DON'T BELIEVE THE HYPE - COLLECTING DATA IN A POST-TRUTH WORLD

Adages about lies, damned or otherwise, and their relationship to statistics are often used to imply a disingenuousness on the part of the accused. But as anyone in a data-heavy industry (and these days, that is most people) will tell you, it is never as simple as counting a complete and consistent dataset.

As the rise of 'big data' shows, datasets are rarely easily found, marshalled or analyzed. It can take complex software, some lateral thinking and a lot of manpower to compile datasets that are comprehensive, consistent and comparable.[1]

Moreover, it is becoming harder rather than easier to do so. Privacy laws are making it more burdensome to collect personal data, even as people's online lives become more intricate than ever before. In the financial world, the same is true: an increasing number of corporate entities are shielded from regulatory or reporting obligations, even as their corporate structures are more complicated than ever before. The relative reporting burdens on different countries, industries and entities can make any financial dataset into a patchwork quilt, making it hard to draw concrete conclusions.

A prime example is the alternative assets industry; a key financial sector, it now holds almost $9tn in assets under management.[2] Yet very few industry participants have mandatory reporting requirements. This is despite the industry having a tangible impact on the wider financial ecosystem: alternative assets funds are of a size to compete with corporate and strategic investors for acquisitions; venture capital funds have helped foster a booming global tech ecosystem; real estate and infrastructure funds build, run and maintain iconic and vital buildings, bridges and tunnels.

Pension funds, insurance companies and government agencies all invest in alternative assets, and so their success or failure can impact even the personal finances of individuals. The high stakes involved create a need for reliable, accurate and ethical data. Investors in alternative assets need to be able to make decisions with access to the best available information. Fund managers must be able to place their activities in the fullest possible context. Regulators, commentators and analysts must be given an accurate assessment of the industry in order to effectively manage or analyze the industry.

At the same time, though, the nature of the industry makes the challenge of collecting and organizing such data significantly more difficult than in almost any other financial sector. It can be hard to know how to trust any data provider when they tell you they know the "truth." How can they claim to have all the information? Where have they gathered it from? And how are they distilling these disparate data points into a semblance of reality? What assumptions have gone into their models?

Mark O'Hare has spent several decades grappling with these issues, first as the founder of equity shareholding information service Citywatch, and now as Chief Executive of Preqin, the alternative assets data provider. In this paper, he discusses some of the issues that face companies seeking to gather financial data, and how their processes and assumptions can affect the intelligence on which financial decisions are made.

DARK MONEY: ALTERNATIVES OVERSIGHT

An Opaque Industry

Even in the complex world of finance, the alternative assets industry is recognized as particularly opaque. Although alternatives are now an integral part of most investors' portfolios, most participants are subjected to relatively few reporting requirements.

[1] There is a large body of academic study into the challenges of processing and studying large and complex datasets. See particularly Chen, Zhang 2014 A Survey on Big Data (https://doi.org/10.1016/j.ins.2014.01.015), and Boyd, Crawford 2011 Critical Questions for Big Data (https://doi.org/10.1080/1369118X.2012.678878).

[2] According to the 2018 Preqin Alternative Assets Performance Monitor, the industry holds assets under management of $8.81tn as at the end of March 2018 (http://docs.preqin.com/press/Perf-Monitor-Sep-18.pdf).

> **Four out of five investors globally have some allocation to alternatives, and the market as a whole is approaching $9tn in assets under management.**

Private institutions must report only to their boards or trustees, while private fund managers below a certain size do not have to report their activities other than to their investors.

There are a few exceptions to this: public pension funds are generally required to regularly disclose their portfolios, and large fund managers must file with regulators in markets like the US and UK. But compared to mutual funds, REITs and other investment vehicles, the alternative assets market faces little in the way of required reporting.

This is a holdover from a time when investing in private equity or hedge funds was a niche activity, in the 'corporate raider' days of the 1980s. The reasoning was that among the handful of managers at that time, the illiquid nature of their holdings, their relatively small size and the individualistic approach they pursued made it unnecessarily burdensome to require them to report in the same way as traditional fund managers.

However, in 2018, four out of five investors globally have some allocation to alternatives,[3] and the market as a whole is approaching $9tn in assets under management. There is, therefore, an obvious need for investors to access accurate, timely and comprehensive information regarding both their own investments and the market as a whole. Activities such as looking for potential new investments, evaluating fund marketing documents and properly benchmarking the performance of portfolios all require that investors can compile figures on an industry that has historically been difficult to assess.

Fund managers themselves also have a need for reliable and actionable intelligence. Whether they are benchmarking themselves against their peers, looking for potential deal opportunities or seeking new investors, they also benefit from being able to know more about the industries and sectors in which they are operating, and about other industry participants.

A Worthwhile Task

There is, then, a mutual need for data provision that encompasses the entire alternatives industry and offers participants a clear and accurate view of what is happening in the space. The stakes are high: public pension funds, for instance, now control an estimated $14tn in retirement savings, of which an average of 6% is invested in private equity and over 7% in hedge funds. Enabling them to make better decisions is a vital and impactful task.

At the same time, many start-ups and new companies are choosing to remain as private entities longer into their lifecycle, in many cases relying on private sources of capital like alternatives funds. Especially in booming emerging markets like China and India, private investment vehicles are a cornerstone of the private sector and are helping to stimulate some of the fastest-growing economies in the world.

With this in mind, the task of providing accurate and insightful data on the industry is as much one of principle as it is business. At Preqin, we believe that having access to data of the best possible quality will allow investors and fund managers to make better-informed decisions. This will make capital markets more efficient, investors will see better returns and more capital will flow to pension pots, government budgets and charitable foundations around the world.

FOLLOW THE MONEY: A PATCHWORK DATA TRAIL

Knowns, Unknowns and Known Unknowns

Although most alternative assets industry participants have few or no regulatory reporting requirements, there are some known sources of information on which data providers and commentators rely. One of the largest is US-based

[3] *According to the H2 2018 Preqin Investor Outlook: Alternative Assets, 80% of institutional investors have an active allocation to at least one alternative asset class, and 50% allocate to three or more alternative asset classes (http://docs.preqin.com/reports/Preqin-Investor-Update-Alternative-Assets-H2-2018.pdf).*

public pension funds, which are required to release quarterly updates on their fund portfolios and the returns they are seeing from them. Given that many of these pension funds are very large institutions with expansive portfolios, this provides fund-level information on a sizeable proportion of the active fund universe.

The difficulty with the information they provide is that it (understandably) specifically relates to the returns those pension funds are getting from their investments. Being large and influential investors, public pension funds are often able to negotiate reduced fee rates, or access to sidecar structures and co-investment opportunities. This may mean that the returns a large public pension fund gets from a fund are different from the returns seen by a small wealth manager, or even another public pension fund that signed a different Limited Partner Agreement (LPA).

It is also challenging in that fund managers are not required to disclose every detail of their fund's activities to investors. They will notify the investor of capital call-ups and distributions, but may not be bound to disclose the levels of leverage they use to augment their buying power, or the use of credit lines to regulate the cash flow of the fund. Where they do disclose details to investors, they may ask that the investor keeps the information confidential. There is some data, in fact, that might never be obtained through impersonal sources, as it will never appear in filings, in the news or in investor statements.

Another important source is government-mandated regulatory repositories. Bodies such as the SEC in the US and Companies House in the UK require funds above a certain size to register with them and submit filings on their fundraising activities. While this provides detailed information on new funds coming to market, there is no centralized categorization or search function for these filings, posing a challenge for those looking to aggregate all the available information. Notices announcing the formation of new alternative investment vehicles or holding companies can easily get buried among thousands of other regulatory filings for other corporations or financial products.

This is also not a globally applicable repository of information. Regulations regarding investment vehicles and their reporting requirements differ widely between countries and are further complicated by funds that are headquartered in one jurisdiction but domiciled in another. There is no single agreed-upon standard for reporting from either fund managers or investors, despite initiatives spearheaded by industry bodies such as the Institutional Limited Partners Association (ILPA)[4] suggesting best practice. This means that the level of information provided to investors, and in turn to the public, can vary widely depending on the particular structure and location of the fund manager in question.

The Rise of the Machines

The most common way to collect this information, where it is available, is to use manual and algorithmic methods to extract quantitative data from news articles, regulatory filings and public pension statements. Machine learning and artificial intelligence tools can increasingly automate this process, being able to scrape web pages to extract the numerical data from text-based reports. Most data providers, including Preqin, are investing in these technologies to speed up the collection process and reduce the reliance on manual research and collection.

However, there are some key problems with this approach. The first is conformity. Assuming that web scraping programs are able to collect 100% of the available information, all companies employing such tools will have a 'complete' dataset that is indistinguishable from its peers – leaving nothing to differentiate between providers. It can also be prone to errors of extraction. An algorithm cannot account for every way that information can be presented or reported and is not able to read the full report

[4] ILPA have released a series of templates and guidelines for fund managers to report to their investors, including resources such as their Fee Reporting Template (https://ilpa.org/wp-content/uploads/2016/01/ILPA-Fee-Reporting-Template-Version-1.0-Guidance-1.pdf).

> **“Every year we have direct conversations with more than 7,000 fund managers, and more than 9,000 investors**

and understand a number in its context. This is why all the information we collect from web scraping programs is cross-checked and curated by human researchers with an understanding of the industry and the dataset.

Algorithms are also unable to collect information held in non-public sources. Some investors, for example, have more information on their portfolios than they release publicly. In the US and UK, though, that information can still be accessed by making Freedom of Information Act (FOIA) requests to the investors concerned – something a machine cannot replicate.

Lastly, there is the significant challenge of corroboration. Most web scraping algorithms do not have sophisticated means of cross-checking information to ensure its accuracy, especially when information may come from different kinds of sources. Where possible, it is more timely and accurate to collect information from fund managers and investors themselves, which can then be corroborated both internally and with external sources to ensure its veracity. Preqin, for instance, encourages fund managers to submit their performance and fundraising information on a monthly or quarterly basis, ensuring that information on our database remains timely and up to date. More than 10,000 alternative assets funds submit their data to us on this basis – representing more than 25% of the known fund universe and more than 60% of its estimated market capitalization.

Most importantly, though, some of the most valuable and actionable information is simply not quantifiable from news sources and regulatory filings. We prioritize direct contact with fund managers and investors precisely because so much granular data on their activities can only be collected this way. It allows us to collect qualitative information on investors' intentions for the next 12 months, their active mandates and their concerns with the industry. It allows us to find out what trends fund managers are seeing regarding their investors, what challenges they are facing, and where they perceive the best opportunities over the coming months.

We aim to speak to all fund managers and investors in our database at least once a year where possible, and more frequently in many cases: high-profile investors or fund managers currently fundraising might be contacted as frequently as every month. Every year we have direct conversations with more than 7,000 fund managers, and more than 9,000 investors, to find out qualitative and often exclusive information about their funds, investments, plans and challenges. This allows us to present industry participants' activities in context – what they have done in the past, what their plans are, what their peers are doing – which is simply not possible when relying on automated data collection.

WHAT HAPPENS WHEN YOU ASSUME

Modelling the Data

Even when all the available data has been collected and logged in the dataset, the work has barely begun. Raw data, especially when it relates to real-world activities, can often paint a confusing, fragmented and contradictory picture. This makes it difficult to know how the data relates to what is really going on.

For example, say that there are three available sets of performance data from three different pension funds all stating the returns they have seen from a particular fund. They are all slightly different, but the fund cannot be performing at three different levels. The differences lie in the specific commitments the pension funds made and the details of the LPAs they signed, but that information is unavailable. How can these disparate figures be reconciled to produce the most accurate picture?

Every data provider, then, has to introduce some curation and modelling to the raw dataset in order to present actionable intelligence rather than a confusing sprawl of data. By its nature, this requires making some assumptions, performing some calculation and extrapolating from known precedents. But this is a double-edged sword: to introduce assumptions and processes is to introduce bias or obfuscation into the dataset which may not be useful to the end-user.

To return to the performance problem, which route is best? To take the performance from the largest investor? The one considered most reputable? To take an average of the figures? And if so, should it be weighted by commitment size? None seem like desirable options, and all can result in providing information that does not benefit someone looking into the performance of the fund in question.

There are two key ways to mitigate this. The first is to make intelligent models that are cross-checked and corroborated to be consistent. Preqin does this by cross-referencing data points with each other. We then query inconsistent data with fund managers or investors themselves, asking them to provide further evidence of the figures in question. This ensures that we present as cohesive a dataset as possible – ultimately, if our researchers are not satisfied with the validity of a data point, we will not include it in our models.

This goes beyond human quality controls and into the aggregating calculations that data providers make. For instance, Preqin has a cash flow modelling tool which helps investors predict when they might expect capital calls and distributions from a fund. This is based on the historical cash flow data of more than 4,300 funds, which enables us to benchmark expected cash flow timings. Having a reliable and robust benchmark means that outliers are more quickly identified and, if needed, flagged for further review.

The other key factor is to be transparent about the sources of data, especially in cases where there are multiple available data points saying different things. Fund performance is a prime example of this, as noted in the example above. Where information is gathered from different sources, it is important to recognize that different users will find different sources more relevant or useful. As such, we make all the source data for performance information available, and when building benchmarks or examining fund performance, users can switch between the information sources that best suit their needs.

Is Bigger Always Better?

There are other assumptions that data providers make on behalf of users that go in the other direction – including more information than is useful, rather than condensing multiple sources. There is a frequently drawn conclusion among both data providers and commentators that big numbers must always be better – an attitude that extends well beyond financial data into all walks of life. However, it is not always a useful exercise to gather as much information as possible and break it into separate data points, and it can be disingenuous to claim that these figures represent a larger known universe.

For example, if a hedge fund has 20 known share classes that all operate on a master/feeder structure, they should plausibly be considered one entity rather than 20. By the same token, if the master fund represents the pooled assets of the feeder fund, to count those assets twice over would not be a fair reflection of the fund's size. Similarly, if an insurance company has 10 sibling entities that all make investments in alternative assets, but through a single investment arm of the parent corporation, it must be considered a single investor rather than 10 exactly-aligned-but-apparently-separate ones.

The inaccuracies in the 'bigger is better' approach go far beyond a misstated headline figure. It can actively obscure the data's relevance to the real world and have material consequences for users of the dataset. A fund manager might approach one of the insurance companies with a fund pitch, only to find out that they do not do any investing on their own behalf. An investor may take a hedge fund's double-counted size as a testament to its appeal and change its investing decisions accordingly.

This highlights the importance of data curation, and that data providers must constantly be relating their figures back to the activities of industry participants to ensure that they reflect reality. Whether the initial information is gathered by human interaction or machine learning, it needs to be checked for quality, corroborated and tied back to reality to be of use.

When investors and fund managers have access to better data, they are able to make more informed decisions.

Preqin believes this requires a human touch. While machines may one day be able to make judgements on information quality and perform corroborating checks, we trust that trained, engaged and intelligent human curation will always result in the best-quality data that is most accurate and actionable.

NUMBERS ON A SCREEN

Can You Repeat the Question?

Once all the available information has been gathered, and moulded to give a fair reflection of the realities of the industry, it may seem that the challenges are mostly over. But in fact some of the hardest tasks are with how the raw data points can be most usefully drawn through into tools and analytical models that are of real help to the user.

One thing that data providers, and indeed companies of all stripes, sometimes forget is the actual tasks that their users are looking to achieve. Very few users will access a dataset by asking questions like "what is the median performance of a North America-based buyout fund with a 2004 vintage?" Instead, users might be asking "what's going on in the healthcare private equity sector," "what regions should I be paying most attention to in the next six months?" or "how is my portfolio doing compared to the market as a whole?"

This is to say that specific data points are needed for some use cases, while trends and patterns are more useful in others. Trying to guide users to the answers that they need is therefore a delicate balance between making the raw line-by-line dataset available and having aggregating tools that can automatically show them relevant trends or higher-level figures. The raw datasets can be overwhelming, but to apply too much built-in calculation is to force the user into accepting assumptions of which they may not even be aware.

Preqin strives to offer users the ability to employ both approaches. They can download the line-by-line unfiltered data themselves in order to apply their own models, or they can use our cutting-edge analysis tools to offer a quick overview and insight. But a binary choice does not reflect the needs of most users, so Preqin allows them to mix the two approaches: when making custom benchmarks or target lists, users can set aggregate parameters to find the funds, firms or institutions they want, and can then add or remove individual entrants according to their preferences. We believe that only this level of control allows users to really get to the heart of what they want to know without Preqin's own biases obscuring the facts.

This matters, because the data taken from providers is frequently used to plug into financial risk models, due diligence processes and decision-making analysis. Many of Preqin's clients receive data through use of an API function, meaning that they will only ever see the data within their own client software. This means it is imperative to get the balance right between providing a clean dataset free from duplication or misclassification and providing an overworked dataset that has been previously modelled.

They Seek Them Here, They Seek Them There

The other key consideration is the connectivity of different datasets. What may be partitioned from a database perspective may well be part of a continuum of information useful to a user. Very few use cases are interested in only a single category of information. For most users, different data points will have bearing on other categories of data, forming several parts of a cohesive understanding of a sector or trend.

For instance, knowing which investors are targeting Sub-Saharan Africa naturally raises questions about how much money is going to the region. This in turns begs the questions as to which fund managers are raising it, what opportunities they are deploying it into, and what returns they are seeing from it. Datasets that may be drawn from different collection methods, and curated by different teams, nonetheless need to be able to work seamlessly to provide a holistic reflection of what is going on.

Likewise, the decisions that an investor makes about where to target for their next private equity investment will have a tangible impact on whether,

THE PREQIN DIFFERENCE

We prioritise human interaction

We maintain complete transparency

We enable confident analysis

when and where they invest in real estate, hedge funds or other asset classes. Preqin believes that keeping data siloed between different asset classes and data types is not an effective way for users to gain intelligence on the market, which is why we have developed at-a-glance profiles and searches that span all regions and asset classes.

THE PREQIN DIFFERENCE

While the methodologies and challenges of collecting data supersets are a recognizable field of study, relatively little attention has thus far been paid to the applications of these concepts in institutional (as opposed to consumer) finance. Less still has been studied about data for the alternative assets industry specifically.

This is despite the fact that alternative assets represent a known challenge to investors and commentators in their opaque nature and widely varying reporting standards. Although industry bodies like ILPA and the AIC are trying to introduce industry best practice, their frameworks are voluntary, and they lack any regulatory power.

This creates a simultaneous need and challenge: the industry would undoubtedly benefit from better access to data, but under current conditions that access is limited and inconsistent.

There are three key things that Preqin does to ensure that its datasets are as comprehensive, consistent and comparable as possible:

- We prioritize human interaction when collecting raw data. This allows us to supplement, cross-check and curate the data gathered from digital sources like web scrapers, news articles and regulatory filings. It also means we can gather qualitative data on participants' priorities and intentions, as well as the challenges they face.
- We maintain complete transparency about the sources of our data and our processes for curating it. This means that participants can choose to include or exclude specific data points or sources as best suits their needs, while still ensuring that the dataset as a whole remains free of duplication or too much inherent bias. We also make our collection methods and calculation methodologies freely and publicly available[5] for inspection.
- We make sure that the data fed into our clients' models and presented on our platform is raw but structured. This enables analysis to be done on the dataset with confidence: the data does not contain overlapping or duplicated data points that would obscure potential findings, but is neutral in terms of analysis or conclusion.

It is sadly true that there is no such thing as the 'perfect' dataset. Any large collection of data will inevitably have flaws – whenever data relates to real-world events there can be contradictions or inexplicable gaps that contribute to an incomplete or inconsistent dataset. However, there are avoidable errors of process or bias that can be eliminated – issues of collection, structure and analysis.

Ultimately, any dataset is only as good as the uses to which you can put it. Collecting more individual data points is useless if they cannot help to provide the user with an accurate reflection of real-world events. Preqin therefore constantly looks to tie its datasets to the real world, be that through conversations with industry participants, providing the basis for risk and financial models, or providing our own analysis of specific markets and sectors.

We strongly believe that a commitment to transparency and openness are paramount qualities for an information provider, and that ensuring all industry participants have access to the best quality data and tools is vital to the continued growth and wellbeing of the industry. When investors and fund managers have access to better data, they are able to make more informed decisions. When they make more informed decisions, capital markets become more effective and efficient and all parties ultimately benefit, including those who rely on institutions like pension funds to look after their personal finances.

[5] *Preqin's data coverage sheet gives some examples of collection methods, while methodologies for analysis tools on Preqin Pro are available to users (http://docs.preqin.com/reports/preqin-global-data-coverage.pdf).*