks adopted legal entity identifiers, then they could decrease the costs of processing trades by at least 10 percent. That translates into an annual savings of over $150 million in the investment banking industry and $500 million for banks issuing letters of credit to finance trade.[45]

Capital markets participants can use open business entity graphs to organize public and private company data in structured, interconnected representations that help them detect anomalies, enhance their competitive analyses, and hone their investment and risk management strategies. Unlike tabular datasets, such graphs model the companies, their financials, ownership hierarchies, and relationships so that firms can do deeper analytics.

Consider the CompanyKG dataset, an open, large-scale heterogeneous graph of over 1.169 million real-world companies and 50.8 million weighted intercompany relationships published under an MIT license by researchers in Sweden.[46] Its users can more easily quantify the similarities and risks of a set of competitors and can screen portfolios in investment workflows.[47] Another open resource, FinReflectKG, offers a financial knowledge graph extracted from US SEC 10-K filings for S&P 100 companies, with structured triplets linking entities, relationships, and temporal facts for tracing risk exposures or detecting abnormal patterns in financial disclosures.[48]

In the broader open data domain, Wikidata's knowledge graph includes 120.7 million entities and their attributes under a public domain dedication (CC0 1.0 Universal), so that analysts can enrich firmographics with consistent identifiers and linkages to external datasets such as sector codes or geolocations.[49] Beyond static graph representations, implementations like Agentic GraphRAG combine structured entity relationships with unstructured filings (e.g., 10-K documents).[50] Along with an open entity dataset, these help capital markets analysts to answer complex questions about regulatory risks, supply chain dependencies, and counterparty exposures in the S&P 100 and to fuel their investment dashboards and anomaly scoring models without proprietary licensing constraints.

Agents and AI model training

"After many years of encouraging collaboration among diverse groups of humans to build federated cloud and data management systems, we're now bringing AI systems to the stakeholders' table," wrote Heidi Picher Dempsey, US research director of Red Hat LLC, in her article on what scientists and practitioners have learned from such collaborations as the Mass Open Cloud and the Large Hadron Collider.[51] Modern AI systems rely on high-quality data to reason reliably and interoperate globally. These use cases illustrate how layered architectures—spanning entity data sources, extraction frameworks, entity resolution, knowledge graphs, provenance tracking, AI models, and agent infrastructure—promote trustworthy, auditable, and adaptive intelligence for analytical and operational workflows.

Trusted AI grounding with canonical reference data. To reduce hallucinations and improve factual accuracy of results, developers must ground AI systems in reliable knowledge.[52] Without such grounding, an LLM might guess whether two names refer to the same entity in generating an answer. With grounding, the model can retrieve the authoritative entity record and relationships, as when a bank's AI compliance assistant gets verified records from sources such as Office of Foreign Assets Control sanctions lists or the Financial Crimes Enforcement Network database, before generating an answer.

If a company stores its product names, components, firmware versions, and documentation in structured repositories such as product catalogs and engineering knowledge bases, then it can ground an AI service agent with those canonical product entities and documentation graphs and enhance the accuracy, utility, and customer experience of those agents.[53]

Retrieval-augmented generation (RAG). RAG is another well studied architectural pattern that couples structured external knowledge with LLMs to deliver more fact-based, context-aware responses.[54] For example, BrightQuery has developed what it calls an "agentic hybrid RAG factual AI platform" that fuses structured and unstructured information and then applies AI models to interpret relationships among entities, detect anomalies, and generate insights that inform decision-making in business and government domains.[55] The platform uses RAG techniques alongside agent-based workflows, and its architecture is modular and open source so that it can scale and adapt across industries. With MCP server access to the entity graph, AI agents can resolve entities in real time and reduce hallucinations.

The platform operates by ingesting and aligning diverse datasets in a unified framework. In capital markets, it blends financial filings with real-time news and opinions to assist forecasters, risk analysts, and compliance officers. In government settings, it merges economic data with public sentiment to inform policy analyses and national security decisions. It delivers several clear benefits to users: higher quality decisions through deeper context and corroborated insights, accelerated analysis across complex data environments, and increased trust through systematic bias reduction. The platform's flexible design accommodates evolving data needs as it maintains consistent performance across sectors.

Standardized entities with global coverage for AI agents. To coordinate reasoning across services and workflows around the world, AI agents depend on consistent semantic representations of entities. Knowledge graphs are a formal model for encoding entities and relationships as structured semantic triples—the basic unit of knowledge representation—so that they can share semantics and reasoning across tools and services.[56] According to the Boston Consulting Group, AI agents could boost banks' profitability by 30 percent and reduce costs by up to 40 percent by 2030.[57]

Data hygiene network effects with ecosystem contributions. Clean, standardized entity data improves AI performance and the reliability of automated decisions. Organizations apply entity resolution and knowledge graphs to remove duplicates, link records about the same real-world entities, and structure relationships. When members of an ecosystem contribute new records and make corrections, they cultivate the dataset over time, and these network effects strengthen training data and downstream AI applications.[58]

Governance, risk, and compliance controls that are traceable and auditable. For trustworthiness, AI models must trace every data transformation, and AI agents must perform as well as or better than human beings against benchmarks.[59] Data provenance and lineage systems document the origins of and changes to data so that stakeholders can audit, verify compliance, and reproduce insights as entity data flows from source through resolution and inference pipelines.[60]

Efficient retrieval of information with authoritative entity resolution. Knowledge graphs unify structured representations of entities and their semantic relationships. When developers feed these graphs such resolved, canonical entity identities, the graphs can do precise contextual retrieval and rich reasoning over relationships, thereby improving the efficiency of searches and the extraction of insights.[61]

Adaptive and emerging use cases such as know your agent. Advanced multiagent infrastructures integrate canonical entity representations with agent-identity frameworks, zero-trust agentic identity and access management architectures, and interoperability protocols such as layered orchestration for knowledgeful agents (LOKA), Agent2Agent, and MCP. With these technologies and knowledge-graph memory layers, interoperable AI agents could authenticate identity, discover tools, coordinate actions, and reason over structured entities to shore up adaptive workflows and governance models for emerging use cases such as know your agent (KYA).

"A lot of people who create AI agents don't necessarily want them to be known as agents. Other LLM creators have scraped different sites, and they purposely obfuscated where the traffic was coming from," said Ramaswami of Google's Data Commons. He described how website owners with large numbers of web pages get a huge charge in one random month because "they're getting hit by all these IP addresses from a certain country or a certain LLM creator," he said. "I would not be surprised if the creators of those agents designed them to masquerade as humans."[62]

Business credit

On the business credit side, graph-based methodologies like AWS Labs' SageMaker credit scoring tool (Apache 2.0 license) show how linking entities via corporate network graphs constructed from SEC 10-K/Q filings with financial ratios can enhance predictive credit rating models.[63] Open entity datasets such as OpenData.Org's supply foundational attributes like incorporation details and director information that users can link to credit monitoring workflows and detect early signs of distress or fraud across related businesses.[64]

According to Isio Nelson, Prosight's members—largely banks, credit unions, and financial services professionals—want "better information on the companies they serve or want to serve so that they can do their jobs better, expand their existing relationships, understand the risk of their existing portfolios better, and put their resources in the right places."[65] By combining open knowledge graph datasets, open graph databases, and open AI techniques, these implementations illustrate how open business entity graphs can strengthen capital raising assessments, risk management, and real-time credit risk monitoring in both public markets and private credit environments.[66]

Sales, marketing, and customer relationship management

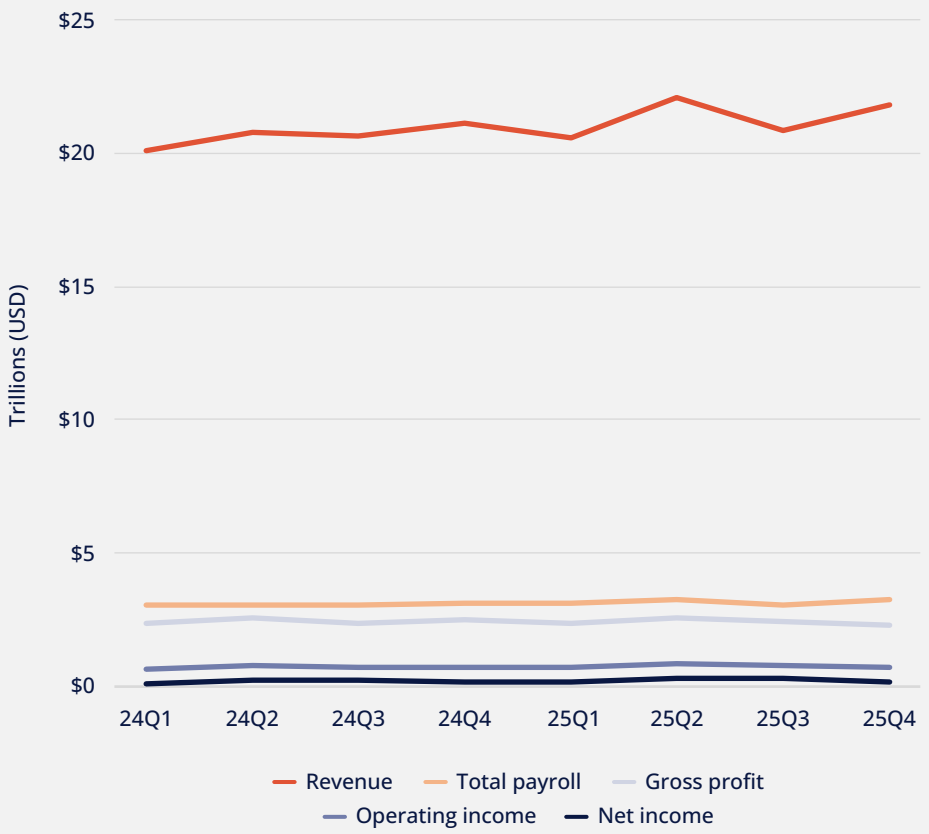
Researchers found that financial institutions often relied on single-entity data to calculate customer lifetime value (CLV). If they used open banking data and analyzed customers' cross-firm behavior, then they could develop customer-centric strategies that boosted per-customer profitability, nurtured long-term relationships, and ultimately increased CLV by over 20 percent.[67]

Nelson of Prosight said that, with all the other data the project is attaching to entities, plus the lineage and the usage rights of the data, "it might help organizations to backfill some of their customer records so that they can better understand the customers they already have." For example, "How many of our current customers own a small business or are a principal at a commercial entity? Are they sole proprietorships? Businesses with employees? What are their cash flows?" Such data would help to prioritize opportunities, Nelson said. "Of these customers, these are the top ten with the highest growth profiles. Or these are the underperforming accounts with the most potential."[68]

By combining entity resolution, relationships, and risk indicators from open data sources into an integrated graph, sales and marketing teams can augment customer profiles and uncover golden opportunities or risks before they reach out to prospects. For instance, organizations could combine the OpenData.Org dataset with the Wikidata knowledge graph—a collaboratively curated global graph of entities (public domain CC0 license)—as a starting dataset of companies, individuals, and attributes that they could query via SPARQL to enrich and segment their prospect records.[69] Likewise, they could add sanctions and risk-related data from OpenSanctions and filter out high-risk or compliance-sensitive targets.[70]

Figure 7: Companies with full-time employees are growing

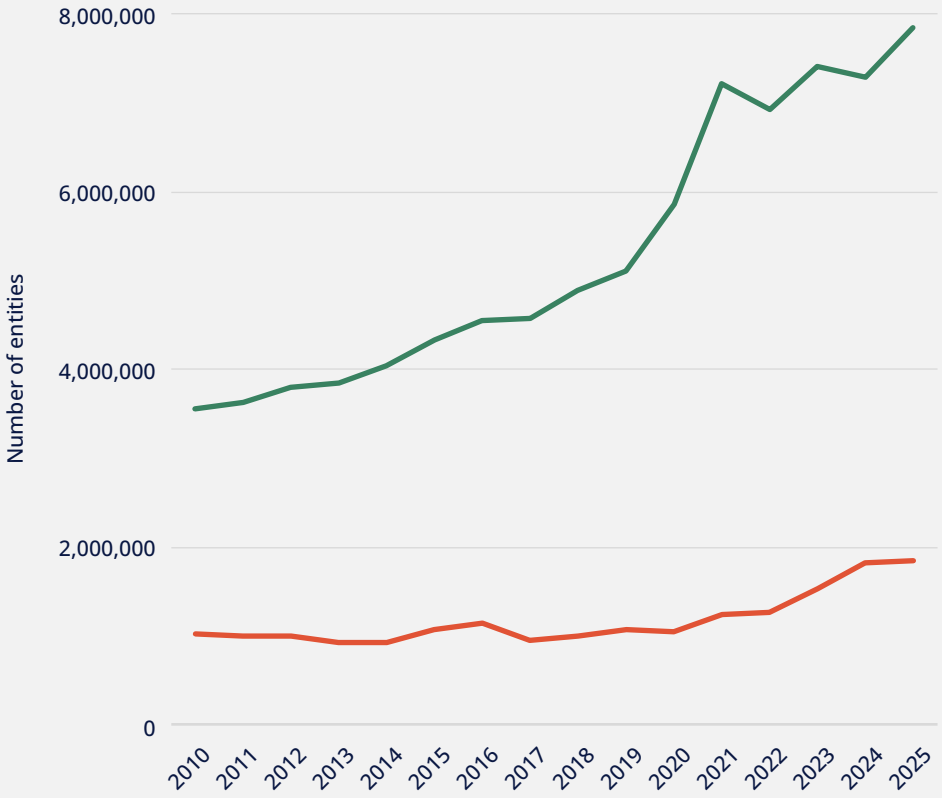
An analysis of ~8 million US small businesses showed growth in revenue (6.43%), payroll (6.24%), and employment (0.64%).



Source of data: BrightQuery Inc. and ProSight Financial Association, "Companies with full-time employment," *BrightQuery–ProSight Economic Analysis Report*, as of 1 Jan. 2026, https://docs.brightquery.com/bq-prosight-economic-analysis-report.

Figure 8: Tracking formations and dissolutions of legal entities, United States

Sales representatives could analyze new business formation by zip code or building address within their territories. The green line indicates new business formations, and the red line indicates dissolutions.



Source of data: BrightQuery Inc. and ProSight Financial Association, "Annual legal entity new formations and dissolutions," *BrightQuery–ProSight Economic Analysis Report*, as of 1 Jan. 2026, **https://docs.brightquery.com/bq-prosight-economic-analysis-report.**

"Platform users could personalize outreach campaigns to no- or low-risk prospects."

Open graph databases such as TerminusDB and JanusGraph give users practical ways to store and query these integrated entity graphs so that they can discover secondary contacts, common affiliations, or shared risk factors among corporate networks.[71] For example, a marketing analytics platform built on these technologies could fuse Wikidata identifiers, OpenSanctions risk flags, BrightQuery unique entity IDs, and customer IDs from a company's CRM system to get a risk-aware 360 degree view of prospects. Platform users could personalize outreach campaigns to no- or low-risk prospects.[72]

These open knowledge graph architectures also backstop federated AI workflows—via natural language query layers or embedding models, for instance—so that sales teams could ask high-level questions (e.g., "show all contacts connected to industry X who are not on any sanctions lists") and gain insights into campaign segmentation and compliance.[73]

Supply chains and ESG reporting

One of the biggest missing pieces of any kind of environmental, social, and governance (ESG) effort is location. "You're trying to put together this knowledge graph of the relationship among these entities, their suppliers, their distributors, the people who work for them, and the location of their materials," all to understand how the resulting supply chain affects the environmental landscape.[74] This open data project is one way to bring all those pieces of information together. "That's really powerful," said Dr. Amy Rose, chief technology officer of the Overture Maps Foundation, "a collaborative open data initiative launched in 2022 and led by software developers, data experts, cartographic engineers, and product managers from dozens of Overture Maps Foundation member companies."[75] "Anything having to do with supply chains is a huge one," she said.[76]

"To analyze a company's performance across ESG factors, investors first need to unambiguously identify the entity in question," wrote Richard Robinson, chief strategist, open data and standards at Bloomberg LP.[77] That may require reconciling and mapping the different identifiers used to track a single legal entity, and that's what open data standards such as LEI and FIGI and the open entity graph of OpenData.Org help to do. Otherwise, analyzing the ESG performance of a specific entity over time "can be exceedingly challenging for investors or regulators."[78]

Enterprise intelligence

Deriving enterprise intelligence from public web signals and alternative data was another theme.[79] Analysts can infer strategic developments—such as mergers, product launches, or shifts in corporate direction—by examining patterns in publicly available web data such as Common Crawl's.[80] For example, changes in website content, job postings, patent filings, or corporate announcements can reveal early signals of acquisitions or new initiatives.[81] From automotive and financial sector analyses to pharmaceutical pricing, different industries have tested such approaches. This type of web-derived intelligence complements or sometimes outpaces traditional structured business datasets because web signals often appear earlier and in greater detail.

Location of retail outlets and warehouses

For bricks-and-mortar retail stores and franchises, location is everything. Pinpointing the optimal spot requires data on the demographic density and median household income of the target audience during hours of business, walkability or drivability score, and data on complementary and competitive businesses nearby.

"Let's say I'm opening a pizza franchise. I'm trying to figure out where I should locate it. To do that, I need answers to, 'Where are all the other pizza serving places?' and 'What are the demographics of the area I'm targeting?' I can get demographic data for free from the US Census. But where do I get the locations of all the other pizza joints?" asked Ramaswami of Data Commons.[82] Without that information, an entrepreneur will likely open the new pizza joint near existing pizza joints. "That's why we get four gas stations next to each other," he said.[83]

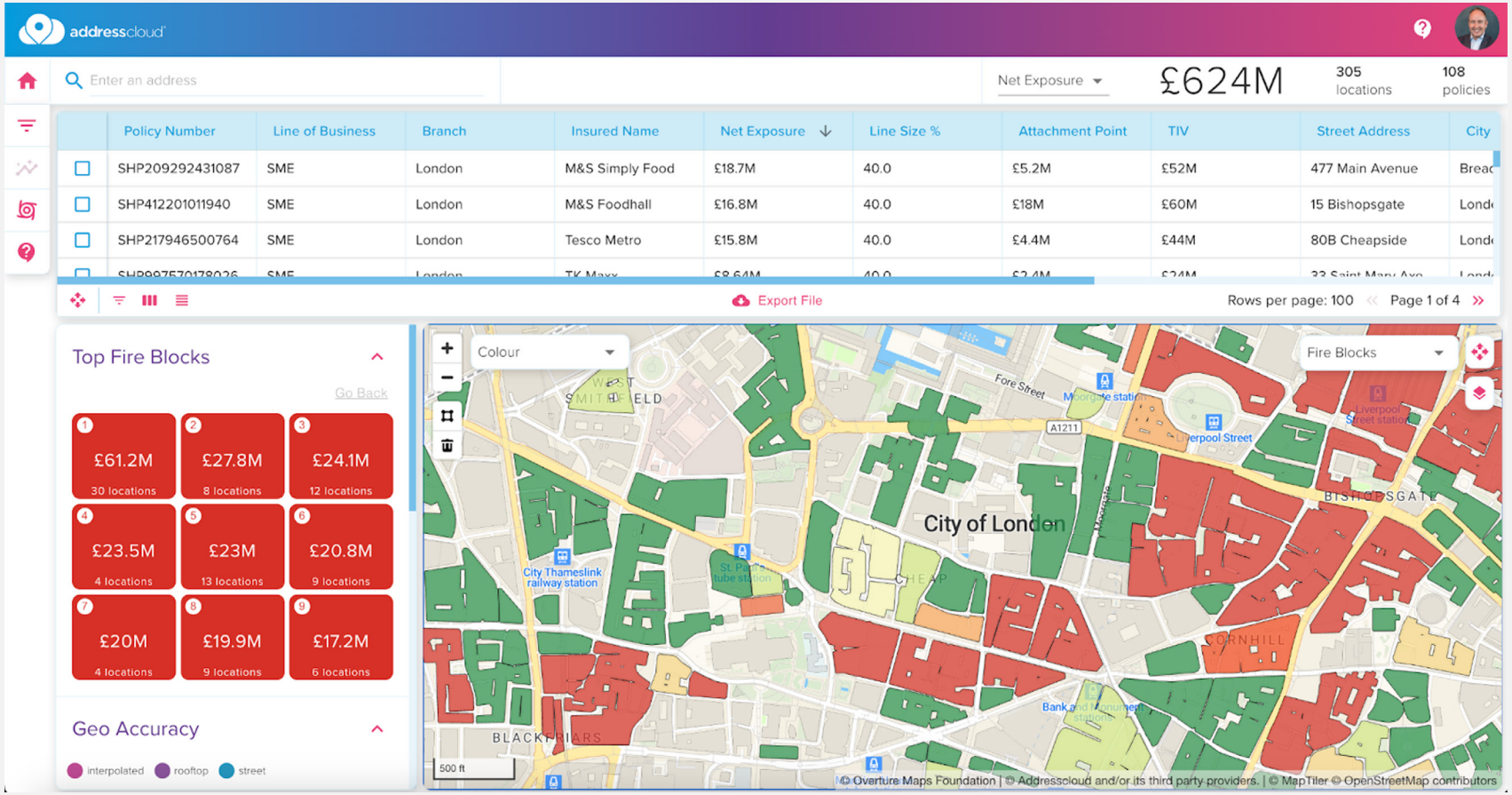
Web standards pioneer Dr. Ramanathan V. Guha described a similar challenge: the nonprofit Feeding America collects excess stock from supermarkets and redistributes it to its network of food pantries. "It runs the back end of roughly half the food pantries in this country," Dr. Guha said.[84] To determine where to locate its warehouses, it needed data on transportation systems and trends in levels of poverty and population across its country-wide network.

They see an opportunity to make much better decisions using open datasets like OpenData.Org's: "With more open data, we'll create more efficient systems, and those more efficient systems will help us all in the long run," said Ramaswami.[85]

Disaster risk management and economic development

Dr. Rose of Overture Maps described how the combination of geospatial and economic development data could help to identify overserved and underserved corridors.[86] Governments could use data and AI not only in drafting and deploying disaster risk management and response plans but in adjusting them with precision as conditions change.[87] After natural disasters or wars, communities could explore not simply how to recover but how to leverage the relationships among their industries and locations to mitigate risks, allocate public resources more consequentially, form more effective public-private partnerships, and close economic deficits. See Figure 9 for an example from Addresscloud.

Figure 9: The Addresscloud Risk platform showing fire accumulation blocks based on Overture Maps data



Source: "How Addresscloud Scaled Geospatial Risk Analysis with GERS," Case Study, Overture Maps Foundation, 10 Sep. 2024, https://overturemaps.org/case-study/2024/how-addresscloud-scaled-geospatial-risk-analysis-with-gers/. Used with Addresscloud's permission.

"What we see very valuable is cross border problem solving," said Anusha Dandapani, chief of the AI Hub of the United Nations International Computing Centre (UNICC). "Sharing open datasets makes our interoperability very easy and strengthens our governance to start with. But we don't need more open data or more knowledge graphs per se. We need open data and open AI models." She explained, "Open data plus open models will have a multiplier effect in cross border capacity building, especially in the Global South, if we consider them together in government and public sector contexts, if the open data propagates proportionally, and if we have clear policies for using and reusing sovereign datasets."[88]

Reasons for participating in the project

Individuals and organizations have several reasons for using and contributing to open data projects new to the Linux Foundation. "In capitalism, information is supposed to be free and moving without any form of friction," said Ramaswami of Data Commons. "Information is supposed to be what creates the most free markets. But we know that's not the case."

Ramaswami pointed to the advantages that come with an organization's scale, size, and ability to pay, and the cost barriers to small businesses, startups, and entrepreneurs. "To some extent, this open data project evens the playing field, capitalistically speaking," he said. "It also opens up opportunities for businesses that probably couldn't flourish yesterday."[89]

Stronger organizational capabilities

"In financial services—and I think it's very similar in other sectors—people want to engage with an initiative because it's transformational, it's strategic, or it's lowering costs," said Jane Gavronsky, chief operating officer of the FINOS, which promotes open innovation among its members and users in the financial services industry across 50 projects and working groups.[90]

Maintaining large high-quality datasets and a robust data infrastructure can weigh down small and midsized organizations. Participating in the OpenData.Org project, for example, could help *ambitious midmarket and smaller organizations* such as community banks to advance their data practices, access high-quality datasets, leverage the fire power of collaboration, and engage with regulators, especially in jurisdictions where the public sector views open data and open government initiatives quite favorably. To the extent that the project can work with global regulators, those regulators would also benefit from this type of dataset.

Positive financial impact

To contribute to projects like this one, organizations need to see a lot of potential for revenues or savings. "Data in general hits both of those quite strongly, but the messaging must clarify whether it's operational or transformational or both," Gavronsky said. "The target audience for this project are the people who understand that open collaboration is good, the product is useful to them, and they're going to use it, contribute to it, and proselytize others. That's what FINOS does. We help to build community."[91]

Greater access

"In the nonprofit context, people don't generally have access to enterprise grade entity resolution software, and nobody has much of a budget to run technical infrastructure," said Friedrich Lindenberg, founder of the OpenSanctions project, an "international database of persons and companies of political, criminal, or economic interest."[92] He sees the OpenData.Org project as a resource for the academic and civil society space, where people want to do more with the useful data they collect but have few resources to do so. "Much of the data world is massively feudal, with fiefdoms of data, each with its own way of naming things. Now anyone with a CRM can use OpenData.Org for enrichment. Companies, startups, and firms with smaller budgets can make great use of this open dataset," he said.

Greater economic growth

Studies have shown that open and spatial data deliver large but underrealized economic value: the consultancy Lateral Economics estimated that G20 countries could gain about $2.6 trillion annually from open data, while the tech firm Spatineo found that Finland had captured under 25 percent of its €13 billion potential; expanded use of national spatial data systems offered additional direct and indirect benefits of substance.[93]

In a more recent study, the McKinsey Global Institute estimated that, by broadly adopting open financial data ecosystems, high-income blocs such as the European Union, the United Kingdom, and the United States could boost their economies by as much as 1.5 percent of GDP in 2030, and lower-middle income countries like India could see a lift of four to five percent.[94]

Focused innovation and differentiation

"Open data allows firms and providers to focus on creating and competing on value-add activities," said Robinson of Bloomberg LP. He told *data.world*, "If something is not open—even if it is a 'standard'—and it is mandated, it creates lock-in, which stifles innovation." He said that "an open data foundation" fosters competition "by forcing firms to compete on quality and breadth of services, not on footprint."[95]

That will be critical in competing on artificial intelligence. The key to AI differentiation is data. The commoditization of frontier AI models is pushing innovation to the top of the stack of AI. WIth datasets like this one, which is nondifferentiating and therefore noncompetitive, companies can mutualize those costs and invest instead in training their models, creating AI agents, and building other solutions. That's where the real competitive differentiation will happen.

As an example, Gurin of CODE pointed to the US Department of Transportation's work with the major vehicle manufacturers around the data needs for automated vehicle (AV) development.[96] CODE facilitated their discussions, which resulted in the Work Zone Data initiative for automated vehicle safety and supported the broader Data for AV Integration initiative "to increase access to data for AV integration and lead to actionable priorities and clear roles in implementation."[97] It was a model of coopetition, where businesses cooperated to "provide data and develop standards in the interest of enabling the government safety initiatives that were essential to their business."[98]

The potential of pooled resources

"Collecting data costs money. Making it usable costs money," said Dr. R.V. Guha, creator of such widely used web standards as RSS, RDF, and Schema.org. "For people to use a dataset, you need a critical mass of data that is easily usable. For people to contribute their data so that the dataset reaches a critical mass, they must believe that the entity running it is neutral and will remain neutral," he said.[99]

Gavronsky added to that idea. "Yes, for any project, we need large institutions that would normally do the project on their own," she said. "They're going to invest in the project because they need it, but it's not their competitive edge." If their investment is open and available to everybody, then it could prove valuable to the market, and they benefit from that indirectly. She explained, "If we all have better KYC data, then everybody benefits. We can avoid the costs of cleaning and using the data in house, and we get better data faster."[100]

How you can help

A new project community typically focuses on key areas such as data quality, dataset adoption, and emerging use cases. For example, initial working groups can outline processes for the following:

Contributing open data to the project. Dr. Plehn believes in the power of collaboration and the knowledge of the crowd. "The data world never has 100 percent accuracy. I'd be very happy if we reached 80 percent accuracy," he said. "But if we engaged the collective knowledge and wisdom of all the enterprises out there, then I think that, together, we could reach pretty darn close to 100 percent."[101] The BrightQuery team envisions an open community where everyone can make updates and fix errors related to websites, LinkedIn accounts, phone numbers, addresses, and so forth.

Developing mapping and testing tools with transparent/drop-in replacement capabilities available so that contributors can map to existing data standards and IDs and reduce the friction of swapping an existing dataset with this open dataset.

"If their investment is open and available to everybody, then it could prove valuable to the market, and they benefit from that indirectly."

Expanding the entity graph. In an open data ecosystem, contributors can add the entity types (e.g., publications, securities, etc.) most meaningful to their work. For example, the academic and literary worlds have authors, copyrights, and publications. The pharmaceutical industry has patents, product lines, batches, and ingredients. Even construction and energy companies, shipping and airline operations, and global economic development agencies like the United Nations Development Programme have approaches to identifying and tracking their considerable physical assets.

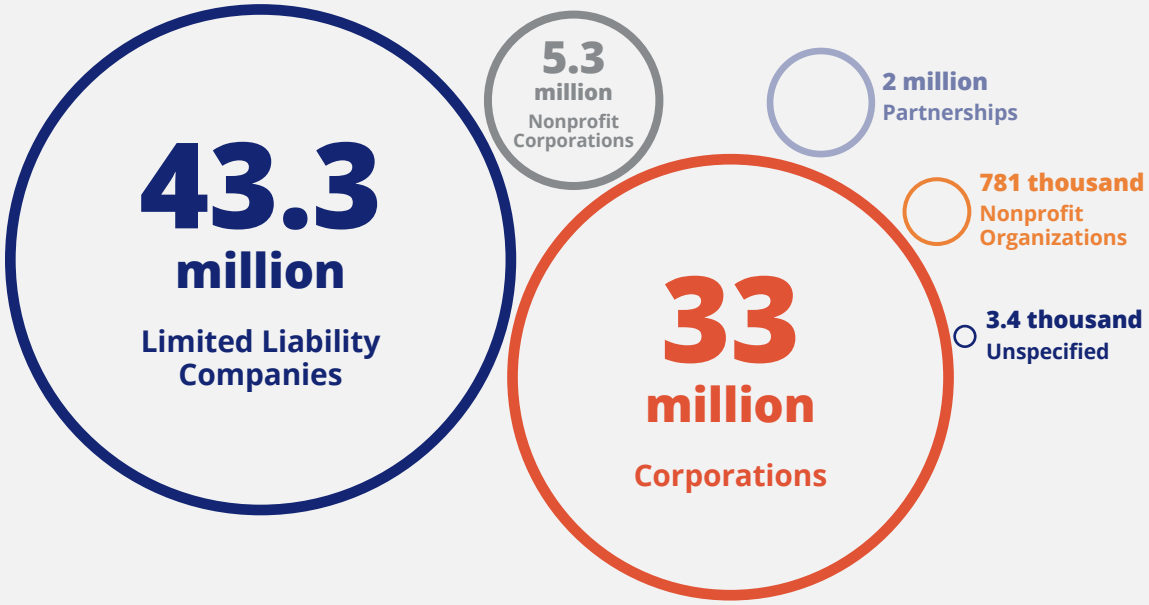
Securities data. OpenData.Org already combines entity-level information with securities and capital information captured in structured regulatory and corporate filings so that users can attach real-time securities data. Securities data would be the most obvious addition. Dr. Frank Nagle, a research scientist at the Initiative on the Digital Economy at the Massachusetts Institute of Technology, thought that the annual (Form 10-K) and quarterly (Form 10-Q) reports, filed with and made publicly available through the US Securities and Exchange Commission (SEC), would be useful.[102] Dr. Nagle wanted to see the relationships among these entities—the financial relationships, ownership relationships, and customer-supplier relationships—as well as board members and seat-level connections.[103]

Sustainability data. Also missing is consolidated sustainability information, especially in terms of the European Union's corporate sustainability reporting directive (CSRD) and other sustainability and climate-focused reporting initiatives in Australia, Canada, China, India, Indonesia, Japan, Korea, Singapore, South Africa, Switzerland, the United Kingdom, and Viet Nam.[104]

Nonprofit data. Ramaswami, head of Google's Data Commons, pointed to the inefficiencies in the market for nonprofit data and suggested including information on charitable and nongovernmental organizations.[105] As of this writing, OpenData.Org covered over six million nonprofits (Figure 10).

Finding out more. To download the dataset, visit https://opendata.org/download. To discuss being part of an open data consortium, please email partners@opendata.org. The next section of this paper walks through the process of standing up open data projects.

Figure 10: OpenData.Org coverage of different types of organizations



Source of data: "Organization counts," BrightQuery Documentation, as of 1 May 2026. https://docs.brightquery.com/org_summary.

Standing up a new data project

Realizing the value of open data

"Why is the Linux kernel so important as a piece of software?" asked Michael Dolan, senior vice president of legal and strategic programs at the Linux Foundation. "Because it's openly available to anyone at any time to instantaneously build things on a proven operating system most of the world's software supports."[106] For example, Google based its Android operating system (OS) on an upstream Linux long-term supported kernel, and NASA put a Linux OS on its small helicopter Ingenuity, which accompanied the Perseverance rover to Mars.[107]

"Having access to a fundamental layer of the technology stack under an open license, so that you can instantaneously start building something, is a huge force multiplier," Dolan said. "At a macro level, sharing technology openly through open source drives both updates for existing software and new feature innovation. The same principle applies to data, which, when shared and improved openly, can enable continuous updates and new use cases."[108]

Figure 11: Open source reference architecture: From data to insight

Organizations can tap open source solutions in creating an entity resolution–knowledge graph system for a variety of use cases, with data as a base layer of the stack.

ANALYTICS	**Apache Superset, Linkurious, Metabase, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
APPLICATIONS	**Cytoscape, Gephi, Jupyter, Streamlit, etc.**
Tools use entity data to power workflows such as risk scoring, compliance, and customer intelligence.	
SERVING / APIS	**GraphQL (Ariadne, Graphene, Hasura), REST, etc.**
Application programing interface services expose resolved identities through queryable interfaces.	
STORAGE / QUERY	**JanusGraph, TerminusDB, ArangoDB, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
ENTITY GRAPH	**Python, Apache Spark, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
ENTITY RESOLUTION	**Splink, Dedupe, Febrl, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
NLP EXTRACTION	**spaCy, Hugging Face transformers, Stanford CoreNLP, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
INGESTION / ORCHESTRATION	**Apache Airflow, Apache NiFi, Dagster, Prefect, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	
DATA SOURCES	**OpenCorporates, OpenData, OpenSanctions, etc.**
Users gain insights from tools that analyze relationships and patterns to identify trends and anomalies.	

Open datasets are another base layer of the stack, as are open artificial intelligence (AI) models and agentic AI; and the quality of the former is critical to improving the quality of the latter. (See Figure 11, previous page.) They are all precompetitive components of the products and services that the world's fiercest competitors and demanding innovators—companies, nonprofits, government agencies, and academic institutions—have created to serve their constituents better. Many of those same organizations collaborate to maintain and strengthen these base layers.[109]

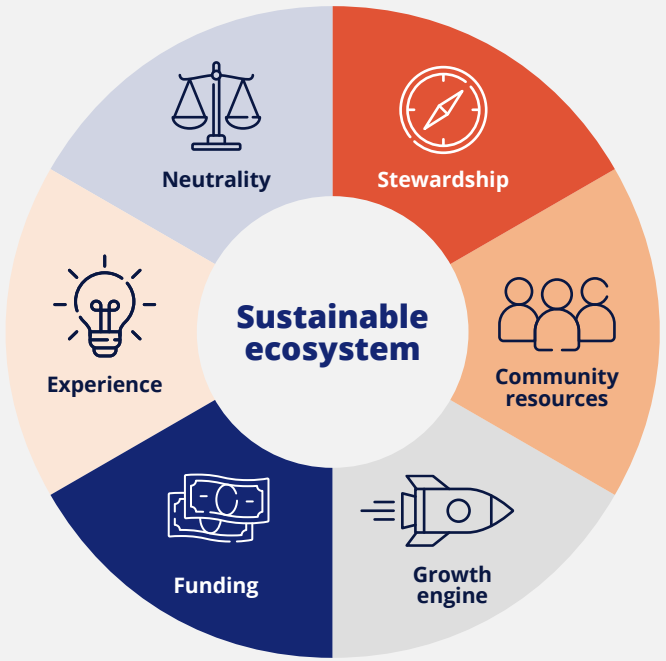
Existing open data sets are not enough for many commercial use cases. "There's a lot of open data out there from scientific research, but it's not generally data needed in a commercial context or an enterprise setting," said Dolan. As a result, innovators must identify who holds the data they need, understand the applicable terms of use, and determine whether those terms permit aggregation with other datasets, redistribution, and deployment within intended business models. "Generally, these answers don't meet the needs of the business, because data providers commonly restrict access and usage," he said. "Historically, data has remained tightly controlled, despite longstanding legal frameworks recognizing that data itself is typically not subject to exclusive ownership rights."[110]

To fill that gap, Dolan pointed to the work of the Overture Maps Foundation, whose members were "powering current and next-generation map solutions by creating reliable, easy-to-use, and interoperable open map data" that developers can use.[111]

"Successful projects have the buy-in of a governing board and supporters so that when they announce it, the community shows up," said Dr. Frank Nagle of MIT. "It's not just one project by itself; it's a consortium ready to grow and invested in growing the project over time."[112]

As members of the FINOS and Overture Maps projects, the BrightQuery team experienced firsthand the Linux Foundation's excellence in curating and cultivating technical communities and in building platforms for managing open projects to high operational standards. The team values what Linux brings to the table: neutral stewardship of projects with impartial managers who help move them forward; commercial engagement to accelerate adoption and growth; clarity around the intellectual property rights of contributors and users; and governance models where the doers are the decisionmakers in the project community (Figure 12).[113] Its DevOps environment and infrastructure are the state of the art, as are its security and maintenance tools. Project stewards have metrics and dashboards for gauging project health and guides to monitoring and maintaining compliance.

Figure 12: Proven system for standing up new projects



Nick Hart, president and CEO of the Data Foundation, suggested not only doing "an accurate audience projection at the front end" and architecting the system for the target audience and its core use cases but also having "a clear sense of directionality of the actual value proposition for that audience."[114]

Nailing the incentive model

In his experience, Joel Gurin of the Center for Open Data Enterprise (CODE) observed, "For open data advocates, figuring out models that really work for large scale sharing can be challenging. It happens largely where it has to happen for the businesses' own interest."[115] And so the first priority is nailing the right incentive models for this dataset so that it evolves continuously in both the breadth and the depth of entities represented and the completeness and the quality of data, on each entity. The Linux Foundation has a well established model for open source projects whereby the incentive for contributing to a project derives from the value that a company or an individual draws from adopting that project.

Individuals also have fairly established incentives to contribute, such as fueling their passion, getting their next job, or demonstrating their technical prowess in public. Organizational users have a vested interest in making the project as bug free as possible, contributing their own data, maintaining data quality, or adding the features they need. Nailing this multifactorial incentive model—where the project asks users for money, sweat equity in the form of adding or validating new data, or both—is the first critical decision point for the project's success.

For projects like OpenData.Org, that includes creating incentive models for the existing data vendors—the Bloombergs, the FactSets, and the MorningStars of the world that supply the financial services industry with massive datasets—to see such a project as additive and not competitive. That might help to overcome inertia or absorb the switching or integrating costs of a new data pipeline for those who "have Bloomberg in their DNA."[116]

Friedrich Lindenberg, founder of the OpenSanctions project, described OpenSanctions as a for-profit company, but it sells licenses, not data. "The people who need sanctions data don't need a data file. They need a vendor relationship. They need a supplier who can guarantee them that the data will be available, updated, and of a certain quality for a commercial use case," he said. "In this universe, selling the data is unnecessary if you can sell the service. The site works as an open data portal where journalists, media organizations, academics, and civil society use it for free."[117]

Setting the project's mission and values

For a project new to the open source community, Dr. Nagle of MIT wanted to see a clear mission with a clear scope, clarity around the boundaries of the project, and clear leadership. He has witnessed the growing pains of projects that shifted from a single entity's "calling all the shots" to a governing board's underwriting.

As an example, he pointed to the machine learning framework PyTorch: "PyTorch was already open source, and it already had a wide community of contributors, but it was owned by Meta. People speculated that contributions would increase when Meta moved the governance of the PyTorch project to the nonprofit PyTorch Foundation in 2022."[118] Dr. Nagle and his colleague found that, while Meta's own contributions to the project significantly decreased, contributions from other for-profit companies increased, especially from complementors like chip manufacturers, whereas the rate of contributions from app developers and cloud providers remained consistent.[119] Dr. Nagle tied that back to a clear vision of the project's primary stakeholders, what they care about, and what they're trying to accomplish through it.[120]

Lane Becker, president of Wikimedia Enterprise, underscored the importance of a value system embodied in a clear mission, "principles we'll live by," and an invitation to "hold us to them."[121] He recommended "committing at a deep level, explicitly writing them out, 'Here's what we believe, here's what you can hold us to,' and make it public, so that you're actually beholden to it." At the Wikimedia Foundation, he said, "We may express our value system in different ways to meet the current moment, but our commitment to the value system itself is unparalleled. We've held on to it for such a long time that dropping it in the service of anything else is unthinkable inside the organization."[122]

Rich Skrenta, executive director of the Common Crawl Foundation, raised an ethical consideration, namely, the balancing of private and public interests.[123] Organizations building open data platforms must ask themselves whether the public benefits of large-scale accessibility outweigh the potential private harms associated with the loss of the "practical obscurity" of paper-based public records that are technically accessible but difficult to retrieve.[124]

The ethical recruitment and treatment of data subjects as well as data gatherers and data enrichers are paramount. Stewards of open data projects must commit to responsible sourcing principles and practices that protect and respect the privacy rights of subjects and the labor rights of workers.[125]

Designing for adaptability

Dr. Hart of the Data Foundation recommended "having an infrastructure that's adaptable, malleable even, but nimble enough to be adjustable to changes in policies or the management of government or business that can affect the underlying data in real time." He said, "Make sure that you plan for all those with good data architecturing." In his view, the statistical infrastructures of the Bureau of Economic Analysis and the Census Bureau have endured because "they have a very clear set of processes, get public advice from experts, and are transparent about their methods."[126]

Dr. Hart acknowledged the continuum of licenses between fully open and totally privileged that data users must navigate: "In our organizational model, we use the word *openness* because we view it as a pipeline where some data will always be fully restricted or proprietary, but it contributes to more open data." He pointed to the GDP as a public domain data point that comprises various public, proprietary, and private data.[127] "It's an example of why we need to recognize the spectrum of access points to different kinds of data," he said. "We'll treat some data as inherently confidential and private, and other data as more open as we aggregate up, into the GDP, which we can show by industry and by different levels of the economy, to inform our decisions from investing to policymaking."[128]

"Data governance is key if you want to have useful information," Dr. Hart added. "If you can't govern the data at the starting point, then everything people ask about the data becomes more expensive." He suggested designing and delivering the service in a way that governs the data well and having a mechanism in place for improving the data. "Data governance is somebody's priority, but it's also integrated in the service delivery." The delivery team builds it into workflows.

For the scale and scope of projects like BrightQuery's, Dr. Hart cautioned against designing it for highly technical users or for highly sophisticated research. He thought the project would need "something like Amazon's rating scale and comment system that lets people give real-time active feedback at a community level, where users can see that, for example, a hundred other people had a data quality problem for their jurisdiction or their kind of company."[129] He envisioned the system as a "very quick quality assurance control, not exclusively on the entity that's producing the data," and he thought the project team could "build it as a rapid mechanism."

Engaging users in governance

"Whichever governance model the project adopts will be a differentiator," said Prem Ramaswami of Google's Data Commons. Users and contributors need transparent processes for updating, adding, and deleting data and substantiating claims and sources. "Inviting organizations to own their own entities gives them the best incentive to keep their data up to date," he said, as long as those updates don't grow stale in the review queue as the project grows.[130]

Lindenberg of OpenSanctions has seen too many open data projects whose organizers spent time developing "comprehensive governance mechanisms but nobody ended up using the data." For him, the key is to "let people play. Create a space where people can play with the data, share what they've done, and allow them to mix it freely." That's how project originators learn. "Once you know the demands on the dataset, you can build mechanisms around it and say, 'This is what we need in place for this dataset to do what it's doing.'"[131]

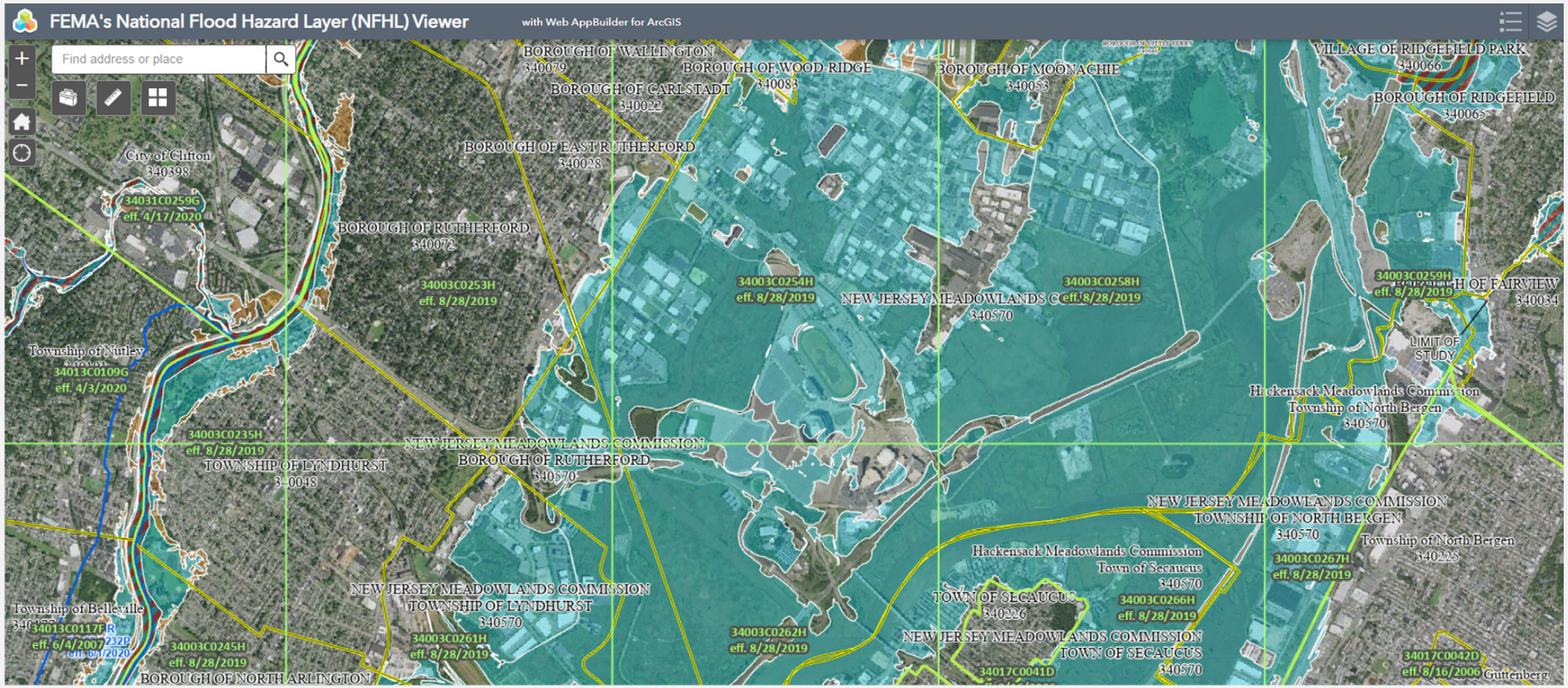
"Data governance is key if you want to have useful information, if you can't govern the data at the starting point, then everything people ask about the data becomes more expensive."

For example, CODE worked with the Federal Emergency Management Agency (FEMA) to engage frontline communities in assessing the value of FEMA's flood maps.[132] The project resulted in important user feedback to the agency and CODE's public Flood Information Products Resource Hub of data, maps, and other assets for mitigating the risks of flooding.[133] For Joel Gurin of CODE, the experience underscored a key tenant of open data stewardship: "When we think about the public review or public governance of datasets, we must think about its potential impact not just on the quality and accuracy of the data but on the people and entities it represents," he said.[134]

Under the Linux Foundation, many projects use a maintainer-led governance model through which anyone can contribute via patches and pull or merge requests.[135] Committers or maintainers review those changes, and authorized maintainers, special interest groups, or technical steering committees decide which changes to make.[136] Each project is independent; the Linux Foundation helps to set up and steward the project as a legal entity with support for the project's licenses, contributor contracts, and intellectual property (e.g., trademarks and brand names), its infrastructure (e.g., continuous integration, security, and compliance), and its events and community.[137]

Figure 13: National Flood Hazard Layer Viewer

In this view centering on the New Jersey Meadowlands, location of the FIFA World Cup 2026 final, blue/cyan indicates a high-risk floodplain, tan/orange indicates a moderate-risk flood area, no shading indicates a lower-risk area, and red stripes denote a floodway or area of strongest flow.



Sources: Federal Emergency Management Agency, National Flood Hazard Layer Viewer, US Dept. of Homeland Security; basemap imagery: US Department of Agriculture and US Geological Survey, accessed 24 May 2026 and used under 17 U.S.C. § 105. https://hazards-fema.maps.arcgis.com/apps/webappviewer/index.html?id=8b0adb51996444d4879338b5529aa9cd.

Large-scale open data projects must also consider complex legal and compliance obligations. When collecting, transforming, and distributing web data, for example, organizations like the Common Crawl Foundation must manage a "stack of liability."[138] It includes tracking provenance, detecting and removing sensitive material such as personally identifiable information or illegal content, and honoring copyright restrictions, web crawling and archiving opt-out preferences, Digital Millennium Copyright Act takedown requests, and "right to be forgotten" requests under such regulations as the European Union's GDPR, all of which Common Crawl does.[139] The lesson for new open data initiatives is that legal governance must evolve alongside technical governance; moving too quickly without safeguards can create substantial legal exposure.

Maintaining high quality data

"Full transparency on the provenance of the data is pretty critical," said Becker of Wikimedia Enterprise. "Where did this information come from? To what degree can this project give complete transparency into the data, its origins, its structures, who contributed to it and how? From the Wikipedia point of view, it's pretty straightforward."[140]

An equally important priority is high quality data. "Why does everybody keep talking about 'getting the data right'? Because it's hard, and data gets 'not right' very fast," said Gavronsky. "Even if it's right for a moment, it gets 'not right' very fast," especially nonreferential data. "Everybody must make it even 'more right' for themselves to some degree and, very importantly, must constantly validate that the data continues to be right."[141] Even data subscribers must prepare the data for their needs and actively manage it as an asset.

"Vendors looking to sell you data will say, 'I have the best and the most amount of data.' That wouldn't excite me too much, because we're talking about quality; and quality is hard to measure and harder to prove. Yet, that's where the real value lies," said Gavronsky. She sees real opportunities in crowdsourcing public data, which vendors have effectively collected, reorganized, cleaned up, and then sold to customers who value those efforts. "Perhaps they're paying for the most current data; but last week's data is open, free, and potentially very valuable for organizations that can't afford it otherwise. If we all put our public data together, then managing it will be cheaper for everybody, and everybody will get a better dataset. That's an opportunity. That's what we're here for," Gavronsky said of FINOS.

Managing risks and building trust

The project stewards and governing board must be candid about the risks from the get-go. "If you look at the risks squarely in the eye, then you will address them," Gavronsky said. "Better to think about all the potential problems in your product now, because if you don't, then somebody else will point them out to you."[142] As an example, she talked about the extent of *openness and data correctness*, comprised of field density and ongoing correctness.

"First, be clear about what's public, what's free, and what's not free." For example, let's say these 300 fields describe a corporation. Under this open data model, users get the first 100 for free—the first 100 fields are open—and users pay for the other 200 fields.

Second, be clear about how many records have totally complete fields. "Let's say we have 50 million records of companies. How many of those 50 million records have those free 100 fields filled out? That's what I call *field density* or *data density*. If you have LEIs for only 100,000 companies, then that's not very interesting," considering that GLEIF has issued 3.23 million LEIs, of which 2.99 million are active.[143]

Third, be clear about how many of those fields are correct. "Let's say those 100 fields are filled out. Are they correct today? What about tomorrow and the day after and the day after?"[144]

Last, how do the project stewards know these fields are correct? Which classic tests are they running? What bullet-proof process have they set up, which people are managing this process, how have they trained these people, which checks have they put in place to make sure those people are following the process, and so on.

"Financial services companies ask such questions of their data vendors as well as themselves because they cannot pass their liabilities on to the vendors. The regulators will fine them, not their vendors," Gavronsky said.[145]

Contingency planning is key. Gurin of CODE described the all-hands-on-deck efforts of the open data community around the Climate and Economic Justice Screening Tool. The Council on Environmental Quality had fully developed and deployed the tool and was using it to guide federal investment under the Justice40 Initiative.[146] Then, in January 2025, federal action removed it from public access. Within a matter of days, the Environmental Data and Governance Initiative—formed in 2016 as an environmental data watchdog and preservation network of sorts—worked through a new collaboration of the Public Environmental Data Partners to reconstruct the tool from archived datasets and restore it online on an independent platform.[147] "Nonprofits put it back up," Gurin said. "They had a huge community input process for that, and they replicated it very effectively."[148]

Choosing a license for open innovation

To get the degree of global entity resolution that OpenData.Org's open entity graph offers, organizations have typically subscribed directly to proprietary data feeds or accessed them via analytics platforms under the proprietor's terms of use, which tend to limit what organizations can do. With an open license such as CDLA Permissive 2.0, users could combine the dataset with others such as OpenSanctions, for example, to develop trade compliance solutions.[149]

The choice of license is especially important in training AI models. According to the Data Provenance Initiative (DPI), many of the datasets developers use to train AI systems lack clear, consistent licensing and attribution.[150] The result is widespread legal and ethical uncertainty around how they're actually using data in model development. The DPI also found that website owners have begun restricting access to their web data in response to AI companies' scraping and using their content to train commercial models; and owners' efforts to regain control over their data is shrinking the open web available for AI training.[151] For transparency, accountability, and lawful use of training data, the DPI calls for the AI community—researchers, companies, and policymakers—to adopt systematic data provenance practices such as better documentation, standardized licensing, and auditable data pipelines.[152]

The goal of using open licenses like CDLA Permissive 2.0 is to make the combining of data from different contributors as permissive as possible so that usage is relatively friction-free. "Even with open licenses, a project can experience licensing friction," said Dr. Rose of the Overture Maps Foundation.[153] For example, under 17 US Code § 105, copyright protection is unavailable for US Government works—text, data, or software—created by US Government officers or employees in their official duties.[154] Instead, those works enter the public domain upon publication within the United States.[155] Therefore, federal entities cannot build upon datasets carrying ShareAlike or other restrictive license terms; nor can they contribute their public domain data through platforms that would impose such terms on the resulting combined dataset. Commercial users such as Google wouldn't touch a ShareAlike license either because they couldn't guarantee always sharing whatever they created.[156]

"Pay very close attention to how you're licensing your data and what downstream impact that might have on combining it with other datasets," Dr. Rose said.[157] Set up policies and processes that prevent ingesting data not licensed expressly under open data agreements, and issue clear attribution and licensing information for contributed data so that users understand the terms of use.[158]

"Think critically about how you want the data to be used and not used," said Lane Becker, president of Wikimedia Enterprise. "Since the beginning, we made a conscious choice to keep our licenses incredibly lenient to enable all pathways that help people access free knowledge. This even includes content forks, which are acceptable in line with our open knowledge ethos—even when they don't necessarily align with our own organization's mission or goals."[159]

For example, Wikipedia makes its text available under a Creative Commons Attribution-ShareAlike 4.0 License (CC BY-SA 4.0).[160] In 2023, a Russian editor forked Wikipedia to create Ruwiki, a version more aligned with the Russian Federation's political priorities and messaging strategies.[161] Another example is Grokipedia, Elon Musk's 2025 fork of Wikipedia. Instead of human authors and editors, it uses xAI's Grok language model. In comparisons of references, researchers found that Grokipedia cited more corporate, government, think tank, and user generated content, including sources deemed unreliable in the Wikipedia community such as the conspiracy outlet Infowars, the neo-Nazi forum Stormfront, and the white nationalist site VDare.[162]

"Revisiting licensing structures has been the third rail of open source," Becker said. "It's as if we've moved to book five of the *Harry Potter* series, where the kids are running around saying, 'Voldemort's back!' And the adults are responding, 'No, he's not.'" The students have an eyewitness and evidence; the professors have faith in their institution. "We're moving into the showdown phase, because it's so bad, people can't ignore it anymore."[163]

Communicating with the community

"Velocity here is a network effect—what happens when like-minded people share resources and compound each other's efforts. Things get built faster. Creators create faster," said Jeff Jonas of Senzing.[164]

For example, the idea for Wikimedia Enterprise emerged as a recommendation from the Wikimedia community through a two-year open consultation process.[165] The community recognized that it needed a revenue strategy to "increase the sustainability of the Wikimedia movement" and the Foundation's mission to "make and keep useful information from its projects available on the Internet free of charge, in perpetuity."[166]

The community prioritized developing an enterprise-grade API and turning it into a revenue source; Wikimedia Enterprise was the mechanism for doing that.[167] Launched in 2021 under a wholly owned limited liability company of the Wikimedia Foundation, Wikimedia Enterprise delivers high-volume Wikimedia and Wikipedia project data as a paid service to commercial customers.[168]

Supporting the participatory process was a whole team dedicated to what Wikimedia calls "movement communications."[169] "Our team spends a lot of time thinking about the best ways to engage in the community and getting their feedback on what we're doing early on," Becker explained. "Not just because their feedback is critical to getting products or services right, but also because we want to enable space for the communities to have the context around our plans before we make something public. It's a big chunk of the work."

"Everything we do at this organization is done in concert with the community in some way, shape, or form," said Becker. "The tension is mostly productive. Even the most negative voices are helpful to have in the ecosystem. At the end of the day, it's still just people talking to each other."

Tapping, training, and resourcing global talent

Since using data, especially in AI models, requires a level of resources, Dandapani of the UNICC advocated for global equity because, "at this moment, open data is disproportionately benefiting actors who have bigger AI capacity or the largest compute."[170] According to the International Monetary Fund, low-income economies lack access to hyperscale computing facilities, ultra-fast networks, high-capacity power generation and cooling systems, and a skilled workforce to use and maintain them.[171] "Improving the capacity of Global South actors is essential," Dandapani said.

Her colleague Emily Bennett, head of Digital Public Solutions and the UNICC Open Source Program Office, agreed. "Probably the biggest challenge is capacity and knowledge. To benefit from the landscape of solutions, we need to look at the educational systems that underpin talent, not just at specific technologies," she said. "Working with the global community around data generally, I see that entities have some kind of data governance but have not really considered sensitivity classifications."[172] They might look more at where the data originated or where it resides than at whether it should be open in the first place.

Both agreed that targeted educational collaborations between academia and global organizations could help. For example, the UNICC and New York University School of Professional Studies partnered to use open data, generative AI, and agile methods in developing a multimodal prototype for detecting, evaluating, and classifying harmful content across six UN languages and media formats.[173] Another approach is engaging students and their professors in hackathons around the use cases for open datasets and inviting them to add or update data relevant and specific to their community.

Promoting the minimum viable product

For a project's debut under the Linux Foundation, the experts interviewed recommended that project stewards spotlight its minimum viable product.

Highlight uniqueness. "Give enough detail so that people can start to see the possibilities," said Dr. Rose of the Overture Maps Foundation.[174] The point of uniqueness could be as simple as the business name, its precise location, and its industry classification code. For example, the OpenData.Org dataset covers those for entities in 222 countries and territories around the world, not just the United States.

Identify gaps. Most existing options for entity resolution data are proprietary in silos. This project meets industry's need for a durable open source solution that connects all the data silos. "Adding a precisely geocoded spatial component to business entity data is really exciting and incredibly useful," said Dr. Rose of the Overture Maps Foundation.[175]

Avoid accruing too much technical debt, described as the trade-offs that developers make, for example, in the robustness, optimization, or reproducibility of the code or the drafting of documentation.[176] "You want to move fast because you want to involve others in the project. You want to put something out there for people to look at," said Dr. Rose. "But be very scrupulous about technical debt because, in an open project, you want to make sure that it's sustainable—sometimes by a few, sometimes by many, but almost always by rotating people. And so, from a technical perspective, you want to make it easy to come in, work on the project, leave, and come back in. You need that good foundation."[177]

Lean into conversations, consensus building, and collaboration. Forums such as task forces and working groups in pre-competitive collaborations are force multipliers of creativity: they free up funding, and they free participants from repeating and reinventing what others have done to surfacing and shaping new ideas, new solutions.[178] "You get people from different companies in a room together to talk and arrive at a solution everybody feels comfortable with. You've reached that solution by consensus, and everybody knows the project has done its due diligence," Dr. Rose explained. "Whether it's external stakeholders, people who are supplying data, organizations that are using the data, or members of the project, make sure you're leaning into those conversations. That builds trust. It also helps you understand how the project should evolve so that everybody still finds it meaningful."[179]

Remember, not all data must be open. "We want to make sure we have data that helps us uniquely identify an entity but doesn't need rich detail, partly because that's actually not commoditized data; that's competitive data," Dr. Rose clarified. "What we're really doing is laying a foundation for everybody to speak the same language, and we're making it very easy for users to attach all sorts of other data to entities, whether it's other open data or private data or sensitive data or even classified data."[180]

Acknowledgments

Many thanks to these experts for sharing their experience in the stewardship and use of open data, all with candor, good cheer, and clear-eyed optimism:

- Lane Becker, president, Wikimedia Enterprise
- Emily Bennett, head, Digital Public Solutions and the United Nations International Computing Centre Open Source Program Office
- Gabriele Columbro, executive director, Fintech Open Source Foundation
- Anusha Dandapani, chief, AI Hub, United Nations International Computing Centre
- Heidi Picher Dempsey, US research director, Red Hat LLC
- Michael Dolan, senior vice president of legal and strategic programs, The Linux Foundation
- Dr. Jose M. Plehn, founder and CEO, BrightQuery Inc.
- Jane Gavronsky, chief operating officer, Fintech Open Source Foundation
- Eloísa Granado, enterprise account executive, Wikimedia Enterprise
- Dr. Ramanathan V. Guha
- Joel Gurin, president and founder, Center for Open Data Exchange
- Dr. Nick Hart, president and CEO, Data Foundation
- Dr. Shay Hershkovitz, vice president of research, BrightQuery Inc.
- Dr. Jeff Jonas, founder, CEO, and chief scientist, Senzing Inc.
- Friedrich Lindenberg, founder, OpenSanctions
- Brian Malone, head of partnerships and product strategy, BrightQuery Inc.
- Dr. Frank Nagle, advising chief economist, The Linux Foundation; and research scientist, Initiative on the Digital Economy, Massachusetts Institute of Technology
- Isio Nelson, managing director of research, fraud, and thought leadership, ProSight Financial Association
- Prem Ramaswami, head, Data Commons, Google LLC
- Dr. Amy Rose, chief technology officer, Overture Maps Foundation
- Rich Skrenta, executive director, Common Crawl Foundation

Special thanks to Hilary Carter, senior vice president, Linux Foundation Research, and Anna Hermansen, senior researcher and ecosystem manager, Linux Foundation Research, for their creative and substantial contributions throughout the research and review process.

Thanks to the following organizations that sponsored this study:

BrightQuery® The Factual AI Company™ FINOS OVERTURE MAPS

Credits

Figures 1 and 3 feature a location icon by gungyoga04, a bank (legal entity) icon by Smashicons, a (home) address icon by Ahmad Roaayala, a user (people) icon by Freepik, and an calendar (organization) icon by Smashicons, all used under a Magnific free license. Figure 12 features a lightbulb alert icon by Michael Irigoyen, a rocket launch icon by Michael Irigoyen, a scale-balance icon by Simran, a compass icon by Google, an account group icon by GreenTurtwig, and a cash multiple icon by Austin Andrews, all from MDI/Pictogrammers and used under Apache License 2.0.

About the author

Kirsten D. Sandberg is a researcher, writer, and editor who collaborates regularly on research projects in the areas of intellectual property, innovation, science, and technology. Kirsten is editor-in-chief of the Blockchain Research Institute and an adjunct faculty member of the graduate publishing program at Pace University. For over a decade, she was an executive editor specializing in strategy, marketing, and finance at Harvard Business Publishing.

Appendix

Table A1 shows the starting items in the open dataset, Table A2 elaborates on the data types included in the open dataset, Table A3 shows items available in the BrightQuery premium dataset, and Table A4 describes the premium data types.

Table A1: Open data available on the main pillars of the economy

This open entity data project spans 222 countries and territories.

FIELD TYPE	ORGANIZATIONS	LOCATIONS	ADDRESSES	PEOPLE
Global	**324 million**	**512 million**	**197 million**	**1.2 billion**
United States	**85.8 million**	**344.9 million**	**45.4 million**	**249 million**
Open data	• Company ID • Company name • Legal name • D/B/A • Website URL • Company LinkedIn URL • Wikipedia URL • Crunchbase URL • Phone number • Number of locations • Ticker symbol • Capital IQ ID • CIK • FIGI ID • Set of LEIs • Set of ISINs • NPI ID • OSM ID • PERM ID • SAM ID Placekey • Google Plus Code IDs • HQ address with latitude-longitude and administrative area	• Location ID • Company ID it relates to • Location name • Website • Location phone number • Location address with latitude-longitude and administrative area and postal code • Placekey • OSM ID • GERS ID • Google Plus Code ID	• Building ID • Company IDs it relates to • Number of organizations • Legal entities • Locations and units associated with the address • Address with latitude-longitude and administrative area • Placekey • GERS ID • OSM ID • Google Plus Code IDs associated with the address	• Person ID • Company IDs it relates to • Person full name • State or province • Country • Person LinkedIn URL • Instagram URL • Wikipedia URL

Sources of data: "The World's Largest Open Global Entity Graph," Open Data Flyer, 10 Oct. 2025, p. 3, https://opendata.org/BrightQuery-Open-Data-Flyer-Oct-2025.pdf; Dashboard, https://docs.brightquery.com/dashboard; Data Dictionary, https://opendata.org/#datadictionary; Data feeds, https://docs.brightquery.com/data-feeds; and Dr. Shay Hershkovitz, vice president of research, BrightQuery, email to Kirsten Sandberg, 31 May 2026.

Table A2: Details of data types in OpenData.Org

ABBREVIATION	FULL NAME	PURPOSE	STEWARD	SOURCE LINK
Capital IQ ID	Standard & Poor's (S&P) Capital IQ Identifier	Tracks entities and securities within the S&P database.	S&P Global Market Intelligence	https://www.spglobal.com
CIK	Central Index Key	Identifies corporations and individuals who file with the SEC.	US Securities and Exchange Commission	https://www.sec.gov
FIGI ID	Financial Instrument Global Identifier	An open standard for identifying financial instruments.	Object Management Group; Bloomberg LP (Registration Authority)	https://www.openfigi.com
GERS ID	Global Entity Reference System	A legacy identifier for global corporate entity mapping.	Overture Maps Foundation	https://overturemaps.org/gers/
IRS	Internal Revenue Service	US agency responsible for tax collection and law enforcement.	US Department of the Treasury	https://www.irs.gov
ISIN	International Securities Identification Number (ISO 6166)	Uniquely identifies a specific securities issue (stocks, bonds).	International Securities Identification Numbers Organization	https://www.isin.org/about/
LEI	Legal Entity Identifier (ISO 17442)	A global 20-character code to identify distinct legal entities.	Global Legal Entity Identifier Foundation (GLEIF)	https://www.gleif.org
NPI ID	National Provider Identifier	A unique 10-digit identification number for covered health care providers.	Centers for Medicare and Medicaid Services (CMS)	https://www.cms.gov
OFAC	Office of Foreign Assets Control	Administers and enforces economic and trade sanctions.	US Department of the Treasury	https://home.treasury.gov
OSM ID	OpenStreetMap Identifier	A unique ID for features (buildings, roads) in the OSM database.	OpenStreetMap Foundation	https://www.openstreetmap.org
PERM ID	Permanent Identifier	An open, permanent ID for entities, people, and organizations.	London Stock Exchange Group plc (LSEG) (formerly Refinitiv/Thomson Reuters)	https://www.permid.org
Placekey	Placekey	A universal standard identifier for physical places/points of interest.	Senzing acquired Placekey in 2025	https://www.placekey.io

Table A2: Details of data types in OpenData.Org (continued)

ABBREVIATION	FULL NAME	PURPOSE	STEWARD	SOURCE LINK
PLEI	Proto Legal Entity Identifier	A global alphanumeric identifier for legal entities	OpenCorporates	https://p-lei.com/
RDI	Residential Delivery Indicator	Identifies whether a specific address is residential or commercial.	United States Postal Service (USPS)	https://www.usps.com
REG ID	Legal Entity Registration ID	Unique number issued by the jurisdiction with which the Legal Entity is registered.	Secretary of State	https://www.sos.ca.gov/
SAM ID	System for Award Management ID	Required for entities doing business with the US Federal Government.	General Services Administration (GSA)	https://sam.gov; https://www.ecfr.gov/current/title-2/subtitle-A/chapter-I/part-25

In 2022, the US General Services Administration switched from Dun & Bradstreet's proprietary Data Universal Numbering System (DUNS) to its own System for Award Management (SAM) ID, an open identifier standard with a unified tracking system for contractors and other third parties doing or seeking to do business with the government. All the US government's other integrated award environment systems—such as CPARS, eSRS, FPDS, and FSRS—use the SAM ID for KYC and AML, and so the open solution has lowered the barriers to entering the federal marketplace.[181]

Table A3: Premium data available from BrightQuery

FIELD TYPE	ORGANIZATIONS	LOCATIONS	ADDRESSES	PEOPLE
Premium data	• Company type • Structure • Employer identification number (EIN) • NAICS, SIC, and IRS primary and all codes • names and sector data • OFAC and tax lien indicators • Year founded • BQ credit and confidence score • Most recent and historic revenue and employment numbers along with income statement and balance sheet items	• Organization name and website tied to the location • Location type • Class location categories (over 10,000) • Location confidence score • Location active indicator	• Building address type • Active indicator • State • Address RDI details • Building unit type • Address valid indicator • Building unit record type and carrier route • Legal entity and location level mapping to the address	• Legal, business, and consumer • Organization name • Website and ticker associated with person • Person's title • Person's role • All publicly available email and phone numbers associated with person • Complete address (often residential)

Source of data: "The World's Largest Open Global Entity Graph," Open Data Flyer, 14 Oct. 2025, p. 3, https://opendata.org/BrightQuery-Open-Data-Flyer-Oct-2025.pdf.

Table A4: Details of additional data types in premium dataset

ABBREVIATION	FULL NAME	PURPOSE	STEWARD	SOURCE LINK
Capital IQ ID	Standard & Poor's (S&P) Capital IQ Identifier	Tracks entities and securities within the S&P database.	S&P Global Market Intelligence	https://www.spglobal.com
EIN	Employer Identification Number	Used by the IRS to identify business entities for tax purposes.	Internal Revenue Service (IRS)	https://www.irs.gov
NAICS	North American Industry Classification System	Classifies business establishments by their primary economic activity; replaced SIC.	US Census Bureau	https://www.census.gov
SIC	Standard Industrial Classification	A four-digit code used to identify the primary business of an entity; replaced by NAICS.	US Department of Labor (OSHA)	https://www.sec.gov/search-filings/standard-industrial-classification-sic-code-list; https://www.osha.gov/data/sic-manual

References

1. "How the Roman Census Worked (and Why It Mattered)," *UNRV Roman History*, n.d., https://www.unrv.com/government/how-the-roman-census-worked-and-why-it-mattered.php; Jeremy M. Norman, "The Domesday Book, Recording the First English Census," *History of Information*, ID 221, n.d., https://www.historyofinformation.com/detail.php?id=221; Ban Gu, "Han Shu," Chinese Text Project, n.d., https://ctext.org/han-shu; "Nonconformist and Non-Parish Registers: Births, Marriages and Deaths, 1567–1969," National Archives, UK Government, n.d., https://www.nationalarchives.gov.uk/help-with-your-research/research-guides/nonconformist-non-parish-births-marriages-deaths-1567-1969/.

2. LexisNexis Risk Solutions Study Reveals Global Financial Crime Compliance Costs for Financial Institutions Totals More Than US$206 Billion," Press Release, LexisNexis Risk Solutions, 26 Sep. 2023, https://risk.lexisnexis.com/global/en/about-us/press-room/press-release/20230926-global-financial-crime-compliance-costs.

3. Tim Berners-Lee, "30 Years On, What's Next #ForTheWeb?" open letter, World Wide Web Foundation, 11 Mar. 2019, https://medium.com/@webfoundation/30-years-on-whats-next-fortheweb-30a1d1ea034c.

4. BrightQuery Open Data Flyer, OpenData.Org, 10 Oct. 2025, https://opendata.org/BrightQuery-Open-Data-Flyer-Oct-2025.pdf. See also R.V. Guha and Vineet Gupta, "Reference by Description," *arXiv*, 7 Mar. 2016, https://arxiv.org/pdf/1511.06341; Nishadi Kirielle, Peter Christen, and Thilina Ranbaduge, "Unsupervised Graph-Based Entity Resolution for Complex Entities," *ACM Transactions on Knowledge Discovery from Data* 17, no. 1 (20 Feb. 2023): 1–30, Association for Computing Machinery, https://doi.org/10.1145/3533016; and Mohammad Hossein Moslemi et al., "Heterogeneity in Entity Matching: A Survey and Experimental Analysis," *Data and Knowledge Engineering* 164 (Jul. 2026), https://doi.org/10.1016/j.datak.2026.102575.

5. Shay Hershkovitz, vice president of research, BrightQuery Inc., email to Kirsten Sandberg, 31 Mar. 2026.

6. Jose M. Plehn Dujowich, interviewed via Zoom by Kirsten Sandberg, 12 Jan. 2026.

7. Sources of data in BrightQuery's open data project: US government offices, https://docs.brightquery.com/government_offices; US local jurisdictions, https://docs.brightquery.com/local_jurisdictions; US federal agencies, https://docs.brightquery.com/federal_agencies; US Secretary of State offices, https://docs.brightquery.com/secretary_of_state_jurisdiction; and Canadian agencies and registries, https://docs.brightquery.com/canadian_agencies_registries, as of 24 Feb. 2026.

8. Shay Hershkovitz, vice president of research, BrightQuery Inc., email to Kirsten Sandberg, 31 Mar. 2026.

9. "What does GDPR stand for?" https://gdpr.eu/what-does-it-stand-for/; "Right to erasure," https://gdpr-info.eu/art-17-gdpr/; California Consumer Privacy Act of 2018, https://oag.ca.gov/privacy/ccpa; and "Right to delete," https://oag.ca.gov/privacy/ccpa#sectiond.

10. Tokenization here refers to substituting a sensitive data element with a nonsensitive surrogate value that can be re-associated with the original data only through separately secured mapping. See ISO/IEC 20889:2018, *Privacy Enhancing Data Deidentification Terminology and Classification of Techniques* (Geneva: International Organization for Standardization, 2018), https://www.iso.org/standard/69373.html; and Simson Garfinkel et al., NIST SP 800-188, *Deidentifying Government Datasets: Techniques and Governance* (Gaithersburg, MD: National Institute of Standards and Technology, Sep. 2023), https://doi.org/10.6028/NIST.SP.800-188.

11. Jose M. Plehn Dujowich, PhD, interviewed via Zoom by Kirsten Sandberg, 12 Jan. 2026. Cade Metz and Karen Weise, "AI Is Getting More Powerful, but Its Hallucinations Are Getting Worse," *New York Times*, 5 May 2025, https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html.

12. See, for example, https://docs.brightquery.com/bulk_data_feed and https://docs.brightquery.com/business_identity_api.

13. Jose M. Plehn, "Laying the Groundwork for Open Data with the Linux Foundation," PowerPoint presentation, BrightQuery, Nov. 2025, p. 52.

14. ChatGPT reached 100 million users in about two months after its launch in late 2022, whereas TikTok took ~9 months and Instagram, ~2.5 years, in part because it leveraged existing Internet, cloud, and smartphone infrastructure, and technology vendors integrated generative AI into their offerings. Krystal Hu, "ChatGPT sets record for fastest-growing user base," Analyst Note, *Reuters*, 2 Feb. 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/; Ravi Mayuram, "Six tips for combatting vendor lock-in and the data gravity challenge," *VentureBeat*, 24 Apr. 2023, https://venturebeat.com/data-infrastructure/6-tips-combat-vendor-lock-in-data-gravity; and Sagar Joshi, "AI Adoption Statistics," *G2 Lean*, G2.com Inc. 28 May 2025, https://learn.g2.com/ai-adoption-statistics.

15. Jim Zemlin, email to Hilary Carter, 17 Mar. 2026.

16. World Bank, "AI-Ready Official Statistics: Opportunities, Challenges, and Recommendations," UNSTATS, Dec. 2025, https://unstats.un.org/UNSDWebsite/statcom/session_57/documents/BG-5h-CCSA_AI_Readiness_Official_Statistics_v2-E.pdf.

17. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

18. "LexisNexis Risk Solutions Study Reveals Global Financial Crime Compliance Costs for Financial Institutions Totals More Than US$206 Billion," Press Release, LexisNexis Risk Solutions, 26 Sep. 2023, https://risk.lexisnexis.com/global/en/about-us/press-room/press-release/20230926-global-financial-crime-compliance-costs.

19. Craig Hale, "Businesses are being 'locked in' to all-in-one platforms, and it's costing them growth and adaptability," *TechRadar*, 20 Mar. 2026, https://www.techradar.com/pro/businesses-are-being-locked-in-to-all-in-one-platforms-and-its-costing-them-growth-and-adaptability.

20. For other data licenses such as C-UDA-1.0, O-UDA-1.0, ODbL-1.0 ODC-By-1.0, and PDDL-1.0, see https://spdx.org/licenses/.

21. Ravi Mayuram, "Six tips for combatting vendor lock-in and the data gravity challenge," *VentureBeat*, 24 Apr. 2023, https://venturebeat.com/data-infrastructure/6-tips-combat-vendor-lock-in-data-gravity; and Superblocks team, "What Is Vendor Lock-in? Five Strategies and Tools to Avoid It," *Superblocks Blog*, DayZero Software Inc., 21 Mar. 2025, https://www.superblocks.com/blog/vendor-lock.

22. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026.

23. "National Secure Data Service Demonstration," National Center for Science and Engineering Statistics, as of 20 Apr. 2026, https://ncses.nsf.gov/initiatives/national-secure-data-service-demo.

24. "BrightQuery and Senzing Partner to Deliver Faster, Smarter Business Data Resolution," Press Release, Senzing Inc., n.d., https://senzing.com/bright-query-partnership-data-resolution/.

25. Jose M. Plehn, "A Shared Entity-Based Infrastructure for the Modern AI Economy: Building the World's Knowledge Graph at the Entity Level," 1 May 2026. Shared with the research project team.

26. Jose M. Plehn Dujowich, PhD, interviewed via Zoom by Kirsten Sandberg, 12 Jan. 2026.

27. A couple of interviewees described how their respective organization's database had multiple instances of a single person or a company. The most common reasons were (1) the lack of a universal unique identifier for each entity across an organization's relational and enterprise systems, (2) variations in data entry and formats such as "IBM," "I.B.M.," or "International Business Machines," and (3) multiple input systems and integration processes that create redundant and sometimes conflicting entries across operational, compliance, and analytics databases. See Gartner Research, "Data Quality: Best Practices for Accurate Insight," Gartner Inc., n.d., https://www.gartner.com/en/data-analytics/topics/data-quality, accessed 1 Mar. 2026.

28. Jeff Jonas, interviewed via Zoom by Kirsten Sandberg, 30 Jan. 2026.

29. Dan Ennis, "$4B in PPP Loans Were Duplicates or Had Mismatched Data, Analysis Finds," *Banking Dive*, TechTarget Inc., Informa PLC, 2 Sep. 2020, https://www.bankingdive.com/news/paycheck-protection-program-mismatched-data-SBA/584578/; and Office of Inspector General (Hannibal Ware), Report 21-09: *Duplicate Loans Made Under the Paycheck Protection Program*, US Small Business Administration, 15 Mar. 2021, https://www.sba.gov/sites/default/files/2021-03/SBA%20OIG%20%20Report%2021-09.pdf.

30. George Anadiotis, "What Is a Knowledge Graph?" Ontotext, n.d., https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/; and Google Data Commons, "About Data Commons," n.d., https://datacommons.org/about, both accessed 1 Mar. 2026.

31. Henrik Liliendahl Sørensen, "MDM and Knowledge Graph," *Liliendahl on Data Management*, 28 Nov. 2021, https://liliendahl.com/2021/11/28/mdm-and-knowledge-graph/; Nishadi Kirielle, Peter Christen, and Thilina Ranbaduge, "Unsupervised Graph-Based Entity Resolution for Complex Entities," *ACM Transactions on Knowledge Discovery from Data* 17, no. 1 (Jan. 2023), https://doi.org/10.1145/3533016; and Prashanth Rao and Paco Nathan, "From Data to Insights: Entity-Resolved Knowledge Graphs with Kùzu and Senzing," *Kùzu Blog*, Kùzu Inc., 9 Jun. 2025, https://blog.kuzudb.com/post/entity-resolved-knowledge-graphs/.

32. Isio Nelson, interviewed via Zoom by Kirsten Sandberg, 13 Feb. 2026.

33. Olivia White et al., *Digital Identification: A Key to Inclusive Growth* (San Francisco: McKinsey Global Institute, 17 Apr. 2019): vi, https://www.mckinsey.com/~/media/mckinsey/business functions/mckinsey digital/our insights/digital identification a key to inclusive growth/mgi-digital-identification-report.pdf.

34. "Companies House celebrates 10 years of open data," UK Government press release, GOV.UK, published 22 Jun. 2025, https://www.gov.uk/government/news/companies-house-celebrates-10-years-of-open-data.

35. Department for Business and Trade and Companies House, *Value of Corporate Transparency in Tackling Crime: Policy Summary* (London: UK Government, 15 Oct. 2024), https://assets.publishing.service.gov.uk/media/670e554d366f494ab2e7b88c/policy_summary_report_value_corporate_transparency_tackling_crime_october_2024.pdf.

36. Department for Business, Energy, and Industrial Strategy and Companies House, *Valuing the User Benefits of Companies House Data: Policy Summary*, BEIS Research Paper Number 2019/015 (London: UK Government, 27 Sep. 2019), https://assets.publishing.service.gov.uk/media/5d8a299aed915d5cff89a4a1/valuing-benefits-companies-house-data-policy-summary.pdf.

37. Alexander Verhagen et al., "How Agentic AI Can Change the Way Banks Fight Financial Crime," *Insights*, McKinsey & Co., 7 Aug. 2025, https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-agentic-ai-can-change-the-way-banks-fight-financial-crime.

38. Hui Gong, "AI Agents in Financial Markets: Architecture, Applications, and Systemic Implications," *arXiv*, 14 Apr. 2026, https://arxiv.org/pdf/2603.13942.

39. Zornitsa Manolova, "Transforming Data into Opportunities: Metric of the Month: Legal Entity Events," *GLEIF Blog*, Global Legal Entity Identifier Foundation, 5 Dec. 2025, https://www.gleif.org/en/newsroom/blog/transforming-data-into-opportunities-metric-of-the-month-legal-entity-events.

40. Experian, "Global Identity and Fraud Report 2024," *Experian Blog*, Experian Information Solutions Inc., 5 Nov. 2024, https://www.experian.com/blogs/global-insights/wp-content/uploads/2024/11/Global_Fraud_Trends_Report_2024_FinalV.pdf.

41. Jessica Hung, Isaac Owusu, and Ross Gabay, "Build Fraud Detection Systems Using AWS Entity Resolution and Amazon Neptune Analytics," *AWS Database Blog*, Amazon Web Services, 5 Feb. 2026, https://aws.amazon.com/blogs/database/build-fraud-detection-systems-using-aws-entity-resolution-and-amazon-neptune-analytics.

42. "Entity Resolution Software," product description, DataWalk, accessed 1 Mar. 2026, https://datawalk.com/product/entity-resolution/.

43. Ian Hook, "OpenCorporates' Data Powers Anti-Money Laundering Knowledge Graph Prototype," *OpenCorporates Blog*, OpenCorporates, 15 Jul. 2021, https://blog.opencorporates.com/2021/07/15/opencorporates-data-powers-anti-money-laundering-knowledge-graph-prototype; and Data Innovation Lab, EDM Council, as of 1 Mar. 2026, https://edmcouncil.org/innovation/innovation-lab/.

44. Mingming Geng, "Enhancing Entity Resolution with Multichannel BERT [bidirectional encoder representations from transformers]: A Comprehensive Approach," *Proc. SPIE 13171*, Third International Conference on Algorithms, Microchips, and Network Applications (AMNA 2024), 1317126 (8 Jun. 2024), https://doi.org/10.1117/12.3031934; and Jwen Fai Low, Benjamin C.M. Fung, and Pulei Xiong, "Better Entity Matching with Transformers Through Ensembles," *Knowledge-Based Systems* 293 (2024), https://doi.org/10.1016/j.knosys.2024.111678.

45. Philip Bruno et al., "The Legal Entity Identifier: The Value of the Unique Counterparty ID," Global LEI Foundation and McKinsey & Co., 15 Oct. 2017, p. 4, https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/the%20legal%20entity%20identifier%20the%20value%20of%20the%20unique%20counterparty%20id/legal-entity-identifier-mckinsey-gleif-2017.pdf.

46. EQT Partners, "CompanyKG: A Large-Scale Company Relation Graph for Investment Industry," GitHub, v1.0, https://github.com/EQTPartners/CompanyKG; and Lele Cao et al., "CompanyKG: A Large-Scale Heterogeneous Graph for Company Similarity Quantification," *arXiv*, 18 Jun. 2023, https://arxiv.org/abs/2306.10649.

47. LLC Research, "A Large-Scale Heterogeneous Graph for Company Similarity Quantification and Relation Prediction, accepted by KDD'2024," GitHub, v2.0, https://github.com/llcresearch/CompanyKG2; and https://zenodo.org/records/11391315.

48. Abhinav Arun et al., "FinReflectKG: Agentic Construction and Evaluation of Financial Knowledge Graphs," in *Sixth ACM International Conference on AI in Finance* (ICAIF '25), Singapore, 15–18 Nov. 2025 (New York, NY: Association for Computing Machinery, 2025): 283–290, https://doi.org/10.1145/3768292.3770363; and the ReadMe page of the Financial Knowledge Graph Dataset: S&P 100 Companies, https://web.archive.org/web/20250826030522/https://anonymous.4open.science/api/repo/KG-Financial-Datasets-SP-100-529B/file/README.md.

49. "Wikidata," Wikipedia, last edited 22 Jan. 2023, https://www.wikidata.org/wiki/Wikidata:Main_Page, accessed 1 Mar. 2026.

50. John Walsh and Ruben Falk, "Agentic GraphRAG for Capital Markets," *AWS for Industries Blog*, Amazon Web Services Inc., 24 Feb. 2026, https://aws.amazon.com/blogs/industries/agentic-graphrag-for-capital-markets/.

51. Heidi Picher Dempsey, "From Particles to Prototypes: What We Learn from Managing Open Clouds," From the Director column, *Red Hat Research Quarterly*, Nov. 2024, https://research.redhat.com/blog/article/from-particles-to-prototypes-what-we-learn-from-managing-open-clouds/.

52. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

53. See ServiceNow, "Create a Knowledge Base," Extend ServiceNow AI Platform Capabilities, 31 Jul. 2025, https://www.servicenow.com/docs/r/servicenow-platform/knowledge-management/create-a-knowledgebase.html.

54. Andrew Brown, Muhammad Roman, and Barry Devereux, "A Systematic Literature Review of Retrieval-Augmented Generation Models," *Journal of Intelligent Systems* (8 Aug. 2025), https://arxiv.org/pdf/2508.06401v1.pdf; and Ehlullah Karakurt and Akhan Akbulut, "Retrieval-Augmented Generation and Large Language Models for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review," *Applied Sciences* 16, no. 1 (29 Dec. 2025), https://doi.org/10.3390/app16010368. See also DEEP-PolyU, "Awesome-GraphRAG," GitHub, last updated 3 Mar. 2026, https://github.com/DEEP-PolyU/Awesome-GraphRAG.

55. "Agentic Hybrid RAG Factual AI Platform," BrightQuery AI Inc., as of 10 Apr. 2026, https://brightquery.ai/agentic/.

56. Shaoxiong Ji et al., "A Survey on Knowledge Graphs: Representation, Acquisition and Applications," *IEEE Transactions on Neural Networks and Learning Systems* 33, no. 2 (2022): 494–514, https://arxiv.org/abs/2002.00388.

57. Stiene Riemer et al., "How Retail Banks Can Put AI Agents to Work," Boston Consulting Group, 9 Mar. 2026, https://www.bcg.com/publications/2026/how-retail-banks-can-put-agentic-ai-to-work.

58. Peter Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection* (Berlin: Springer, 2012), https://doi.org/10.1007/978-3-642-31164-2; and Stefano Marchesin, Gianmaria Silvello, and Omar Alonso, "Large Language Models and Data Quality for Knowledge Graphs," *Information Processing and Management* 62, no. 6 (Nov. 2025): 104281, https://doi.org/10.1016/j.ipm.2025.104281.

59. Jam Kraprayoon, Zoe Williams, and Rida Fayyaz, "AI Agent Governance: A Field Guide," Institute for AI Policy and Strategy, *arXiv*, 27 May 2025, https://arxiv.org/pdf/2505.21808.

60. Elevate, "AI Data Governance: Provenance, Quality, and Model Lineage," *Elevate Consult Blog*, Elevate Consult LLP, 30 Dec. 2025, https://elevateconsult.com/insights/ai-data-governance-provenance-quality-and-model-lineage/.

61. WNS Analytics, "Information Retrieval Using Knowledge Graphs," *WNS Analytics AI Lab*, WNS Holdings Ltd., n.d., https://www.wns.com/capabilities/analytics/wns-ai-lab/enhancing-information-retrieval-using-knowledge-graphs; Ciyuan Peng et al., "Knowledge Graphs: Opportunities and Challenges," *Artificial Intelligence Review*, 3 Apr. 2023: 1–32, https://pmc.ncbi.nlm.nih.gov/articles/PMC10068207/; and Yilun Zheng et al., "Less Is More: Denoising Knowledge Graphs for Retrieval-Augmented Generation," *arXiv*, 16 Oct. 2025, https://arxiv.org/pdf/2510.14271.

62. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

63. AWSLabs, "AWS SageMaker Graph-Based Credit Scoring," GitHub, last updated 27 Feb. 2023, https://github.com/awslabs/sagemaker-graph-based-credit-scoring, accessed 1 Mar. 2026.

64. OpenCorporates Ltd., "OpenCorporates," OpenCorporates, accessed 3 Mar. 2026, https://opencorporates.com/.

65. Isio Nelson, interviewed via Zoom by Kirsten Sandberg, 13 Feb. 2026.

66. See, for example, Wikimedia contributors, "TerminusDB," Wikipedia, last modified 18 Jan. 2026, https://en.wikipedia.org/wiki/TerminusDB; and Wikimedia contributors, "JanusGraph," Wikipedia, last modified 10 Feb. 2026, https://en.wikipedia.org/wiki/JanusGraph.

67. João B.G. de Brito et al., "Potential Customer Lifetime Value in Financial Institutions: The Usage of Open Banking Data to Improve CLV Estimation," *arXiv*, 28 Jun. 2025, https://arxiv.org/abs/2506.22711.

68. Isio Nelson, interviewed via Zoom by Kirsten Sandberg, 13 Feb. 2026.

69. Last edited on 19 Feb. 2026, https://en.wikipedia.org/wiki/Wikipedia:Wikidata.

70. Friedrich Lindenberg, "Inside OpenSanctions' Open Source Approach to Compliance," *21 Analytics Blog*, 21 Analytics AG, 8 Feb. 2026, https://www.21analytics.co/blog/opensanction-open-source-approach-compliance/.

71. Wikimedia contributors, "TerminusDB," Wikipedia, last modified 18 Jan. 2026, https://en.wikipedia.org/wiki/TerminusDB; Apache 2.0 License, https://github.com/terminusdb/terminusdb/blob/main/LICENSE; Wikimedia contributors, "JanusGraph," Wikipedia, last modified 10 Feb. 2026, https://en.wikipedia.org/wiki/JanusGraph; and Apache 2.0 License, https://github.com/JanusGraph/janusgraph?tab=License-1-ov-file#readme.

72. "Improving Different Ultimate Beneficial Ownership Investigations," *GraphAware Blog*, Graph Aware Ltd., 24 Jun. 2024, https://graphaware.com/blog/ubo-investigation-with-graph-technology-copy/.

73. Wikimedia Deutschland, "Wikidata: Embedding Project," Wikidata, Wikimedia Foundation, last modified 27 Jan. 2026, https://www.wikidata.org/wiki/Wikidata:Embedding_Project. See also Dianne Esber et al., "Reinventing marketing workflows with agentic AI," *McKinsey Insights*, 21 Apr. 2026, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/reinventing-marketing-workflows-with-agentic-ai.

74. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

75. "Who We Are," Overture Maps, as of 16 Feb. 2026, https://overturemaps.org/about/who-we-are/.

76. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

77. Richard Robinson, "How Two Unique Identifiers Can Unlock the Value of ESG Data Faster," *Chief Data Officer Magazine*, 29 Apr. 2025, https://www.cdomagazine.tech/opinion-analysis/how-two-unique-identifiers-can-unlock-the-value-of-esg-data-faster.

78. Richard Robinson, "How Two Unique Identifiers Can Unlock the Value of ESG Data Faster," *Chief Data Officer Magazine*, 29 Apr. 2025, https://www.cdomagazine.tech/opinion-analysis/how-two-unique-identifiers-can-unlock-the-value-of-esg-data-faster.

79. Israel Griol-Barres et al., "Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing," *Sustainability* 12, no. 19 (20 Sep. 2020): 7848, https://doi.org/10.3390/su12197848.

80. Christian Mühlroth, Laura Kölbl, and Michael Grottke, "Innovation Signals: Leveraging Machine Learning to Separate Noise from News," *Scientometrics* 128, no. 5 (12 Apr. 2023): 2649–2676, https://doi.org/10.1007/s11192-023-04672-y.

81. Yuchen Zhao, Xiaogang Bi, and Qing-Ping Ma, "Predicting Mergers and Acquisitions: A Machine Learning-Based Approach," *International Review of Financial Analysis* 99 (Mar. 2025), https://doi.org/10.1016/j.irfa.2025.103933; Karen S. Markel, Mihir Tale, and Andrea Belz, "JobPulse: A Big Data Approach to Real-Time Engineering Workforce Analysis and National Industrial Policy," *arXiv*, 14 Aug. 2025, https://arxiv.org/pdf/2508.11014; Sebastian Heinrich, "Deriving Technology Indicators from Corporate Websites: A Comparative Assessment Using Patents," *Applied Economics Letters* 32, no. 1 (14 Aug. 2023): 28–41, https://www.tandfonline.com/doi/full/10.1080/13504851.2023.2244228.

82. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

83. Harold Hotelling, "Stability in Competition," *Economic Journal* 39, no. 153 (1929): 41–57, https://doi.org/10.2307/2224214; and Anthony Downs, *An Economic Theory of Democracy* (New York: Harper & Row, 1957), https://archive.org/details/economictheoryof00down.

84. Ramanathan V. Guha, interviewed via Zoom by Kirsten Sandberg, 30 Jan. 2026.

85. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

86. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

87. Anna Hermansen and Kirsten Sandberg, "The Value of Open Source AI for APEC Economies: A Review of Industry, Academic, and Open Source Evidence," The Linux Foundation, 31 Oct. 2025, https://www.linuxfoundation.org/hubfs/Research%20Reports/LFResearch_OSAI_APEC_Version2_103125.pdf.

88. Emily Bennett and Anusha Dandapani, interviewed via Zoom by Kirsten Sandberg, 10 Feb. 2026.

89. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

90. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026; FINOS Landscape of projects, https://landscape.finos.org/stats.

91. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

92. Friedrich Lindenberg, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026; and "About OpenSanctions," https://www.opensanctions.org/docs/about/, as of 14 Mar. 2026.

93. Nicholas Gruen and John Houghton, *Open for Business: How Open Data Can Help Achieve the G20 Growth Target* (Port Melbourne, VIC: A Lateral Economics report commissioned by Omidyar Network, Jun. 2014), https://lateraleconomics.com.au/wp-content/uploads/omidyar_open_business.pdf; and Jaana Mäkelä and Luukas Raatikainen, *The Economic Value of Spatially Enabled Services in Finland* (Helsinki: Ministry of Agriculture and Forestry and Finnish Geospatial Research Institute, 2018), https://geoforum.fi/wp-content/uploads/2023/04/The-economic-value-of-spatially-enabled-services-in-Finland.pdf.

94. James Manyika, Sven Smit, and Jonathan Woetzel, "Financial Data Unbound: The Value of Open Data for Individuals and Institutions," McKinsey Global Institute, McKinsey & Co., Jun. 2021, p. 6, https://www.mckinsey.com/~/media/mckinsey/industries/financial services/our insights/financial data unbound the value of open data for individuals and institutions/financial-data-unbound-discussion-paper-june-2021.pdf.

95. Len Fishman, "How Open Data Induces Financial Firms to Compete on Quality: An Interview with Bloomberg's Richard Robinson," *data.world blog*, data.world Inc., 15 Aug. 2017, https://data.world/blog/how-open-data-induces-financial-firms-to-compete-on-quality-an-interview-with-bloombergs-richard/.

96. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026. See also "Roundtable on Data for Automated Vehicle Safety Summary Report," foreword by Derek Kan, *Transportation.gov*, US Dept. of Transportation, 12 Feb. 2018, https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/automated-vehicles/304471/roundtable-data-automated-vehicle-safety-report.pdf; and ITS Joint Program Office, "ITS DataHub: About," *Intelligent Transportation Systems*, US Dept. of Transportation, as of 15 Mar. 2026, https://its.dot.gov/data/about.

97. "Work Zone Data Exchange Overview," *Transportation.gov*, US Dept. of Transportation, 7 Nov. 2018, https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/automated-vehicles/325341/work-zone-data-exchange-overview.pdf; and "Data for Automated Vehicle Integration (DAVI)," *Transportation.gov*, US Dept. of Transportation, n.d., https://www.transportation.gov/av/data.

98. Joel Gurin, via email to Kirsten Sandberg, 31 Mar. 2026.

99. Ramanathan V. Guha, interviewed via Zoom by Kirsten Sandberg, 30 Jan. 2026.

100. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

101. Jose M. Plehn Dujowich, PhD, interviewed via Zoom by Kirsten Sandberg, 25 Nov. 2025.

102. US Securities and Exchange Commission, Form 10-K (Annual report pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934), *SEC.gov*, https://www.sec.gov/files/form10-k.pdf; SEC, Form 10-Q (Quarterly report pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934), *SEC.gov*, https://www.sec.gov/files/form10-q.pdf; and SEC, "Webmaster Frequently Asked Questions," *SEC.gov*, last updated 23 Aug. 2024, https://www.sec.gov/about/webmaster-frequently-asked-questions#reuse. These reports synthesize corporate financial performance and include audited financial statements, risk factors, disclosures about operations and governance, and management discussion and analysis.

103. Frank Nagle, interviewed via Zoom by Kirsten Sandberg, 7 Jan. 2026.

104. Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 Dec. 2022 … as regards corporate sustainability reporting, *EUR-Lex*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464; "Corporate sustainability reporting," *European Commission: Finance*, 9 Dec. 2025, https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en; OECD, *Board Responsibility and Sustainability-Related Disclosure in Asia*, OECD Publishing, Paris, 27 Nov. 2025, https://doi.org/10.1787/8d2672e7-en; and Samantha J. Rowe et al., "Worldwide Sustainability Reporting: Current State of Play," *Debevoise & Plimpton Insights*, Debevoise & Plimpton LLP, 24 Mar. 2026, https://www.debevoise.com/insights/publications/2026/03/worldwide-sustainability-reporting-current-state.

105. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

106. Michael Dolan, interviewed via Zoom by Kirsten Sandberg, 15 Jan. 2026.

107. "Kernel Overview," Android Open Source Project, as of 16 Jan. 2026, https://source.android.com/docs/core/architecture/kernel; Evan Ackerman, "How NASA Designed a Helicopter That Could Fly Autonomously on Mars," *IEEE Spectrum*, 17 Feb. 2021, https://spectrum.ieee.org/nasa-designed-perseverance-helicopter-rover-fly-autonomously-mars; and Taylor Hill, "Meet the Open-Source Software Powering NASA's Ingenuity Mars Helicopter," *JPL News*, NASA Jet Propulsion Laboratory, 8 Jul. 2021, https://www.jpl.nasa.gov/news/meet-the-open-source-software-powering-nasas-ingenuity-mars-helicopter/.

108. The Linux kernel is available under GNU General Public License version 2.0 (GPL-2.0), https://docs.kernel.org/process/license-rules.html.

109. Members, LF AI & Data Foundation, https://lfaidata.foundation/about/members/. For more on precompetitive collaborations, see Institute of Medicine Roundtable on Translating Genomic-Based Research for Health, "Requisites for Successful Precompetitive Collaboration," *Establishing Precompetitive Collaborations to Stimulate Genomics-Driven Product Development: Workshop Summary* (Washington DC: National Academies Press; 2011). https://www.ncbi.nlm.nih.gov/books/NBK54320/; and Jeffrey S. Barrett, "The Precompetitive Space for Drug or Vaccine Development: What Does It Look Like Now and What Could It Look Like in the Future?" *Journal of Pediatric Pharmacology and Therapeutics* 28, no. 5 (2023): 465–472, https://pmc.ncbi.nlm.nih.gov/articles/PMC10731930/#.

110. Michael Dolan, interviewed via Zoom by Kirsten Sandberg, 15 Jan. 2026. See *Feist Publications Inc. v. Rural Tel. Service Co.*, 499 U.S. 340 (1991), https://www.loc.gov/resource/usrep.usrep499340/.

111. "Who is Overture for?" Overture Maps Foundation, as of 15 Feb. 2026, https://overturemaps.org/.

112. Frank Nagle, interviewed via Zoom by Kirsten Sandberg, 7 Jan. 2026.

113. "Why Host Your Project at the Linux Foundation?" The Linux Foundation, 2023, https://www.linuxfoundation.org/hubfs/Why%20Host%20a%20Project.pdf.

114. Nick Hart, interviewed via Google Meet by Kirsten Sandberg, 19 Feb. 2026.

115. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026.

116. A phrase quoted by Frank Nagle, advising chief economist at the Linux Foundation and a research scientist at the Initiative on the Digital Economy at the Massachusetts Institute of Technology. Frank Nagle, interviewed via Zoom by Kirsten Sandberg, 7 Jan. 2026. See also https://ide.mit.edu/people/frank-nagle/.

117. Friedrich Lindenberg, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

118. Frank Nagle, interviewed via Zoom by Kirsten Sandberg, 7 Jan. 2026. See also "Meta Transitions PyTorch to the Linux Foundation, Further Accelerating AI/ML Open Source Collaboration," press release, The Linux Foundation, 12 Sep. 2022, https://www.linuxfoundation.org/press/press-release/meta-transitions-pytorch-to-the-linux-foundation.

119. Daniel Yue and Frank Nagle, "Igniting Innovation: Evidence from PyTorch on Technology Control in Open Collaboration," Harvard Business School Working Paper No. 25-013, 10 Sep. 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4960578.

120. Frank Nagle, interviewed via Zoom by Kirsten Sandberg, 7 Jan. 2026.

121. Please find Wikimedia Foundation's mission here, as of 27 Feb. 2026, https://wikimediafoundation.org/who-we-are/mission/; and its principles here, last revised 16 Sep. 2025, https://en.wikipedia.org/wiki/Wikipedia:Principles.

122. Lane Becker, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

123. Rich Skrenta, interviewed via Zoom by Kirsten Sandberg, 3 Feb. 2026.

124. See the landmark case of US Department of Justice, *Reporters Committee for Freedom of the Press v. Department of Justice*, 489 US 749 (1989), https://supreme.justia.com/cases/federal/us/489/749/; and David S. Ardia, "Privacy and Court Records: Online Access and the Loss of Practical Obscurity," *University of Illinois Law Review* 2017, no. 5 (4 Aug. 2017): 1385–1454, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3013704.

125. Mohammad Amir Anwar, "Africa's Data Workers Are Being Exploited by Foreign Tech Firms: Four Ways to Protect Them," *The Conversation*, The Conversation US Inc., 31 Mar. 2025, https://theconversation.com/africas-data-workers-are-being-exploited-by-foreign-tech-firms-4-ways-to-protect-them-252957; and Michelle Du and Chinasa T. Okolo, "Reimagining the Future of Data and AI Labor in the Global South," Brookings Institution, 7 Oct. 2025, https://www.brookings.edu/articles/reimagining-the-future-of-data-and-ai-labor-in-the-global-south/.

126. Nick Hart, interviewed via Google Meet by Kirsten Sandberg, 19 Feb. 2026.

127. See US Bureau of Economic Analysis, "The Making of GDP," *BEA in Brief*, US Dept. of Commerce, rev. 2 Apr. 2020, https://www.bea.gov/sites/default/files/2020-04/BEA-in-Brief-GDP-Final.pdf; and US Bureau of Economic Analysis, "Input-Output Accounts: Who Sells What to Whom," *BEA in Brief*, US Dept. of Commerce, 23 Feb. 2021, https://www.bea.gov/sites/default/files/2021-02/BEA-Input-Output-Accounts.pdf.

128. Nick Hart, interviewed via Google Meet by Kirsten Sandberg, 19 Feb. 2026.

129. Nick Hart, interviewed via Google Meet by Kirsten Sandberg, 19 Feb. 2026.

130. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

131. Friedrich Lindenberg, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

132. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026.

133. Center for Open Data Enterprise, "CODE's Flood Resource Hub: A Guide for Flood Information Products," *Open Data Enterprise Blog*, 14 Jun. 2023, https://odenterprise.medium.com/codes-flood-resource-hub-a-guide-for-flood-information-products-8ed26b36273d; and Center for Open Data Enterprise, "Flood Resource Hub," *Airtable*, as of 15 Mar. 2026, https://airtable.com/app7R95PbZ6EDQDUn/shrvJvJPxBYwTmpo5/tbliIyvW01TuPdnUi/viwTJ0UJ86nLjuqPQ.

134. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026.

135. Stormy Peters and Nithya Ruff, "Participating in Open Source Communities," *Open Source Guides*, The Linux Foundation, n.d., https://www.linuxfoundation.org/resources/open-source-guides/participating-in-open-source-communities, as of 1 Mar. 2026.

136. See, for example, "Contributor Guide," The Kubernetes Authors, https://www.kubernetes.dev/docs/guide/; and "Special Interest Groups," Community Groups, The Kubernetes Authors, https://www.kubernetes.dev/community/community-groups/, as of 1 Mar. 2026.

137. "Why Host Your Project at the Linux Foundation?" The Linux Foundation, 2023, https://www.linuxfoundation.org/hubfs/Why%20Host%20a%20Project.pdf.

138. Rich Skrenta, interviewed via Zoom by Kirsten Sandberg, 3 Feb. 2026.

139. Internet Engineering Task Force, "RFC 9309: Robots Exclusion Protocol," *RFC Editor*, Sep. 2022, https://www.rfc-editor.org/rfc/rfc9309.html; US Copyright Office, "Digital Millennium Copyright Act," https://www.copyright.gov/dmca/; European Parliament and Council of the European Union, "Right to be Forgotten," https://gdpr.eu/right-to-be-forgotten/; "General Data Protection Regulation," *EUR-Lex*, as of 4 May 2016, https://eur-lex.europa.eu/eli/reg/2016/679/oj; and Rich Skrenta, "Setting the Record Straight: Common Crawl's Commitment to Transparency, Fair Use, and the Public Good," *Common Crawl Blog*, Common Crawl Foundation, 4 Nov. 2025, https://commoncrawl.org/blog/setting-the-record-straight-common-crawls-commitment-to-transparency-fair-use-and-the-public-good.

140. Lane Becker, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

141. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

142. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

143. LEI Statistics, as of 3 Mar. 2026, https://www.gleif.org/en/lei-data/global-lei-index/lei-statistics. See also the Global LEI System Statistics Dashboard, https://www.gleif.org/assets/components/global-lei-system-statistics-dashboard/tableau-dashboard.html.

144. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

145. Jane Gavronsky, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

146. Sara L. Jordan, "Climate and Economic Justice Screening Tool: Frequently Asked Questions," *Biden White House Archives*, Council on Environmental Quality and the White House, Feb. 2022, rev. 14 Mar. 2022, https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/02/CEQ-CEJST-QandA.pdf.

147. Public Environmental Data Partners, "Climate and Economic Justice Screening Tool," *CEJST 2.0*, GitHub.io, n.d., https://public-environmental-data-partners.github.io/j40-cejst-2/en/#3/33.47/-97.5.

148. Joel Gurin, interviewed via Zoom by Kirsten Sandberg, 5 Feb. 2026.

149. "About OpenSanctions," as of 2 Mar. 2026, https://www.opensanctions.org/docs/about/.

150. Shayne Longpre et al., "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing and Attribution in AI," arXiv, 25 Oct. 2023, https://arxiv.org/abs/2310.16787.

151. Shayne Longpre et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons," arXiv, 24 Jul. 2024, https://arxiv.org/pdf/2407.14933.

152. Shayne Longpre et al., "A Large-Scale Audit of Dataset Licensing and Attribution in AI," *Nature Machine Intelligence* 6 (30 Aug. 2024): 975–987, https://doi.org/10.1038/s42256-024-00878-8.

153. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

154. "17 U.S.C. § 101 (Definitions)," Legal Information Institute, Cornell Law School Legal Information Institute, https://www.law.cornell.edu/uscode/text/17/101.

155. "17 U.S.C. § 105 (Subject Matter of Copyright: US Government Works)," Legal Information Institute, Cornell Law School Legal Information Institute, https://www.law.cornell.edu/uscode/text/17/105.

156. Prem Ramaswami, interviewed via Zoom by Kirsten Sandberg, 2 Feb. 2026.

157. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

158. See, for example, Overture Maps' "Attribution and Licensing" page, last updated 21 Jan. 2026. https://docs.overturemaps.org/attribution/.

159. Lane Becker, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

160. Creative Commons Attribution-ShareAlike 4.0 License Deed, as of 10 Apr. 2026, https://creativecommons.org/licenses/by-sa/4.0/deed.en.

161. Smallbones, "Knowledge Manipulation on Ruwiki, the Russian Wikipedia Fork," in *Wikimedia Research Newsletter* 15(7) Jul. 2025, lasted edited 26 Jul. 2025, https://meta.wikimedia.org/wiki/Research:Newsletter/2025/July#Knowledge_manipulation_on_Ruwiki,_the_Russian_Wikipedia_fork; Yiriy Marin, "How Russia Is Creating Its Own Alternative to Wikipedia," *Russia Post*, 25 Jun. 2024, https://russiapost.info/society/alt_wikipedia; "Russian Wikipedia's Top Editor Leaves to Launch a Putin-Friendly Clone," *Bloomberg*, Bloomberg LP, 12 Jul. 2023, https://www.bloomberg.com/news/articles/2023-07-12/russian-wikipedia-editor-leaves-to-launch-a-putin-friendly-clone; and Sergey Volkov, "At the Russian Internet Forum, Wikimedian Vladimir Medeiko presented an alternative to the Russian Wikipedia," News, *World of Encyclopedias*, 24 May 2023, https://www.encyclopedia.ru/news/enc/detail/82756/.

162. Tilman Bayer, "Comparing Comparisons of Grokipedia vs. Wikipedia by Three Different Research Teams," in Wikimedia Research Newsletter 15(12) Dec. 2025, last edited 3 Feb. 2026, https://meta.wikimedia.org/wiki/Research%3ANewsletter/2025/December; Reece Rogers, "Elon Musk's Grokipedia Pushes Far-Right Talking Points," WIRED, 27 Oct. 2025, https://www.wired.com/story/elon-musk-launches-grokipedia-wikipedia-competitor/; Harold Triedman and Alexios Mantzarlis, "What Did Elon Change? A Comprehensive Analysis of Grokipedia," CorarXiv:2511.09685, 12 Nov. 2025, https://arxiv.org/pdf/2511.09685; and David Ingram, "Elon Musk's Grokipedia Cites Neo-Nazi Website 42 Times—Study," NBC News, 20 Nov. 2025, https://www.nbcnews.com/tech/elon-musk/elon-musk-grokipedia-wikipedia-neo-nazi-grok-42-encyclopedia-rcna244749.

163. Lane Becker, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

164. Jeff Jonas, interviewed via Zoom by Kirsten Sandberg, 30 Jan. 2026.

165. "About," *Movement Strategy*, Wikimedia Meta-Wiki, last edited on 25 Jun. 2024, https://meta.wikimedia.org/wiki/Movement_Strategy/About.

166. "Increase the Sustainability of Our Movement," *Movement Strategy*, Wikimedia Meta-Wiki, last edited on 18 Jan. 2026, https://meta.wikimedia.org/wiki/Movement_Strategy/Recommendations/Increase_the_Sustainability_of_Our_Movement; and "Mission," Wikimedia Foundation, as of 27 Feb. 2026, https://wikimediafoundation.org/who-we-are/mission/.

167. "Revenue Streams," *Movement Strategy*, Wikimedia Meta-Wiki, last edited 21 Apr. 2023, https://meta.wikimedia.org/wiki/Movement_Strategy/Recommendations/Iteration_2/Revenue_Streams/1-8; and "Wikimedia Enterprise," Wikimedia Meta-Wiki, last edited 24 Feb. 2026, https://meta.wikimedia.org/wiki/Wikimedia_Enterprise.

168. "Wikimedia Foundation launches Wikimedia Enterprise," *News*, Wikimedia Foundation, 25 Oct. 2021, https://wikimediafoundation.org/news/2021/10/25/wikimedia-foundation-launches-wikimedia-enterprise-the-new-opt-in-product-for-companies-and-organizations-to-easily-reuse-content-from-wikipedia-and-wikimedia-projects/; "About Wikimedia Enterprise," Wikimedia Enterprise, as of 26 Feb. 2026, https://enterprise.wikimedia.com/api/.

169. Lane Becker, interviewed via Zoom by Kirsten Sandberg, 4 Feb. 2026.

170. Emily Bennett and Anusha Dandapani, interviewed via Zoom by Kirsten Sandberg, 10 Feb. 2026.

171. Eugenio Cerutti et al., *The Global Impact of AI: Mind the Gap*, IMF Working Paper No. 25/76 (Washington, DC: International Monetary Fund, 8 Apr. 2025), https://www.imf.org/-/media/files/publications/wp/2025/english/wpiea2025076-print-pdf.pdf.

172. Emily Bennett and Anusha Dandapani, interviewed via Zoom by Kirsten Sandberg, 10 Feb. 2026.

173. Anusha Dandapani, Shambhavi Mohan, Devyani Rastogi, and Andres Fortino. *Responsible AI Innovation for Social Impact: A Case Study in Multilingual Media Moderation*. United Nations International Computing Centre and New York University School of Professional Studies, 8 Sep. 2025, https://www.unicc.org/wp-content/uploads/2025/09/UNICC-White-Paper_Responsible-AI-Innovation-for-Social-Impact-1.pdf.

174. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

175. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

176. "Shipping first time code is like going into debt. A little debt speeds development so long as it is paid back promptly with a rewrite," said Ward Cunningham, computer programmer and inventor of the wiki. "The danger occurs when the debt is not repaid. Every minute spent on not-quite-right code counts as interest on that debt. Entire engineering organizations can be brought to a stand-still under the debt load of an unconsolidated implementation." See Ward Cunningham, "The WyCash Portfolio Management System," *OOPSLA '92 Experience Report*, ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (26 Mar. 1992), C2.com, https://c2.com/doc/oopsla92.html.

177. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

178. Paul A. David and Francesco Rullani, "Dynamics of Innovation in an 'Open Source' Collaboration Environment: Lurking, Laboring, and Launching FLOSS Projects on SourceForge," *Industrial and Corporate Change* 17, no 4 (Aug. 2008): 647–710, https://doi.org/10.1093/icc/dtn026; Sheen Levine and Michael Prietula, "Open Collaboration for Innovation: Principles and Performance," *arXiv*, 8 Jan. 2014, https://arxiv.org/pdf/1406.7541; Xiaohong Chen and Yuan Zhou, "Open-Source Collaboration and Technological Innovation in the Industrial Software Industry: A Multi-Case Study," *Systems* 13, no. 6 (3 Jun. 2025): 433, https://www.mdpi.com/2079-8954/13/6/433; Hongbo Fang, et al., "Weak Ties Explain Open Source Innovation," *arXiv*, 29 Oct. 2025, https://arxiv.org/pdf/2411.05646.

179. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

180. Amy Rose, interviewed via Zoom by Kirsten Sandberg, 22 Dec. 2025.

181. US General Services Administration, "Unique Entity ID Is Here," *Integrated Award Environment Systems Information Kit*, last updated 23 Jul. 2025, https://www.gsa.gov/about-us/organization/federal-acquisition-service/fas-initiatives/integrated-award-environment/iae-systems-information-kit/unique-entity-id-is-here; and Office of Management and Budget, "Universal Identifier and System for Award Management," 2 C.F.R. pt. 25, *Electronic Code of Federal Regulations*, 14 Sep. 2010, https://www.ecfr.gov/current/title-2/subtitle-A/chapter-I/part-25.

BrightQuery delivers the most comprehensive, government-sourced view of global businesses, with coverage spanning 324 million organizations, 512 million locations, and 1.2 billion contacts across 222 countries. Specializing in private company data, BrightQuery sources information from over 100,000 federal, state, and local government agencies through official public filings. BrightQuery serves as the lead government data partner for the US National Secure Data Service (NSDS) and is a primary government data supplier for leading AI companies. BrightQuery is also a member of many AI and open source communities including **AI Alliance**, **Linux Foundation**, **FINOS**, **Overture Maps Foundation**, **Agentic AI Foundation**, **Computing Research Association**, **EDM Council**, and **Data Foundation**.

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. By leveraging project databases and networks and committing 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 organisations worldwide.

x.com/linuxfoundation

facebook.com/TheLinuxFoundation

linkedin.com/company/the-linux-foundation

youtube.com/user/TheLinuxFoundation

github.com/LF-Engineering

To reference this work, please cite as follows: Kirsten D. Sandberg, "The Value of Open Data on Global Entities: Turning the World's Largest Open Entity Graph into a Catalyst for Collaboration, Innovation, and Transformation," foreword by Gabriele Columbro, The Linux Foundation, June 2026.