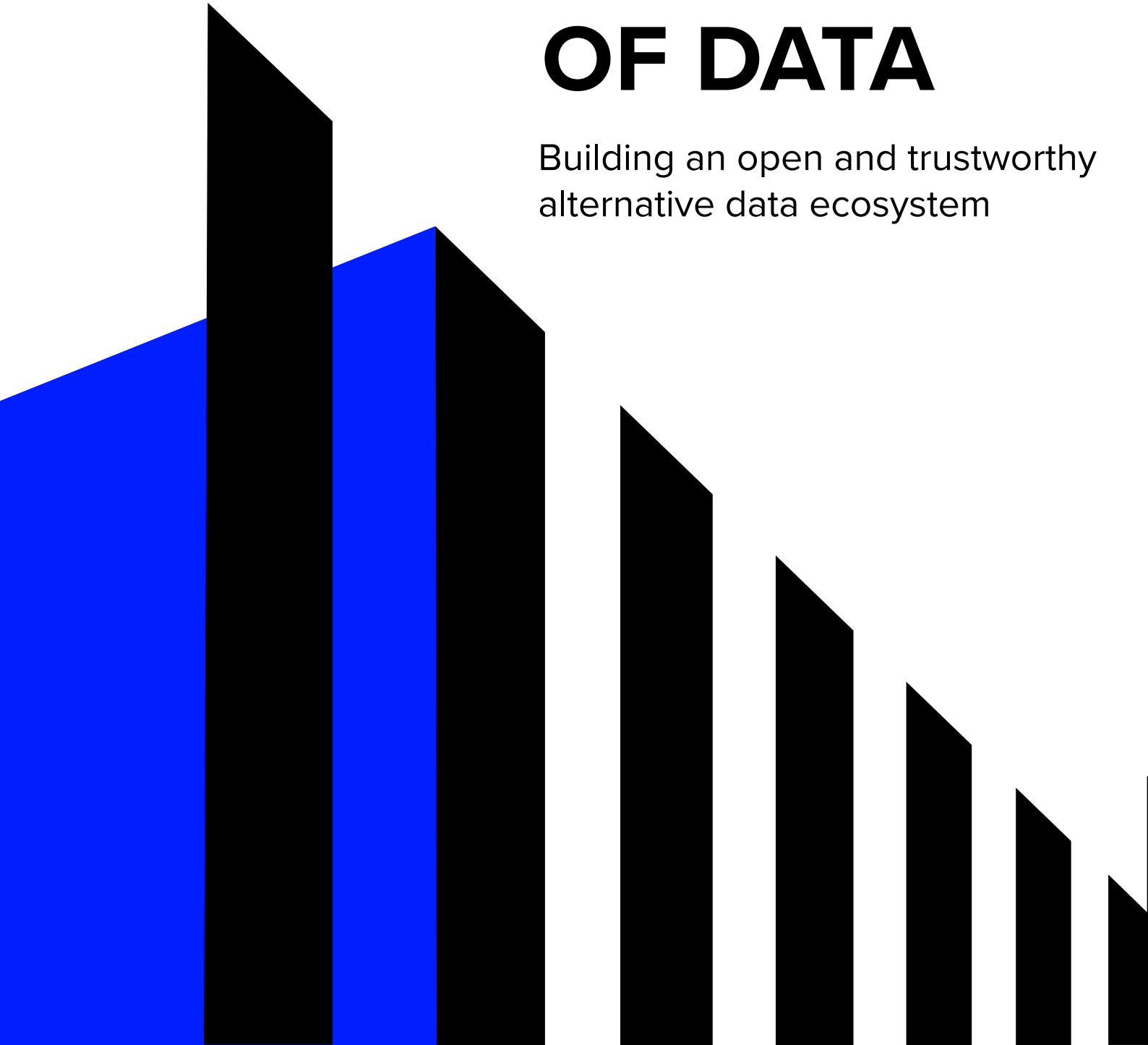


A NEW DIMENSION OF DATA

Building an open and trustworthy
alternative data ecosystem



Contents

Introduction	4
About this report	5
What is “alternative data”?	5
How is alt data currently being accessed, used and shared?	6
Alt data from individuals	6
Alt data from business processes	7
Alt data from sensors and satellite imagery	7
Ethical and legal considerations in alt data	8
Privacy and personal data	8
Data rights and licensing	10
Fairness and material nonpublic information	11
Improving access to alt data	12
The challenges of standardizing alt data	12
Standards in the market today	13
Increasing adoption of standards	14
Recommendations and next steps	16
Recommendations for alt data providers	17
Recommendations for alt data users	17
Recommendations for both alt data providers and users	17
Appendix: Contributors	18
References	19



Introduction

Alternative data, or “alt data,” has been commonly discussed in the offices of hedge funds and other investment firms for years, but has recently seen a spike in popularity and interest in the broader community. Its definition is unclear.

Some see it as a new approach to creating investment insights and value from exhaust data, generated as a side effect of other operations and transactions. To others, it is simply a new element of the buzz about big data in investment research.

For many years, analysts have relied on primary research as a method of collecting data that could identify when a stock was potentially mispriced. From using humans counting cars in parking lots of major retailers, to physical inspection of store shelves, these primary research practices first emerged in the coverage of retail stocks. With the rise of online commerce, the proliferation of cellular technology, satellite imagery and the Internet of Things (IoT), it is now possible to track human activity and commerce on a massive scale. This data can now be processed, cleaned and used to create insights that give investors a better understanding of how firms are performing and make decisions based on well-informed predictions. However, the ability to collect and use data on this scale comes with additional risks to privacy, and potential for public fear about the actual or perceived harms from using this data.

Market research estimates that hundreds of investment firms are already using alt data to some extent. In a recent global artificial intelligence/machine learning (AI/ML) survey conducted by Refinitiv of the top financial institutions using AI/ML today, 70% of firms are using alt data. The U.S. leads the way with 97% adoption. Adoption in Europe (67%) and Asia (53%) is further behind but continues to increase.¹

Hundreds of new data providers have entered the alt data space, hoping to build a long-term business selling data

to financial services companies, while existing financial information companies have introduced their own offerings, looking to capitalize on the emergence of this new segment of the market.

With over a thousand purported alt data sources, hundreds of case studies and millions in global spend, the alt data sector seems to be flourishing. Despite its growth, however, the market remains a very small part of the multibillion dollar market for access to stock market and other financial data. While the governance and norms for the licensing and distribution of traditional data are well established in many countries, the alt data market is at an early stage. Best practices, codes of practice, regulations and standards have not yet been fully established.

Enabling the integration and use of such a disparate set of data sources will require the development of a new set of standards. The alt data industry also needs to adopt codes of practice: ethical and legal frameworks that will build trust in how data is being accessed, used and shared.

These standards and best practices should be developed by the alt data industry in collaboration with regulators. Without taking these steps, users of these new data sources may be exposing themselves to a range of legal risks around the use of personal data, or material nonpublic breach of information. Adoption of these standards will help all alt data industry participants and ensure long-term sustainability of alt data use in financial services.

This report provides a number of recommendations for both alt data providers and users, with the aim of helping to create a more open, trustworthy data ecosystem.

About this report

This report was jointly produced by the Open Data Institute (ODI) and Refinitiv.

Founded in 2012, the ODI is an international, independent and not-for-profit organization based in London, UK. The ODI works with companies and governments to build an open, trustworthy data ecosystem, where people can make better decisions using data and manage any harmful impacts.²

Refinitiv is a new company built on a unique open platform, high-performance products and best-in-class data. In the face of unparalleled industry change, Refinitiv draws on its deep knowledge and heritage of objectivity to drive

performance and innovation with customers and partners.³

This report was written using extensive desk research, user research interviews with 13 professionals in the alternative data space and an expert roundtable of 12 participants. A list of contributors is in the appendix. The views expressed are those of the contributors and do not necessarily reflect the views of the organizations. This article is for general information purposes and is not intended to be and should not be relied upon or taken as legal advice.

What is “alternative data”?

Alternative data is a term increasingly being used in the finance and investment sectors. As a relatively new and changing concept, there is currently no definition that is universally agreed upon.

AlternativeData.org defines alternative data as “*data used by investors to evaluate a company or investment that is not within their traditional data sources (financial statements, SEC filings, management presentations, press releases, etc.)*.”⁴

While some definitions of alt data focus on new data sources, traditional data sets could still be considered alt data if the source or method of analysis was new, as was suggested on a panel at the Financial Management Association Conference in 2018. For example, the use of credit card data to inform investments is now common practice in the industry, but it is still labeled as “alt data.”

J.P. Morgan states that, “*The definition of alternative data can also change with time. As a data source becomes widely available, it becomes part of the financial mainstream and is often not deemed alternative.*”⁵

“Alternative data” then is largely a sector-specific term that relates to the use of new data sources, combining existing data sets together to create new insights and the application of new analytical techniques. Data sources that provide “traditional data” in the agriculture or retail sectors are considered to be “alt data” in the context of finance and investing. Simply put, alternative data is data that is commonly used and analyzed in a certain domain, put to a different or new use.

How is alt data currently being accessed, used and shared?


ALT DATA FROM INDIVIDUALS

Alt data from individuals is data generated from online consumer activity that might help inform investment decisions. This category of alt data generally includes social media data (e.g., Twitter®, LinkedIn®, blogs); data from specialized sites (e.g., news media, product reviews); and Web searches and volunteered personal data (e.g., Google® search, email receipts).⁶ Users of social media platforms agree to their terms and conditions but few understand the extent to which data about them on these platforms is monetized. Growing awareness of this use could increase the [distrust in social media](#) that has already been observed across the UK and Europe.⁷ It is not clear whether all of these uses would be legal under the various legislative environments that exist around the world.

Social media can provide a wealth of insights into consumer behavior. An example of an alt data provider in this space is the social media API aggregation company Gnip, which was purchased by Twitter in 2014.⁸ Gnip now provides an enterprise API for Twitter that offers a single interface to access Twitter, Facebook®, YouTube, Flickr, Google Buzz, Vimeo and more.

Companies like iSentium collect real-time sentiment data from Twitter. By analyzing positive or negative social media posts, investors aim to judge company performance before any traditional data is released from the company. From this, J.P. Morgan constructed an index on the [S&P 500](#) called the JPUSISEN Index based on information obtained from iSentium. Through historical analysis, J.P. Morgan claims that the JPUSISEN Index would have provided a 13.7% return annually since 2013, 1.6% more than the S&P return of 12.1%.⁹

From a legal and ethical standpoint, given the prevalence of personal data in this category, social media is one of the riskier types of alt data to work with. Without proper privacy and security practices in place, issues around reidentification and identity theft will exist. This topic will be explored later in the paper.



Business process data, such as consumer transactions, may contain personal data. This means that appropriate privacy and security measures are needed to prevent harm and steps taken to ensure it is used in a trustworthy way.

ALT DATA FROM BUSINESS PROCESSES

Alt data from business processes is data generated from the typical activities of organizations that is not directly related to investing behavior. This category of alt data generally includes “*data made available by public agencies (e.g., federal and state governments); commercial transactions (including e-commerce and credit card spending, exchange transaction data); and data from other private agencies (e.g., industry-specific supply chain data).*”¹⁰

Customer transaction data can be a very strong predictor of company performance. Companies involved in payments, such as credit card network companies or point-of-sale terminal companies, can provide this data to investment firms.

Eagle Alpha has created a model called RevCast which incorporates alternative and traditional data to make forecasts on company performance, including consumer transaction data, online search data and historical financials. By combining these data sets, the RevCast model forecasted the car rental company Hertz would have second quarter 2018 revenues of USD 2.45 billion, differing from the market consensus estimate of USD 2.3 billion. A month later, Hertz reported revenues of USD 2.4 billion, a figure that was twice as close to the predicted number than the market consensus produced.¹³

Business process data, such as consumer transactions, may contain personal data. This means that appropriate privacy and security measures are needed to prevent harm and steps taken to ensure it is used in a trustworthy way.

Another major issue regarding business process data is the debate around what is considered to be public information and the social, legal and ethical issues of repurposing data that has been shared for other reasons. Scraping data from public websites, observing store and factory activity, and the use of employee information are all examples of where this tension exists.

Companies like Matchdeck use publicly shared personal data to create products based on analyzing online trends in organizations, such as employee growth and executive turnover, and correlating with business performance. Individuals are able to have their profiles removed from these products, and Matchdeck does not crawl password-protected sites, but there is still a lack of understanding in the market regarding these practices.

ALT DATA FROM SENSORS AND SATELLITE IMAGERY

Alt data in this category includes data generated from smartphones, portable electronic devices, satellite data, geolocation data as well as data from other sensors and “smart” or networked devices.

The use of satellite and other types of geospatial data for investing is a growing trend in finance. The ODI captured several use cases in a recent paper on geospatial data infrastructure, such as using aggregating activity data via apps in smartphones and GPS devices and matching the points against a street to indicate footfall.¹² The increasing accuracy and timeliness of satellite data collected by commercial organizations also provides the ability to confirm real-time data such as crop harvests, port usage, parking lot occupancy and factory production. Previously, this type of data was collected by having people physically observe and report on activity in ports, factories and stores, but machine learning is allowing insights to be extracted from satellite images.

The company SpaceKnow has a product that uses satellite image processing technology originally intended for agricultural use to monitor the level of Chinese manufacturing. Through algorithmic imagery analysis, SpaceKnow created the China Satellite Manufacturing Index (SMI) to compete with the current state-run indices – the China Purchasing Managers Index (PMI) and the Caixin PMI. The two PMIs are created by using survey data from managers, collected by the National Bureau of Statistics of China. SpaceKnow uses satellite imagery data sets consisting of over two billion observation points to create an index that they claim more accurately predicts Chinese manufacturing for investors. Analysis of historical data over a 10-year period shows a very strong correlation to both PMIs.¹³

Sensor data collected from personal devices might be useful for measuring footfall in cities and retail districts, but as with other sensitive personal data, it presents significant privacy challenges. Similar to the scraping of data from websites, the ability to remotely observe commercial facilities by satellites raises ethical questions around what is public information and what might be considered trespassing.

The alt data use cases above showcase the variety of ways data can be sourced and used to provide financial insights. A common feature among the examples is the existence of ethical, and possibly legal, challenges. Organizations in the alt data market need to pay attention to the ethical use of data in their investing decisions to avoid causing societal harm and avoid legal, financial and reputational risks.

Ethical and legal considerations in alt data

Faced with public criticism and debate around the use of data, the individual practitioners and organizations involved in collecting, sharing and working with data are exploring the ethics of their practices. Data ethics is a branch of ethics that evaluates data practices with the potential to adversely impact not only individuals, but businesses and wider society too.¹⁴ While the debate around data ethics often centers on the impacts of the use of data about individuals, it also applies to our growing ability to remotely monitor business operations.

Organizations are using emerging tools, such as the ODI's [Data Ethics Canvas](#), to help identify potential ethical issues associated with a data project or activity. The canvas promotes understanding and debate around the foundation, intention and potential impact of any piece of work, and helps identify the steps needed to act ethically.¹⁵

Trust is equally essential for any organization and for the alt data sector as a whole. When trust in organizations breaks down, they may suffer reputational damage which can lead to loss of business. When trust in a sector as a whole decreases, there is a danger that it cannot realize the full benefits that innovative use of data could bring. Data might not be collected, shared or used to the extent it could be because of concerns it may be misused. Individuals withdrawing consent could lead to data that is biased and misleading. Countries could introduce regulation that limits data collection or use, or makes it significantly more burdensome, presenting challenges to the existing data market.¹⁶

There are three areas where we feel that the alt data ecosystem needs to focus attention to create an open, trustworthy data ecosystem that benefits everyone: privacy, rights and fairness.

PRIVACY AND PERSONAL DATA

One of the most controversial aspects of the use of alt data in investing revolves around the collection and use, intentionally or otherwise, of personal data. Even if not damaging to the people the data is about, breach of data privacy laws can be financially destructive to firms. In the European Union (EU), the new General Data Protection Regulation (GDPR) legislation stipulates that noncompliant businesses can be fined up to €20 million or up to 4% of their annual worldwide turnover of the preceding financial year, whichever is greater.¹⁷

Other activity, although legal, can cause reputational damage through negative press coverage and upset people. Much of this harm comes from the difference between how people expect data will be used and the reality of data sharing between businesses. People may be happy with personal data being used for benefits to society, [as noted in a recent poll](#),¹⁸ but may not want that data being used to assist hedge funds' investment decisions.¹⁹ Being open and transparent when using, accessing and sharing data helps avoid the backlash that comes when unexpected data use gets revealed. In addition to this, assessing the ethical use involves engagement with those who might be affected by the use of data.

With scandals such as Facebook and Cambridge Analytica²⁰ and the implementation of the GDPR and the California Consumer Privacy Act (CCPA) of 2018, personal

data came to the forefront of public and media discussion. Though most of the press regarding alt data is not negative, there has been growing scrutiny of the risks, with groups such as Big Brother Watch, Privacy International and Open Rights Group identifying serious privacy and human rights risks.²¹

These concerns are not unfounded. Sentiment data from social media, credit card transactions from retail stores and geolocation data from sensors can all be used to identify people and it can scare people to think they are being watched. Personal data needs to be properly anonymized, based on a reidentification risk assessment, when shared with or used by alt data providers. If we are to retain trust, it must be clear to people what is happening to data about them, and both regulators and civil society need to be able to scrutinize the claims.

A high-profile example of this potential abuse of alt data occurred in January 2019, when a scandal concerning the use of people's location data occurred between IBM's The Weather Channel and the City of Los Angeles. At the time of this writing, the municipal government is suing IBM for using location data gathered by The Weather Channel mobile application for commercial purposes despite telling users it was only for localized weather services.²² IBM have denied this claim.²³

For both commercial and compliance reasons, users of alt data do not want to access personal data. They want to use aggregated data that may provide insights into overall market trends.²⁴ Often, to unlock value from the source data sets, alt data providers and users need to combine data sets from multiple sources. This creates potential linkage risks, including the reidentification of individuals, despite efforts to anonymize the data sets. Currently, alt data providers are selling many data sets that can easily be reidentified, such as geolocation data, with only the name removed. Using a mixture of suppression, generalization and disruption could improve the anonymization in data sets in the market. It is also important for data providers to understand that anonymization is an in-depth process, involving research, legal and ethical considerations, risk analysis and testing.²⁵ Adoption of standards and best practices to improve anonymization, such as those found in the UK Anonymisation Network's [Anonymisation Decision-Making Framework](#),²⁶ or the ODI guide to [Anonymisation and open data](#),²⁷ and to ensure consideration of the ethical impacts of using personal data, will be important as the alt data ecosystem matures.

DATA RIGHTS AND LICENSING

A significant legal concern regarding alt data is the right to access, use and share the data. A robust data ecosystem needs clear data rights so that companies can operate without high legal or operational risks, while minimizing harm to society. Alt data providers need to be clear about their rights to collect and repackage data. Alt data users need to be clear about the rights that govern the data supplied to them and what they can legally do with it. They also need certainty that their supply of alt data will not be interrupted, for example due to legal action against their alt data provider.

Web scraping, a process of extracting data from websites, is a common methodology for collecting alt data from public websites. For example, companies crawl e-commerce websites to compile data on their current inventory, pricing and product reviews. This data is then analyzed through a process called “text and data mining,” in order to “discover patterns, trends and other useful information that cannot be detected through usual ‘human’ reading.”²⁸

A lack of clear licensing or legal basis for using this data is causing concern in the alt data community, specifically in the degree to which this public information can be reused. Many in the industry believe that data collected from public websites is public data and can be collected and reused for any purpose.²⁹ Others believe that even data protected by a username and password, can be used if it is amenable to Web scraping.³⁰

The lawsuit at the forefront of this debate is *LinkedIn vs. hiQ Labs*. hiQ Labs is a data science company that uses scraped data from public LinkedIn profiles in order to develop tools to help corporate HR departments keep tabs on their workforces.³¹ LinkedIn has sued hiQ for breaching their terms of use (ToU). Although from a ToU-perspective there is a breach of terms, it is unclear if this is a breach of the law, as the Web-scraped data is currently considered public information in the United States.³² So far, the lower courts have supported hiQ and issued an injunction ordering LinkedIn to grant access, however the case is now with an appellate court.³³ The outcome of this case could set a strong precedent for Web-scraping businesses in that jurisdiction that could influence the wider market.

Legal issues aside, companies whose websites are being scraped may attempt to stop this by tightening up their terms and conditions, asking companies to cease scraping operations or by applying more sophisticated technical measures – by detecting and blocking the software that is scraping Web pages or deliberately serving them incorrect information, for example. This lack of trust inevitably leads to an arms race between scrapers and websites.

In a landscape where both legal and ethical frameworks are evolving, there are various ways to mitigate problems and build trust.

Proactive publishing of machine-readable, open and aggregate data by organizations will avoid the need for data to be scraped at all, reducing issues with the accuracy and robustness of scraped data and providing more clarity around reuse rights. Data licensing agreements can also help organizations to share data with clear reuse rights where data cannot be published more openly. The Standards Board for Alternative Investments (SBAI) data licensing agreement³⁴ is one such agreement that exists specifically for alt data.



Web scraping companies can be open with sites that are being scraped, by engaging directly with the organizations. For example, maintaining a mail trail showing that websites have been alerted to data being collected and lack of subsequent objections may be useful in a court case. Halting Web scraping, and providing easy ways for that to be requested, can also help to avoid legal issues. Negotiating ongoing access to data – after prototyping data collection using Web scraping, for example – will also help to ensure ongoing access.

Organizations like Eagle Alpha are attempting to define best practices for Web scraping to curtail negative behavior, for example by ensuring Web scraping does not harm the business of the website through direct costs from increased traffic load or direct competition.³⁵ Alt data providers relying on Web scraping are also aligning their approaches to those adopted by the larger search engines such as Google, with a view that these are more accepted behaviors and that there is legal safety in numbers.

Clearly describing the provenance of data will be important in building trust across the alt data ecosystem. The variety of ways in which data is sourced, and the use of multiple data sets to create alt data products, makes this difficult to achieve. Without clarity around how data has been collected, organizations cannot be sure that their use is compliant or understand the risk of an interruption to supply, which increases their operational risks.

Organizations in the alt data market need to collaborate to discuss potential risks and work to increase effort around provenance and rights. This will help to build confidence and trust with consumers of the data and reduce risks of additional lawsuits and potentially negative media coverage. Overall, it will help to build a stronger, more sustainable and more trustworthy data ecosystem.

FAIRNESS AND MATERIAL NONPUBLIC INFORMATION

A strong data ecosystem provides equitable access to data. Access to data and information promotes fair competition and informed markets and empowers people as consumers, creators and citizens.³⁶

The legal consideration about the use of alt data that most concerns hedge fund managers involves insider trading.

Even the accusation or investigation into insider trading can financially harm a company. Investing activity is subject to insider trading laws that stipulate that any information used for commercial gain must be publicly available to ensure fairness and competitiveness in the market. Data is material nonpublic information (MNPI) when it earns returns, is clearly not licensed for use and is acquired exclusively. It is illegal for holders of MNPI to use it to their advantage in investing, or doing similarly for others.³⁷

Differences in the legal systems between the two largest alt data markets, the U.S. and the EU, contribute to uncertainty as operations become international, but also to exclusivity issues.³⁸ So far, there have been few cases regarding insider trading and the use of alt data. The U.S. Securities and Exchange Commission (SEC) has only successfully prosecuted a single case: the case of *SEC vs. Huang* involved the use of MNPI in credit card transactions to inform investment decisions relating to an outdoor goods retailer.³⁹

Initially, insider trading was a minor concern for firms, as most online and social media data being sold is public information. However, as the data categories of alt data have expanded into credit card transaction or geospatial data, these concerns have increased.⁴⁰

U.S. regulators have been more proactive in the alt data space than their EU counterparts, where the rules for insider trading are broader, making it more difficult to prove illegal activity. Regulatory organizations such as the SEC, the Competition and Markets Authority (CMA) and others need to be at the forefront of increasing market fairness by regulating access to data, balancing this against potential consumer harms.

As highlighted in the previous section, ensuring clarity around both the provenance of data and its licensing will help to address concerns around insider trading. Data provenance will help to clearly identify sources and adoption of open licenses and/or standard data sharing agreements, and access methods will increase access to data.

Improving access to alt data

Alt data providers are building new data infrastructure for the investment sector. Data infrastructure consists of data assets such as: data sets; identifiers and registers; the standards and technologies used to curate and provide access to those data assets; the guidance and policies that inform the use and management of data assets; the organizations that govern the data infrastructure and the communities involved in contributing to or maintaining it; and those who are impacted by decisions that are made using it.

In previous sections, we highlighted the need for clearer policies and guidelines. In this section, we look at other ways to strengthen data infrastructure, specifically through the adoption of open standards.

THE CHALLENGES OF STANDARDIZING ALT DATA

[Open standards](#) for data are reusable agreements that make it easier for people and organizations to publish, access, share and use better quality data.⁴¹ Participants in our research consistently highlighted a lack of standards as an issue when consuming alt data, and reported that a lack of standards is contributing to increased costs when consuming data.

Alt data includes vast amounts of nonstandardized and unstructured data that takes time, talent and technology to properly analyze. This means that the quality and limitations of the data sets being sold in the market today are often not known by alt data users until this investment has been made.

The use of Web scraping and novel sources of data might allow alt data providers to quickly build up data sets, but this does not give any guarantees that the data is correct, has the necessary detail or coverage to inform decision making, or is free from bias.

Increasing standardization could help to mitigate some of these issues. By reducing technical integration costs by standardizing data formats, or making due diligence easier by standardizing provenance information, it may reduce the effort needed to explore innovative uses of these new data sources or assess their validity.

However, creating standards for alt data sets is challenging.

First, the sheer range of data sets being explored means there is a wide variety of potential areas for standardization. While some of these data types are well understood by the industry, others are still being explored.

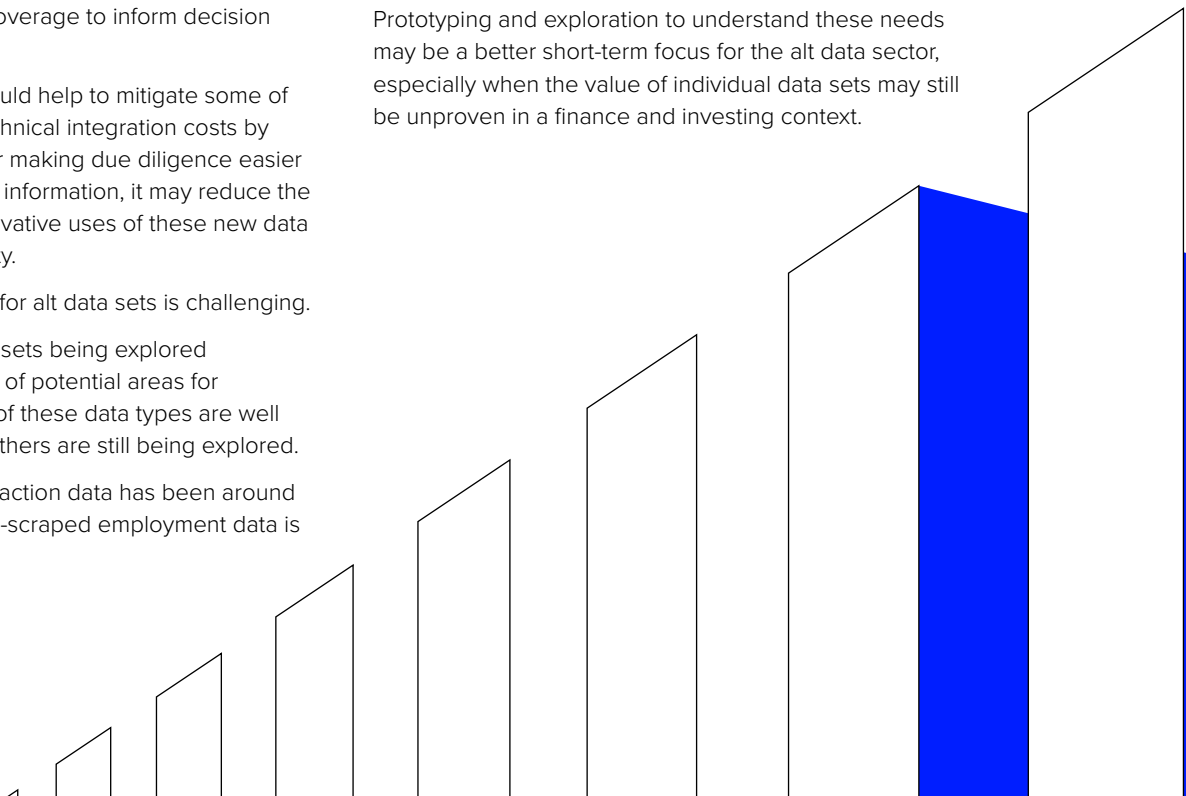
Satellite and credit card transaction data has been around for over a decade, while Web-scraped employment data is still growing in popularity.⁴²

Second, the data sets produced by alt data providers are often highly bespoke and untested in the market. Potential buyers often do not have a concrete use in mind for alt data sets and are looking to be led by vendors. The process of acquiring and exploring an alt data set, and understanding how it can be combined with existing sources, is currently quite iterative.

Finally, while standards have been highlighted as a general issue, community buy-in and support for creating or setting standards has been inconsistent. While there is interest and agreement in the need for standards, there has so far been little commitment from stakeholders.

Premature standardization may hinder the ability to explore new ways to structure and publish these data sets. As the ODI's open standards guidebook highlights, when needs are unclear, then standardization may not be the right approach.⁴³

Prototyping and exploration to understand these needs may be a better short-term focus for the alt data sector, especially when the value of individual data sets may still be unproven in a finance and investing context.



STANDARDS IN THE MARKET TODAY

Despite these challenges, the alt data sector has been making some steps towards creating and adopting standards. The initial steps have focused primarily on the processes supporting due diligence and data acquisition, with trial data agreements and due diligence questionnaires (DDQs)⁴⁴ being the major focus.

There are several organizations that are trying to coordinate broader activities around standards for alt data. These include the UK-based Standards Board for Alternative Investments (SBAI), and U.S.-based Financial & Information Services Association (FISD) and Investment Data Standards Organization (IDSO).

The SBAI is “the custodian of the standards and brings investors, managers and regulators together to collaboratively improve the standards.”⁴⁵ Their standards in the alt data space are more related to conduct rather than technical standards, focusing on disclosure, valuation, risk management, fund governance and shareholder management. Their main contribution to the alt data ecosystem so far has been the Standardised Trial Data License Agreement, published in collaboration with Eagle Alpha. The license agreement has been designed to speed up the process of evaluating data sets before purchasing.

A standardized agreement will lower the risk on trialing data and improve the efficiency of the current trial negotiation process. This will reduce the time spent in these processes, ideally allowing for a more efficient industry and a higher level of innovation through more rapid testing processes.⁴⁶ SBAI is focusing on standardizing one aspect of reusing third-party data sets, as opposed to more technical specifications around file formats and schemas.

The FISD is “the global forum of choice for industry participants to discuss, understand and facilitate the evolution of financial information for the key players in the value chain including consumer firms, third-party groups and data providers.”⁴⁷ FISD developed the Market Data Definition Language (MDDL) metadata standard⁴⁸ as an open industry standard for securities market data⁴⁹ that is considered for use with alt data. MDDL is already used for financial instruments, corporate events and market-related data,⁵⁰ so it seems natural to extend its use as a metadata standard into alt data. The information it covers “includes pricing, descriptive and reference information, and statistics about financial instruments, exchanges and the organizations that trade through them, the economy in general, and other related economic and business factors.”⁵¹

IDSO seeks to improve standardization by publishing best practices, checklists and technical specifications. Their [Web crawling checklist](#) touches on regulation and compliance, website assessment criteria and risk management among other areas.⁵²

However, their main focus has been on the handling of personal data where they have recommended the use of existing privacy standards.⁵³ IDSO has defined a non-exhaustive list of personal data categories and examples such as name, address, phone numbers, account info, personal characteristics, linked information and more.⁵⁴ IDSO has also defined a three-tiered personally identifying information (PII) identifiability scale, ranging from direct identification (Level 1), to the ability to contact or impersonate (Level 2), to personal information that does not identify an individual (Level 3). There are numerous methods recommended for anonymization in the guide such as pseudonymization, hashing and swapping.

INCREASING ADOPTION OF STANDARDS

As the current activities highlight, while there are challenges around standardizing some aspects of the technical infrastructure around alt data, there are clearly some areas where adoption of common standards and best practices could be beneficial.

Data sets are often mapped and combined at great expense in terms of time and money, and every standardized identifier and practice could contribute to market efficiency.⁵⁵ The sector should continue to focus on agreeing to use common standards for areas that are already well-defined, and where there needs to be broad agreement to help reduce risks and unlock benefits.

As the figure on the next page shows, we can standardize a variety of different components of data infrastructure.⁵⁶

In the context of alt data, the sector could choose to create and adopt standards for:

- Representing basic data types, for example dates, and use of common data formats to help reduce unnecessary friction when analyzing data
- Identifying entities such as organizations, geographic areas, products. Using existing identifier schemes like [PermlD](#)⁵⁷ could make it easier to combine data from different sources, even if the content and structure of data sets varies
- Codes of practice and technical methods for Web scraping and other forms of data collection
- Documenting data sets, including key metadata standards, that will help to describe the provenance of data sets and the processes by which data has been collected and analyzed
- Describing the quality, coverage and known limitations of a data set

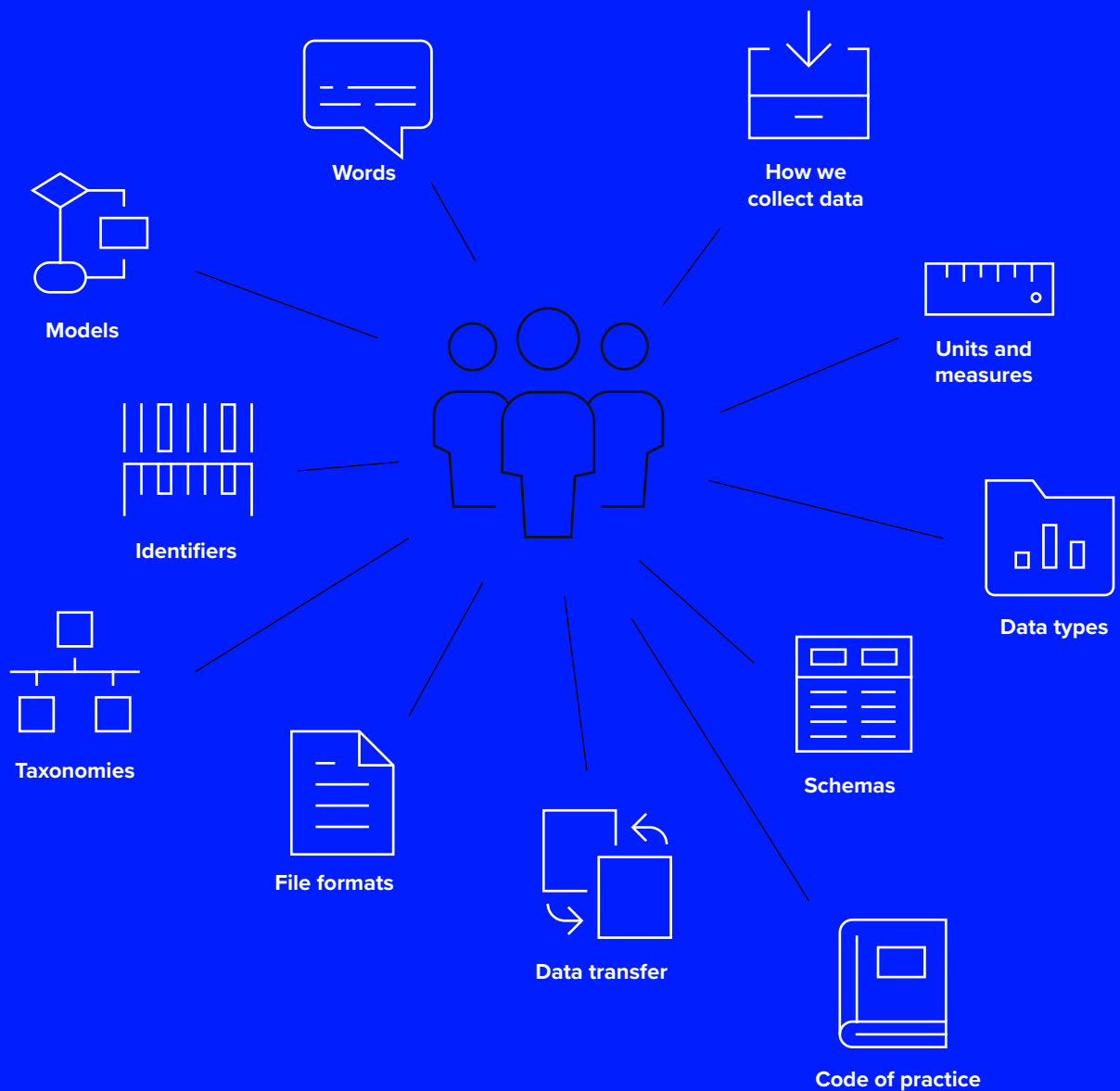
Alt data practitioners should also look to other parts of the broader data ecosystem to understand best practices in other domains. Statistical agencies, for example the Office of National Statistics (ONS) in the UK, provide good examples of how to document the provenance of data sets and identify the limitations of their use in other contexts.⁵⁸

Creating standards should be done through open processes. Collaborative models build trust, reduce cost and create more value than other approaches. Being open improves quality, as more people can contribute to the outcome, and increases the number of connections that can be made.⁵⁹ This approach has worked in adjacent industries before, the best example of which is the development of the Open Banking Standard. The CMA convened the nine largest consumer banks in the UK, and supported by organizations like the ODI, was able to implement an industry-wide set of standards around open APIs.⁶⁰

Collaborative models build trust, reduce cost and create more value than other approaches. Being open improves quality, as more people can contribute to the outcome, and increases the number of connections that can be made

We can standardize ...

Open standards for data consist of many different types of agreements. More complex standards are made up of smaller building blocks.

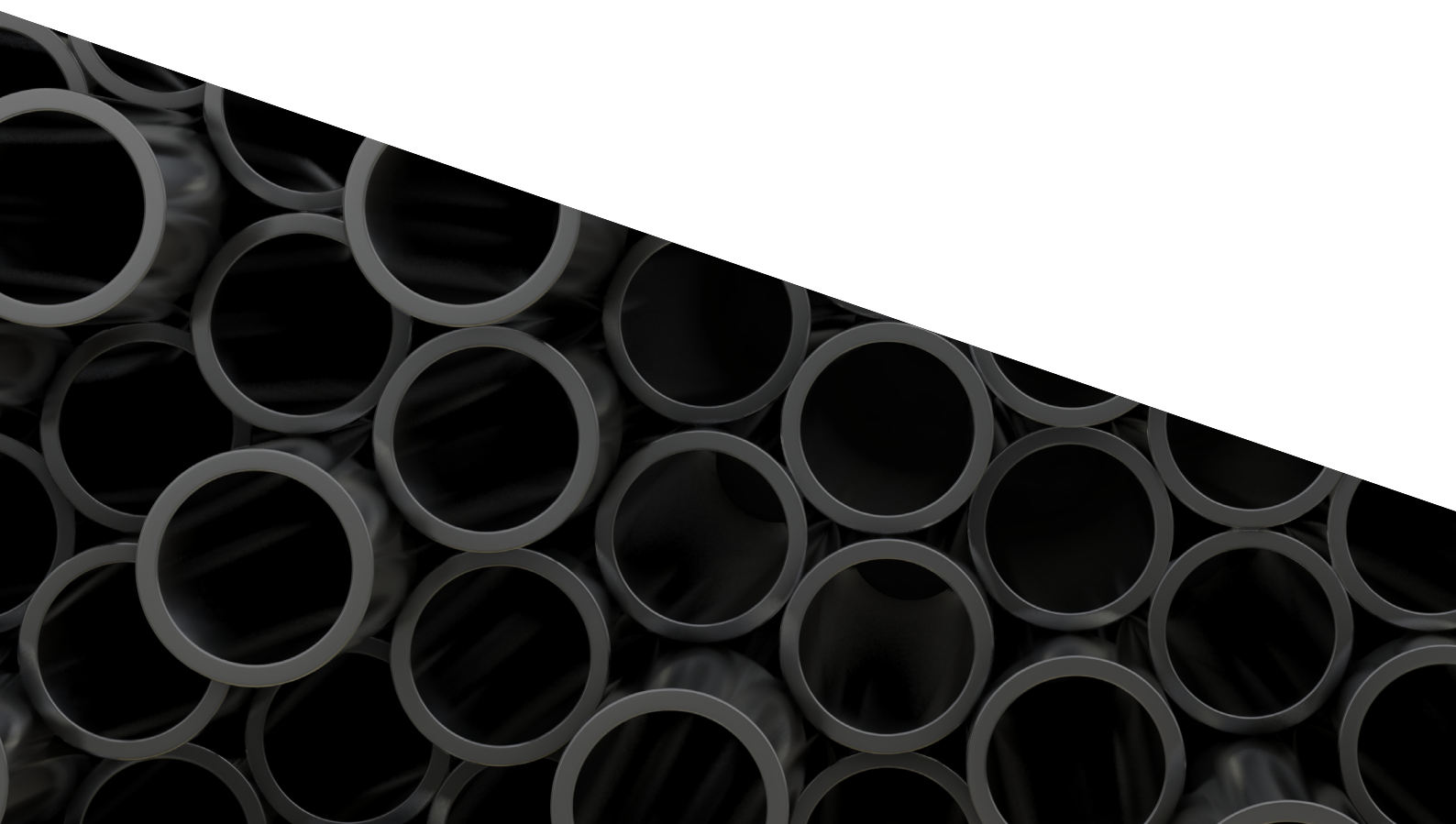


Recommendations and next steps

As we have shown, there are a number of issues around privacy, rights, ethical uses of data and the role of standards. Based on that, we recommend that the alt data industry focuses on several areas to create a more open, trustworthy data ecosystem, as set out below.

Both alt data providers and users need to ensure a strong commitment to ingraining legal and ethical practices across the alt data sector. Legal compliance and ethical behavior serve as a differentiator, a premium value proposition that helps ensure low-risk and high-quality products and services. Adhering to the law and demonstrating ethical behavior will reduce harm to the company and to society, and it will demonstrate the high value of one's investment services.

Organizations in the alt data space also need to collaborate better to improve access to data. Although there is currently some agreement across the industry on best practices, primarily related to the due diligence process and provenance, this remains limited. More collaboration around standards for technical aspects and transparent processes is both needed and achievable by an industry that is currently desiring it.



RECOMMENDATIONS FOR ALT DATA PROVIDERS

- **Alt data providers need to help their customers manage and mitigate legal and operational risks associated with the use of their data sets.** This can be done by ensuring they have appropriate rights/permissions to collect and distribute data, conducting risk assessments to help customers understand potential ethical or legal issues, providing clear provenance information, and, where rights are unclear, helping their customers stay on top of changing terms.
- **Alt data providers need to be transparent with users about the quality and limitations of their data sets.** This is important so that decisions based on their data sets are appropriate to their accuracy, coverage and timeliness. If provided, case studies and sample data need to be carefully documented so that users can make a fair assessment of the value of the data.
- **Alt data providers should be open with the organizations, individuals and communities impacted by the data they are scraping/repurposing.** This will help create more trust and transparency across the data ecosystem.

RECOMMENDATIONS FOR ALT DATA USERS

- **Alt data users should require data suppliers to provide detailed metadata, provenance information and documentation about the data sets they are supplying.** By setting expectations with suppliers, the users of alt data can help to encourage good behavior and create a more transparent and sustainable data ecosystem.
- **Alt data users should build robust processes to assess whether alt data sources are properly licensed.** Licenses are likely to be nonstandard but need to describe key areas such as usage rights (rights over redistribution), representations and warranties about the authenticity of the data, and the rights of the seller to license such data.

RECOMMENDATIONS FOR BOTH ALT DATA PROVIDERS AND USERS

- **All organizations in the alt data market should implement ethics assessments that go beyond compliance and legal issues.** Use tools such as the ODI's Data Ethics Canvas to help identify and make decisions about potential ethical issues associated with alt data-informed investment activity. Publish your findings openly, if you can.
- **Convene the key industry players and regulators in the alt data market to agree on a road map for standard adoption to improve technical integration and data set due diligence.** Similar to the implementation of the Open Banking Standard in the UK, have regulators and neutral third parties help decide on a framework for implementation.⁶¹ Standards will need to be stewarded by an independent organization and organizations may need support in implementing them. In the banking sector, this role is fulfilled by the Open Banking Implementation Entity (OBIE).⁶²
- **Identify where existing (open) standards in the market can be used and where new (open) standards need to be created.** Consider existing standards such as PermID, Market Data Definition Language (MDDL), DDQs and the Standardized Trial Data License Agreement before creating new standards. Identify if the SBAI, FISO, IDSO or other groups have other standards to be leveraged. Where new standards are required, follow open processes, such as those described in the "Open Standards for Data" handbook. Starting small, for example by agreeing on common identifiers and basic data formatting, may help to build momentum.

Appendix: Contributors

CONTRIBUTOR	ORGANIZATION	ROLE
Josh D'Addario	ODI	Consultant, Principal Author
Leigh Dodds	ODI	Director of Advisory, Principal Editor
Peter Wells	ODI	Director of Public Policy
Jeni Tennison	ODI	CEO
Anna Scott	ODI	Head of Content
John Walsh	Refinitiv	Director of Strategy & Innovation
Geoffrey Horrell	Refinitiv	Director of Applied Innovation
Adam Baron	Refinitiv	Director, Big Data Quantitative Research
Bob Bailey	Refinitiv	Chief Information Architect, Enterprise Architecture
Mahesh Narayan	Refinitiv	Global Head Proposition – Research and Portfolio Management
David Aubuchon	Refinitiv	Director – Market Development, Buy-Side & Wealth
Austin Burkett	Refinitiv	Global Head of Quant and Feeds
Michael Mayhew	Integrity Research	Founder and Director of Research
Emmett Kilduff	Eagle Alpha	Chief Executive Officer
Tim Harrington	BattleFin	President
Rob Martinez	BattleFin	Chief Revenue Officer
Matt Carpenter	Vertical Knowledge	Chairman and Chief Executive Officer
Peter Greene	Lowenstein Sandler	Vice Chair, Investment Management Group
Chris Pilling	Matchdeck	Founder and Chief Executive Officer

References

- ¹ Refinitiv (April 17, 2019), "Insights from the Refinitiv 2019 Artificial Intelligence/Machine Learning Global Study," https://refinitiv.com/en/resources/special-report/refinitiv-2019-artificial-intelligence-machine-learning-global-study?utm_source=Refinitivperspective_blog&utm_medium=blog&utm_campaign=107263_AISurveyReport&utm_term=&utm_content=Reqlp&eqCampaignId=6848
- ² Open Data Institute (2012), "The Open Data Institute," theodi.org
- ³ Refinitiv (2018), "About Us," <https://refinitiv.com/en/about-us>
- ⁴ AlternativeData.org (2018), "Get Started," <https://alternativedata.org/alternative-data/>
- ⁵ J.P. Morgan (May 2017), "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing"
- ⁶ J.P. Morgan (May 2017), "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing"
- ⁷ Open Data Institute (Jul 5, 2018), "Who do we trust with personal data?," <https://theodi.org/article/who-do-we-trust-with-personal-data-odi-commissioned-survey-reveals-most-and-least-trusted-sectors-across-europe/>
- ⁸ Thomson Reuters (April 15, 2014), "Twitter buys social data provider Gnip, stock soars," <https://www.reuters.com/article/us-twitter-gnip/twitter-buys-social-data-provider-gnip-stock-soars-idUSBREA3E17D20140415>
- ⁹ J.P. Morgan (May 2017), "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing"
- ¹⁰ J.P. Morgan (May 2017), "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing"
- ¹¹ Eagle Alpha (April 2018), "Alternative Data Use Cases Edition 6," https://s3-eu-west-1.amazonaws.com/ea-pdf-items/Alternative+Data+Use+Cases_Edition6.pdf
- ¹² Open Data Institute (Feb 2019), "Using geospatial data: a guide to licenses," https://docs.google.com/document/d/1N_y0Zhc583T8YJ4k2XnhZpFwnncDIDkUHP8gV3myes/edit#heading=h.4my9b7oijmma
- ¹³ SpaceKnow (Jan 2018), "China Satellite Manufacturing Index," <https://www.spaceknow.com/case-studies/china/SpaceKnow-China-Satellite-Manufacturing-Index.pdf>
- ¹⁴ Open Data Institute (2018), "Data Ethics," <https://theodi.org/service/data-ethics/>
- ¹⁵ Open Data Institute (Aug 5, 2017), "The Data Ethics Canvas," <https://theodi.org/article/data-ethics-canvas/>
- ¹⁶ Open Data Institute (2018), "Our theory of change," <https://theodi.org/about-the-odi/our-vision-and-manifesto/our-theory-of-change/>
- ¹⁷ European Union (May 2018), "The EU General Data Protection Regulation (GDPR)," <https://eur-lex.europa.eu/TodayOJ/index.html>
- ¹⁸ Open Data Institute (Feb 12, 2018), "ODI survey reveals British consumer attitudes to sharing personal data," <https://theodi.org/article/odi-survey-reveals-british-consumer-attitudes-to-sharing-personal-data/>
- ¹⁹ Integrity Research (Sept 12, 2017), "Privacy Watchdogs Worry About Hedge Fund Use of Geolocation Data," <http://www.integrity-research.com/privacy-watchdogs-worry-hedge-fund-use-geolocation-data/>
- ²⁰ Facebook-Cambridge Analytica scandal (2019) <https://www.bbc.co.uk/news/topics/c81zyn0888lt/facebooks-cambridge-analytica-scandal>
- ²¹ Integrity Research (Jan 2018), "Mitigating Legal Risks Associated With Alternative Data," <https://www.integrity-research.com/wp-content/uploads/2018/01/Mitigating-Legal-Risks-Alternative-Data-January-2018-2.pdf>
- ²² New York Times (Jan 3, 2019), "Los Angeles Accuses Weather Channel App of Covertly Mining User Data," <https://www.nytimes.com/2019/01/03/technology/weather-channel-app-lawsuit.html>
- ²³ Insurance Journal (Jan 7, 2019), "Los Angeles Sues IBM's Weather Channel for Use of Location Tracking," <https://www.insurancejournal.com/news/national/2019/01/07/514074.htm>
- ²⁴ Interview with Integrity Research (Feb 20, 2017), Peter Greene (Mar 13, 2019)
- ²⁵ The Open Data Institute (April 2019), "Anonymisation and open data: An introduction to managing the risk of re-identification," https://docs.google.com/document/d/1CoXniaTnQL_4ZyQuji9_MA_YCEEIQjx4z1SEdBO8c2M/edit#
- ²⁶ UK Anonymisation Network (UKAN) (2016), "Anonymisation Decision-making Framework," <https://ukanon.net/ukan-resources/ukan-decision-making-framework/>
- ²⁷ The Open Data Institute (April 2019), "Anonymisation and open data: An introduction to managing the risk of re-identification," https://docs.google.com/document/d/1CoXniaTnQL_4ZyQuji9_MA_YCEEIQjx4z1SEdBO8c2M/edit#
- ²⁸ CopyrightUser (May 18, 2017), "Text & Data Mining," <https://www.copyrightuser.org/understand/exceptions/text-data-mining/>
- ²⁹ AltData.TV (Sept 20, 2018), "Peter D. Greene on web crawling," <https://aitdata.tv/2018/09/20/web-crawling/>
- ³⁰ Integrity Research (Feb 20, 2019)
- ³¹ hiQ Labs (2018), "Who we are," <https://www.hiqlabs.com/new-who-we-are>
- ³² Lexology (Dec 3, 2018), "Data Scraping: Theft or Fair Game?," <https://www.lexology.com/library/detail.aspx?q=5e951f2d-55c7-42a3-a539-fbe88165ea5a>
- ³³ Electronic Privacy Information Center (2018), "hiQ Labs, Inc. v. LinkedIn Corp.," <https://epic.org/amicus/cfaa/linkedin/>
- ³⁴ SBAI (Feb 2019), "SBAI Publishes Standardised Trial Data License Agreement," <https://www.sbai.org/wp-content/uploads/2019/02/SBAI-Press-Release-SBAI-Publishes-Big-Data-Trial-Agreement-6-Feb-2019.pdf>
- ³⁵ Eagle Alpha (June 21, 2016), "Web Crawling as alternative data, a regulatory perspective," https://www.celent.com/system/media_documents/documents/399/944/216/original/554254227.pdf?1466607837
- ³⁶ The Open Data Institute (2018), "Our manifesto," <https://theodi.org/about-the-odi/our-vision-and-manifesto/our-manifesto/>
- ³⁷ Investopedia (Apr 30, 2018), "Material Insider Information," <https://www.investopedia.com/terms/m/materialinsiderinformation.asp>
- ³⁸ Integrity Research (Jan 2018), "Mitigating Legal Risks Associated With Alternative Data," <https://www.integrity-research.com/wp-content/uploads/2018/01/Mitigating-Legal-Risks-Alternative-Data-January-2018-2.pdf>
- ³⁹ SEC (Jan 21, 2015), "Securities and Exchange Commission vs. Bonan Huang and Nan Huang," <https://www.sec.gov/litigation/complaints/2015/comp23216.pdf>
- ⁴⁰ Integrity Research (Jan 2018), "Mitigating Legal Risks Associated With Alternative Data," <https://www.integrity-research.com/wp-content/uploads/2018/01/Mitigating-Legal-Risks-Alternative-Data-January-2018-2.pdf>
- ⁴¹ Open Data Institute (2018), "Open standards for data," <http://standards.theodi.org/>
- ⁴² Eagle Alpha (April 2018), "Alternative Data Use Cases Edition 6," https://s3-eu-west-1.amazonaws.com/ea-pdf-items/Alternative+Data+Use+Cases_Edition6.pdf
- ⁴³ The Open Data Institute (2018), "When not to create new standards," <http://standards.theodi.org/introduction/when-not-to-create-new-standards/>
- ⁴⁴ Alternative Investment Management Association (Oct 13, 2017), "AIMA launches new due diligence template," <https://www.aima.org/article/aima-launches-new-due-diligence-template.html>
- ⁴⁵ SBAI, "About Us," <https://www.sbai.org/about-us/>
- ⁴⁶ Financial & Information Services Association (2019), "About FISD," <http://www.siaa.net/Divisions/FISD-Financial-Information-Services-Association/About>
- ⁴⁷ MDDL (2014), "The Market Data Definition Language," <https://web.archive.org/web/20130512024014/http://v3-beta.mddl.org/>
- ⁴⁸ Finextra (May 30, 2007), "FISD launches market data definition language 3.0 beta," <https://www.finextra.com/pressarticle/15193/fisd-launches-market-data-definition-language-30-beta>
- ⁴⁹ MDDL (2014), "The Market Data Definition Language," <https://web.archive.org/web/20130512024014/http://v3-beta.mddl.org/>
- ⁵⁰ Finextra (May 30, 2007), "FISD launches market data definition language 3.0 beta," <https://www.finextra.com/pressarticle/15193/fisd-launches-market-data-definition-language-30-beta>
- ⁵¹ IDSO (Jan 2018), "IDSO Best Practices: Web Crawling," https://docs.wixstatic.com/ugd/c6ff57_43135666c05c49f8ff7c2763ef846f0.pdf
- ⁵² The National Institute of Standards and Technology (April 2010), "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)," <https://nvlpubs.nist.gov/nistpubs/legacysp/nistspecialpublication800-122.pdf>
- ⁵³ IDSO (May 2018), "IDSO Best Practises: Personally Identifiable Information (PII)," https://docs.wixstatic.com/ugd/78bb6b_8c273efa9b934df796fd309bb5f5de8.pdf
- ⁵⁴ Amelia Axelsen (Feb 20, 2019), "Crowded Alt Data Market Makes Standing Out Difficult for Providers," <https://www.waterstechnology.com/data-management/4162636/crowded-alt-data-market-makes-standing-out-difficult-for-providers>
- ⁵⁵ The Open Data Institute (2018), "Types of open standards for data," <http://standards.theodi.org/introduction/types-of-open-standards-for-data/>
- ⁵⁶ Permid.org, <https://permid.org/>
- ⁵⁷ Office for National Statistics (2011), "Methodology and variables," <https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/methodologyandvariables>
- ⁵⁸ Open Data Institute (Aug 31, 2016), "Principles for strengthening our data infrastructure," <https://theodi.org/article/principles-for-strengthening-our-data-infrastructure/>
- ⁵⁹ Open Data Institute (2016), "Open banking: setting a standard and enabling innovation," <https://theodi.org/project/open-banking-setting-a-standard-and-enabling-innovation/>
- ⁶⁰ Open Data Institute (2016), "Open banking: setting a standard and enabling innovation," <https://theodi.org/project/open-banking-setting-a-standard-and-enabling-innovation/>
- ⁶¹ Open Banking Ltd. (2016), "About Us," <https://www.openbanking.org.uk/about-us/>

Refinitiv serves more than 40,000 institutions in approximately 190 countries. Refinitiv is committed to providing information, insights and technology that drive innovation and performance in global financial markets. Our heritage of integrity enables our customers to make critical decisions with confidence – while our best-in-class data and cutting-edge technologies enable greater opportunity.

Quite simply, we enable the financial community to trade smarter and faster. Our service can help you overcome regulatory challenges, scale intelligently, manage risk, develop differentiated investment strategies and, we believe, invest more intelligently than ever before.

Visit refinitiv.com/alternativedata



@Refinitiv



Refinitiv



The Open Data Institute, 3rd Floor, 65 Clifton Street,
London EC2A 4JE, UK theodi.org

REFINITIV®
LABS

