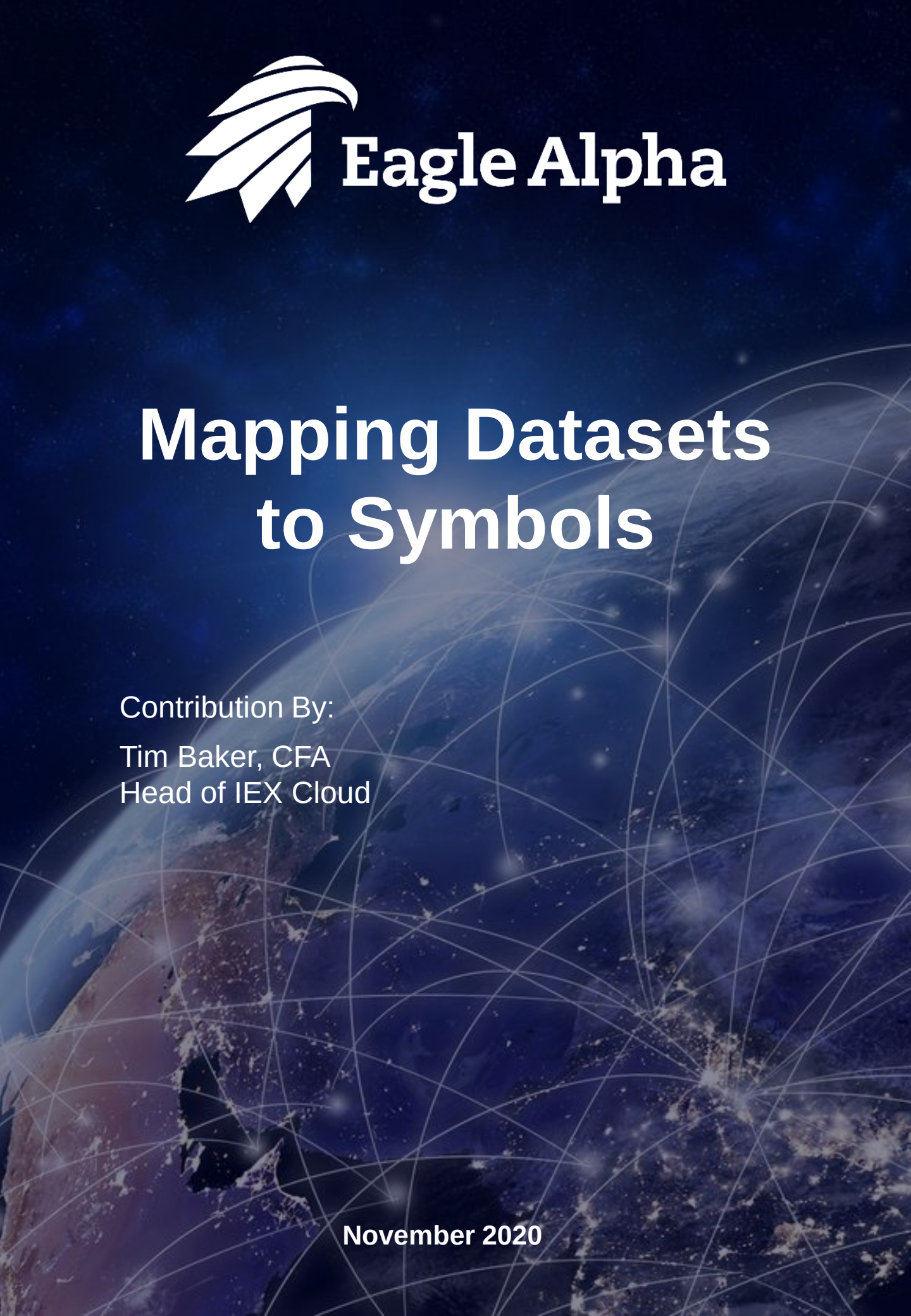


Eagle Alpha

Mapping Datasets to Symbols

Contribution By:

Tim Baker, CFA
Head of IEX Cloud

November 2020

Table Of Contents

Introduction and Purpose of This Report

This report addresses the topic of ticker mapping as it relates to the alternative data marketplace. The reports objective is to inform both buyers and sellers of alternative data of the complexities and solutions of ticker mapping. Broadly speaking ticker mapping involves the application of identifiers, symbology or tickers to companies, or entities, in a dataset.

As a leading aggregator in the alternative data market Eagle Alpha sits at the junction between data vendors and buyers of data. Alternative data vendors often do not understand the nuances of the fund management industry and the need of the buyside to have datasets mapped to some identifier, symbol or ticker. The buyside in this instance is the financial services industry, particularly hedge funds, quant funds, pension funds and mutual funds. Across the industry the ubiquitous question is – "is the data mapped to tickers?" The term mapped to ticker or "tickerization" is a generic term used to indicate if a dataset has been mapped to a financial symbol. It may not mean it is mapped specifically to a stock ticker, but it does need to be mapped to some "identifier" that the buyside can in turn map to in their securities master file. The important point is that datasets that are not mapped to some type of recognized securities identifier can face obstacles in pre-sales meetings, trialling of data and ultimately sales.

The team at Eagle Alpha impress on vendors the necessity to map their data to some type of identifier, symbol or ticker. Over the years buyside clients have made comments to us that reinforce the necessity of ticker mapping. For example;

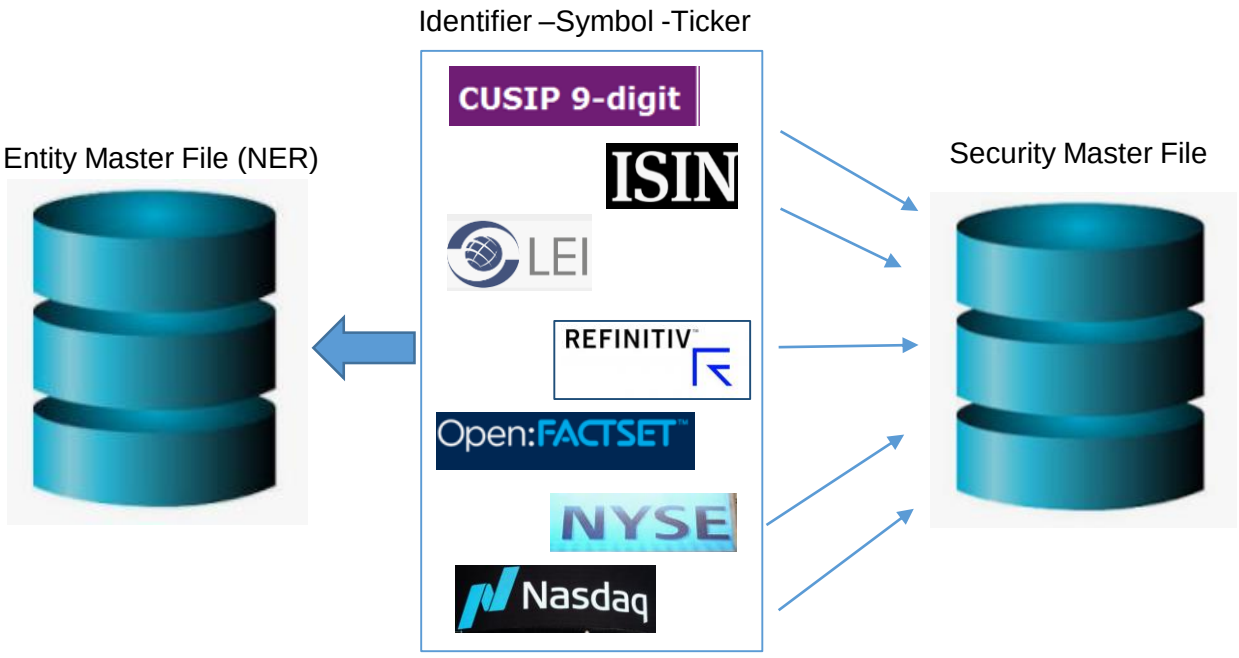
- *"A lot of the time the missing ingredient is vendors having any identifier at all as a lot of the time we get a messy string based on how they view the world. So it is helpful if they make any effort of getting some sort of an identifier that we can work with".*
- *"Vendors tell us something like 'just give us a name and we will get you everything on that company'. We need to explain that is not how we work. For example, we need to send emails explaining what point in time data means and what we are looking for".*
- *"You can't do equity mapping if you can't keep up with corporate actions…they should be outsourcing it".*
- *"It is not only important to have alternative datasets map to something, it's how easily you can link that mapping to an internal system".*

We asked the FISD Alternative Data Council [1] what the latest feedback is from the buyside in relation to ticker mapping and symbology. Tracey Shumpert, who leads the FISD's efforts in this regard, responded by saying "our mission is to establish best practices and standards for the delivery of alternative data to the investment industry so that the data can flow more easily. Mapping data to security identifiers is a key requirement identified by investment managers. The "Data Vendor Tear Sheet" developed by the Council points out this requirement. The mapping of security identifiers to alternative datasets goes a long way to making them easier to integrate into buy-side data management processes".

[1] *The Alternative Data Council is series of working groups and information-sharing forums within FISD and was created as part of the strategic initiative to engage the alternative data community. They are focused on establishing best practices and standards for the delivery of alternative data to the investment industry.*

The core of this document focuses on two key aspects that are represented in the diagram below. First and foremost it is important to map relevant entities in a dataset to an entity master file. This is performed by a process called Named Entity Recognition (NER). Entities are then tagged in some form to at least one commonly used identifier, symbol or ticker. The problem is the industry is replete with symbols, which brings another layer of complexity to the exercise. For the buyside all of this symbology can be held in a securities master file, which is used to map to the identifier(s) presented by a vendor in their alternative dataset.

Figure 1: Symbology is at the Centre of Mapping



Source: Eagle Alpha

This paper address the intricacies surrounding identifiers and also examines the importance of building an entity master file. Head of IEX Cloud, Tim Baker, begins with the complexities of identifiers, symbology and ticker mapping. The analyst and data science team at Eagle Alpha discuss building an entity master file and named entity recognition (NER).

We follow this with some experience and advice from some large alternative data vendors who have successfully addressed mapping their datasets. Lastly, our research has led us to a handful of technology solutions providers that have the ability to perform ticker mapping as a service on alternative datasets and we present a matrix of their capabilities.

Eagle Alpha

Section 1

Identifiers and Symbology

Identifiers, symbology and tickers. A run down on a complex issue from Tim Baker, the Head of IEX Cloud

iex cloud

Identity Crisis

A few years ago, my wife wanted a sculpture for our front yard garden for her birthday. She found what she liked, and we commissioned the piece that duly arrived on a truck from Houston two months later. The artist had titled the piece Symbolism – I suppose because it incorporated the "∞" symbol. Since I'd found myself immersed in the topic of Symbology, we renamed it as such: "Symbology."

I think about Symbology almost every day. According to Wikipedia "a **symbol** in computer programming is a primitive data type that's instances have a unique human-readable form. Uniqueness is enforced by holding them in a symbol table. … and most common indirectly is their use to create object linkages."

Bloomberg's OpenFIGI site states: "Symbology refers to more than a code – it is the methodology and system for defining how data is related and how that information is conveyed".

Modelling the World

- **What is an ontology?** In computer science and information science, an ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains of discourse.
- An ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.
- **What is meta data?** The "data about data." A set of data that describes and gives information about other data (e.g. this news article is about Amazon and Jeff Bezos; this building is occupied by McDonalds).
- **What is an identifier?** A specific label used to reference an object, such as a product, company, person, building or security.
- **What is symbology?** Symbology relates to identifiers for securities or "ticker symbols."

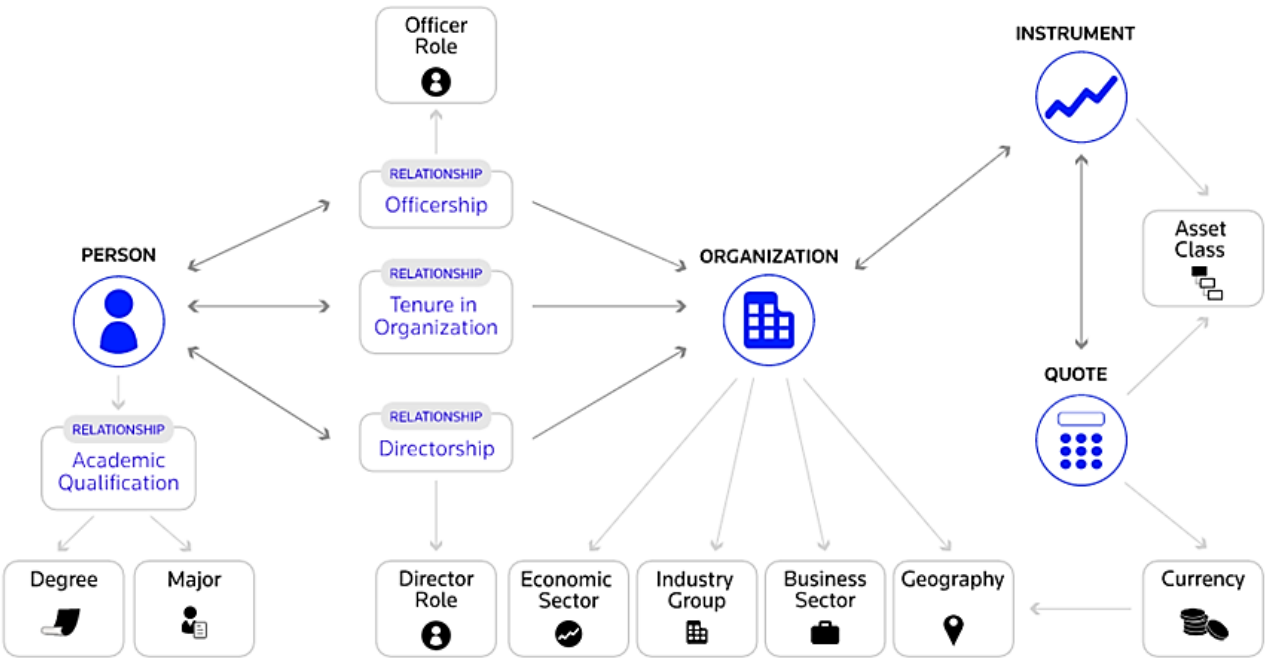
The past dozen or so years I have followed the evolution of this topic as data integration and provenance has become mission critical in the financial data space. There remain persistent challenges and costs associated with something so basic as disambiguating whether a security is the one you think it is – before you trade it, analyze it, connect it with associated data, study it in a chart or regress it against other time series data.

Challenges with symbology are part of a bigger issue around reference data – with no agreed upon standards, vendors are left up to their own devices to create their own ontologies, treatments, and yes, symbology. Some attempts to create utilities to solve for these challenges have had mixed results. EDM Council's FIBO, for instance, looks to create consensus across the industry. Finding a solution is made more difficult perhaps by the presence of "too many cooks" – and potentially the involvement of some of the vendors in these initiatives. But then how do you come to a consensus when the stakes are so high, and so many people and organizations must agree?

For entities (companies and other organizations) the Global Legal Identity Foundation (GLIEF) has made great strides to help the financial services industry solve the entity identification issue – and has openly licensed the data for use and reuse freely across the industry. But the reality is that challenges continue as vendors keep proprietary identifiers locked down which they believe keeps clients tied into their data offering, or from licensing revenues of the identifier in its own right.

I get it – maintaining this complex network of identifiers and symbology is a non-trivial thing and can be expensive. Firms have started to open a little – take the open permid.org from Refinitiv, or FIGI from Bloomberg. Both are somewhat "open," and certainly useful in their own right. They also differ considerably in their scope. For instance, FIGI is solely securities orientated, whereas PermID not only addresses equities but also includes related identifier classes such as companies, people, and sectors.

Figure 2: Relationships and Complexity of the Ecosystem



Source: permid.org / Refinitiv

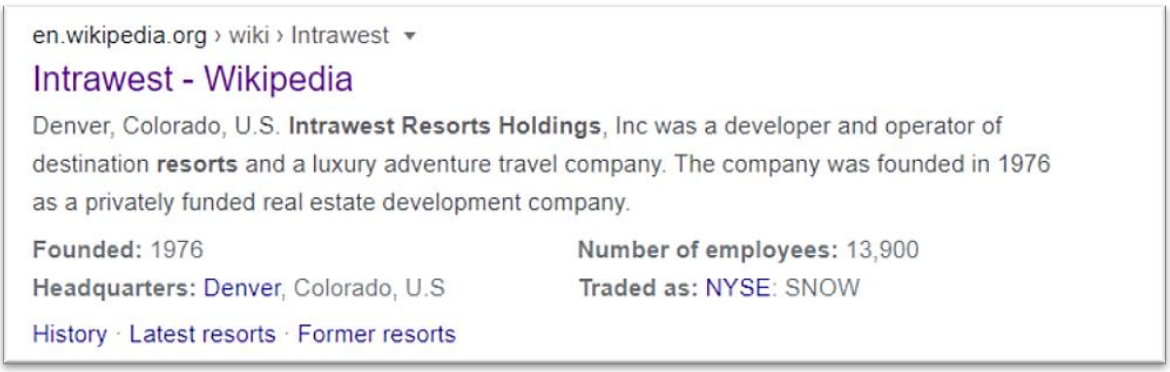
So Why Is This So Important?

We live in an ever more interconnected and complex world, where there are more and more complex securities and information about those securities. Even while the number of listed companies has been falling, the advent of derivatives, ETFs and indices means that a breakage in one place can have a ripple effect across the industry. There are now more ETFs and funds in the U.S. than there are underlying stocks. The whole value chain of financial services is dependent on tickers and symbology to work – so that pre-trade, trade, and post-trade processes work seamlessly.

But fundamental flaws in symbology still make for brittle systems across the industry. Downstream fixes and a lack of standards for these fixes means that there are fundamental inconsistencies in the output from vendors on which professionals and systems rely. Such ambiguity makes it hard for quants to build reliable models, while systems to process the vast amounts of unstructured data require high-quality entity data for the purpose of named entity recognition (NER). More generally, machine learning applications require clean data to improve the signal, so poor and patchy entity data will worsen results.

Perhaps a topical example will help make my point: take the recent listing of Snowflake on the New York Stock Exchange (NYSE) – ticker SNOW. You might be surprised to hear that as recently as 2017, SNOW was the NYSE listed ticker for a ski resort developer called Intrawest until it was bought by private equity group Fortress and delisted.

A Google search for Intrawest will give the impression that it's still listed with the ticker SNOW:



It's worse! – go back to 2000 and you'll find SNOW used to be IGN Entertainment – in 2002 it changed its ticker from SNOW to IGNX! There is also an Amsterdam listed stock that still trades with the ticker SNOW (the ISIN is NL0010627865).

The solution, of course, is not to use human readable tickers as your primary identifier for securities in the first place. Better, instead, to rely on the unique identifiers assigned to a security when it is listed or created. The human readable ticker and market identifiers should be metadata associated with such identifiers, along with other descriptive information (metadata) about the security. In most markets, this will typically be an ISIN or a CUSIP for the U.S. and Canada.

So, back to our Snowflake example – Intrawest's common stock has a CUSIP of 46090K109, whereas the CUSIP for Snowflake's Class B stock is TC8Q0SS64. Neither are very recognizable to the human eye – but to a machine they work.

I won't get into how ISINs and CUSIPs are licensed (here is a handy link that gives more information), but needless to say it is certainly not free for any systematic or business use of the data. Redistribution of such identifiers also needs to be licensed, which is why there are few, if any, free services providing this lookup service.

Most data vendors won't deliver reference or price data to a customer unless they attest that they have a CUSIP or ISIN license. What that means is that even if you have decided to adopt one of the commercial open identifiers like FIGI or PermID, you'll still struggle to map these back to the corresponding CUSIP or ISIN without that license to the vendor and the CUSIP Service Bureau.

To summarize, most security identifiers have their flaws, are tightly licensed, are expensive to use, and aren't easily cross referenced to each other.

Beyond Securities

Needless to say, as the world becomes even more complex, the need to uniquely identify other data objects and to link them to investible securities becomes increasingly important – perhaps the most important domain is that for companies. Again, lots of open and closed identifiers are below, neatly summarized in the following table from Alacra. You can outsource a lot of your cross-referencing needs to Alacra – for a price of course.

Figure 3: Alacra Entity and Security Identifiers

Source: Alacra

As mentioned, securities are issued to companies, and the Legal Entity Identifier (LEI) is a sound and open identifier for issuers and financial counterparties. GLEIF and the Association of National Numbering Agencies (ANNA) launched a daily file linking ISINs to their issuer's LEIs. It's a big file and too large to be loaded into Excel. While useful, it is just the ISIN and the CUSIP with no other metadata. It also currently doesn't include retired ISINs – so I assume it's not that useful to help disambiguate my SNOW example. GLEIF also provides a one-to-one mapping table to a company's BIC. Refinitiv also publishes a link between the LEI and the firm's open identifier PermID – which does open more possibilities.

Schemas to classify companies are also very challenging, notwithstanding most approaches to classifying a company into a sector require a one to one mapping. This is difficult for a firm like Apple, which can be categorized simultaneously as Software, Hardware, Gaming, or Banking! The public schemas are limited, and both private schemas (such as Global Industry Classification Standard, or GICS, and Industrial Classification Benchmark, or ICB) can be pricey and are not openly licensed. GICS (owned by MSCI) and ICB (owned by FTSE) are multi-tiered schemas and aren't that different in their approach. Other vendors have their own industry classifications to help reduce costs. For example, Refinitiv has a multi-level schema called Thomson Reuters Business Classification (TRBC), and the Primary Business Sector is mapped on permid.org (this field is only licensed for non-commercial use). Of course, the real value of these schemas is the mapping of them to the companies, which requires a process and people.

So Where Are We Heading?

I wish there was good news here – there isn't much yet. Despite the flaws and challenges I have identified, the industry seems very set in its ways. Exchanges allow the reuse of tickers and have outsourced the issuance and monetization of identifiers to commercial entities. Vendors have probably gone as far as they will go to open up their internal schemas. Mapping across these schemas is left up to the better staffed customers. There is no indication that these competing firms will collaborate in the space.

The good news is that with more and more advanced system availability to match entity data, the easier it becomes for firms to reliably spot errors and disambiguate entity data. Larger firms can afford to properly license datasets, so that larger vendors will deliver cross referencing files to them. Smaller firms, however, find the costs prohibitive or fly under the radar for as long as possible – and we know where that could end up!

Personally, I'd like to see tighter standards and a more modern approach to making source data more accessible and open. For example, it should be easy to query public data to identify securities that have been delisted. Ticker re-use should be disallowed. Yes, it's fun to have SNOW as a ticker – but was there any consideration as to the downstream consequences?

Source: SEC File No. 4-533, effective under rule 608(b)(3)(iii):

On August 24, 2015, Financial Industry Regulatory Authority, Inc.

(``FINRA''), on behalf [a group of US Exchanges],filed with the Securities and Exchange Commission [an amendment] unanimously approved by the Parties.\4\ The Amendment to the Plan proposes to revise Section IV(d) of the Plan (Reuse of a Symbol) to provide that, where a Party ceases to use a symbol, such party may elect to release the symbol and that such symbol may not be reused to identify a new security (other than the security that has been trading under such symbol) within 90 calendar days from the last day of its use to identify the old security, without the consent of the Party that released the symbol. In addition, a Party may not reuse (or consent to the reuse of) a symbol to identify a new security unless such

There are rules out there – in Canada "symbols previously used by other issuers cannot be reassigned for 53 weeks." In the U.S., reuse is permitted after 90 days, unless the change causes investor confusion.

Confused?

Advice for the Alternative Data Community

Needless to say, the more that an alternative data provider can do to ease the customer burden to test and integrate the data the better. Symbology is a major pain point so take the time to understand the space. If your data is at the entity level, append the PermID or LEI – it is open, and customers will be able to resolve to a stock ticker if required. Or better still, add the ticker's PermID and FIGI in there for good measure - the more the better.

Make sure you document your schema, and any adjustments you have made. Keep an eye on things too – as you now know – identifiers change through time!

Finally, leverage as many of the open standards out there as you can:

- Legal Entity Identifier (LEI): Good for banks and big companies and provides a cross reference to ISIN and BIC.
- FIGI (Bloomberg): Has become more developer-friendly and has gained traction across the industry.
- PermID (Refinitiv): Great API, and a matching tool that helps create common licensing. Includes LEI mapping.
- Open FactSet: Symbology service, for a fee. Although, I think for data partners, you will get a pass.
- Most other large vendors provide matching and cross-referencing services for a fee, and these are often bundled with other products – don't forget to ask if you are a customer!

I'm not explicitly saying avoid the fee-liable identifiers like the CUSIP and ISIN. It is just that they can be expensive, especially for a startup. There may also be cases where the use of CUSIPs or ISINs is unavoidable – for instance, when working with fixed income data – but there are approaches you can use to keep costs down and stay within the licensing regime. Also, just because a CUSIP is published on the SEC website, it does not mean you are licensed to use it or distribute it.

And finally

Here is Symbology – outside our front door (and a deer).

Any ideas expressed in this article are personal opinions and are not official views of IEX Group or IEX Cloud.

Eagle Alpha

Section 2

Named Entity Recognition

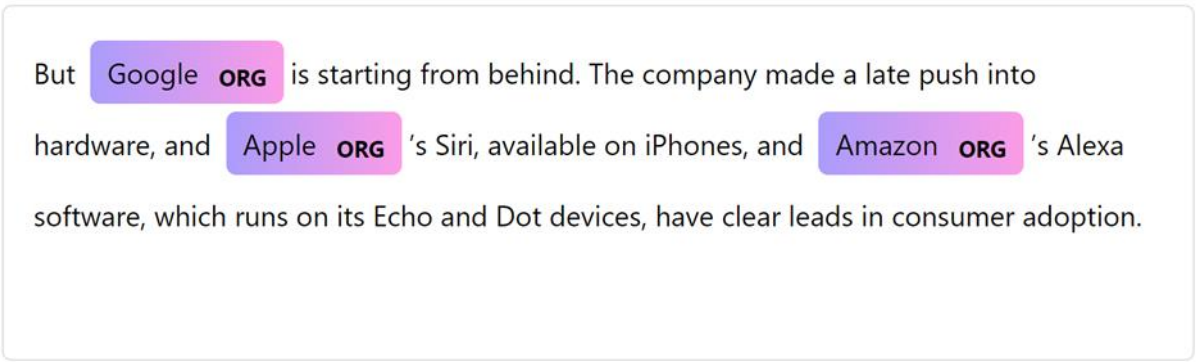
The importance of an entity master file and named entity recognition (NER), from the data science and analyst teams at Eagle Alpha.

What Is An Entity Master File?

As outlined in the previous section, symbology is a complex subject, and it can be difficult to map a dataset to identifiers, symbols and tickers. As Tim alluded to, a dataset requires high quality entity data in the first place. Alexander Denev and Saeed Amen say in The Book of Alternative Data, "regardless of the origin of data (individual, institution and sensors), making it useable requires it to be converted into a structured form" …and "name entity recognition is also key to identify proper nouns of interest, such as people, places and brands". A dataset needs to be mapped to company, place or person before you even start to map it to identifiers and tickers to extract value from it. This requires a dataset to undergo a named entity recognition (NER) process. Like symbology this is a complex task.

Constructing a NER model for a data source most often involves natural language processing (NLP) to extract entities by using a set of semantic rules about the structure of the text. For traditional raw data sources such as "plain text" language corpuses, there has been significant research progress made. Released in the early 2000's with continued support since, the Stanford Named Entity Recognizer uses a Conditional Random Field (CRF) model and provides general implementations and pre-trained models for extraction of people, entities and locations in multiple languages and operating systems. More recently, Google's open-source Bidirectional Encoder Representations from Transformers (BERT) and implementations of OpenAI's API-accessible GPT-3 language model have been shown to perform at state-of-the-art levels for many NLP tasks including Named Entity Recognition tasks, having the benefit of being trained on internet-scale text corpuses. Popular libraries in Python such as spaCy and NLTK contain a significant range of tools for practical entity recognition tasks and are often more user-friendly and widely supported than state-of-the-art methods. Groups such as Hugging Face focus specifically on translating the state of the art in NLP methods into open-source, easily-approachable libraries.

Figure 3: How spaCY Represents Entities



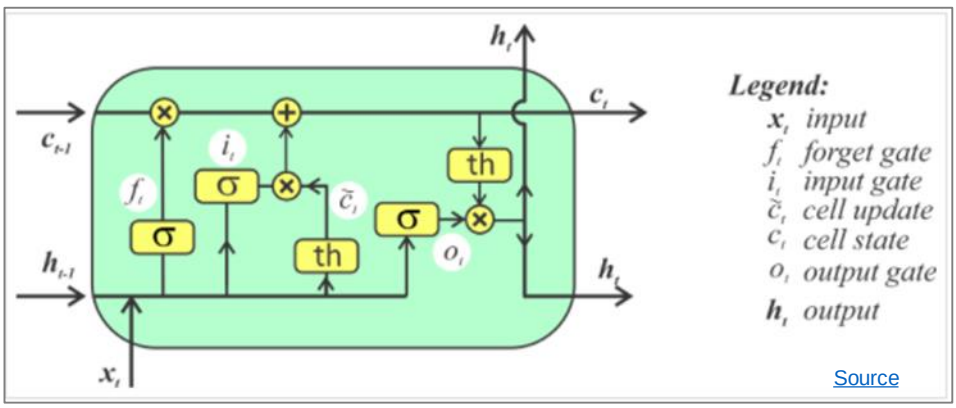
Example of "Organization" recognition in a sample text visualized with spaCy. https://spacy.io/usage/visualizers

It is worth noting that many of these advances train on "plain text" natural written or spoken language, and these models often break down when applied to uniquely abbreviated text, machine-generated text, or text styles not commonly available to the public – forms commonly seen in alternative data. Examples of these areas where it is challenging to find quality pre-trained models (or labelled training data) for NER tasks include email receipt text and card transaction descriptions, both of which often contain esoteric and inconsistent formats for presenting entities in text.

Named entity recognition is a classification task where an algorithm infers the class of a target word or phrase using information about a number of context words surrounding the target. Common word classes for NER can include company, person, organization, product, brand, city, or country among others. More complex models such as Recurrent Neural Networks, Long-Short Term Memory networks, and Auto-Encoder networks can utilize not only multi-word, but also multi-sentence and multi-paragraph context inclusive of word order, while simpler algorithms might only use the appearance of nearby words (regardless of word order) or part-of-speech classification to make assumptions about the type of the target.

As described here, there are advantages to LSTM and one of the main advantages "is that it can provide a constant error flow. In order to provide a constant error flow, the LSTM cell contains set of memory blocks, which have the ability to store the temporal state of the network. The LSTM also has special multiplicative units called gates that control the information flow". This is represented in the schematic plot below. One other key advantage in the learning process is that the forget gate facilities what should be filtered out and what should be remembered for the next iteration of input values.

Figure 4: A representation of a Long Short-Term Memory Network



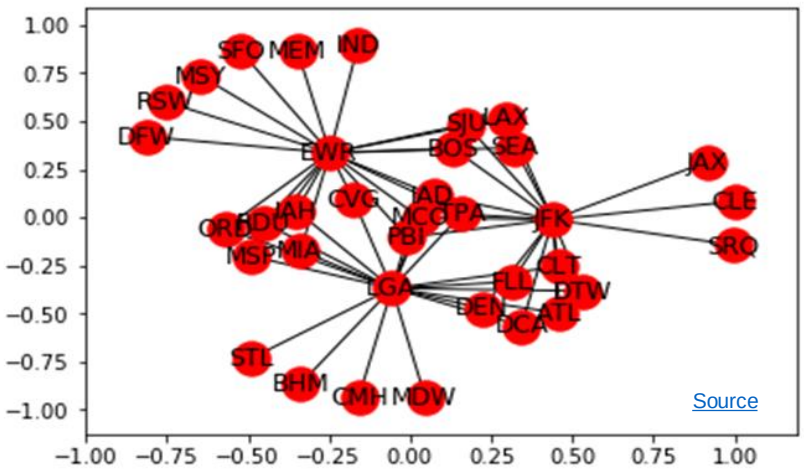
The next step in a NER pipeline, once the generic classification of words into entities is complete, is to link the identified entities to an internal knowledge base. As is often the case in natural language text, a single known company may be represented by a multitude of different variations of the company name, all of which could be classified as "a company" by a NER system. It is the task of an entity-linking system to disambiguate these variations into a single entity. Approaches for this disambiguation task range from simple fuzzy-matching to character and word embedding models that are commonly built up in an unsupervised manner on the training data once classified into types. The embedding approaches have the added benefit of allowing for unsupervised relationship-building between companies, products, people, and locations if the training data has sufficient adjacent mentions of these entities in the text. The downside of these approaches is the need to collect greatly more training data in order to use a truly unsupervised approach, therefore it is common to add a "human-in-the-loop" to associate products to companies in the reference entity knowledge base building upon an initial unsupervised attempt at association.

We spoke to Evan Schnidman, Founder and ex CEO of Prattle, who used a bi-directional LSTM to construct an NER for Prattle. The team at Prattle queried the web using NLP to identify named entities, pulling 5 million documents a day in the process. It took the company

nearly a year, with a dedicated lead data scientist, to complete the NER project. When completed, resource had to be dedicated to the project as a NER model needs to be retrained and applied on an ongoing basis to stay current. Like Prattle, the alternative data vendors we spoke to for this paper (see section 4) all constructed their own entity master file. For example, Owlin started with a base library of entities and ended up "building our own reference entity dataset". 1010 Data built their own entity recognition system as they wanted "full control of the process". LinkUp built their own entity database "in house" because when they started selling into the financial market "about 5 or 6 years ago, there weren't really any viable options" to manage their data assets.

Some of the third-party technology solutions providers we spoke to for this report use graph models to construct a NER file. Graph techniques have been around for quite some time and there is lot of literature on the subject. From this quick overview of the topic "graphs are mathematical structures used to study pairwise relationships between objects and entities". Probabilistic graphical modelling layers in Bayesian probability theory to accelerate the mapping of likely connections between the variables, or entities, in the graph. This is represented in the plot below that depicts how entities are interconnected. A practical example of a graph model might show how any product, brand or entity related to Old Navy, Banana Republic or Gap would ultimately map up to parent entity, The Gap Inc.

Figure 5: Representation of a Graph Model



The process is complex and can be costly to deploy and requires the right technology and human resources. Unlike LinkUp's problem from 5 to 6 years ago there are now software technology solutions providers in the market. Eagle Alpha has partnered with some of these solutions providers that use graph models and other NER model building. These vendors are quick to point out that building a robust NER model does not end with entity recognition and entity linking. Like Tim's discussion on the temporal nature of tickers there is a temporal aspect to entities that layers on another level of complexity. Simply put, entities are not static. There are many things that can influence an entity over time, such as M&A, spin offs and other corporate actions. Parent child relationships and minority or controlling stakes by large companies can also have an impact on an entity at any given point in time. The entity database may be "as at" rather than "as of". A NER that is "as of" is more useful to the buyside for back testing as it provides a view of the entity as of a particular point in time.

One needs to be careful of the symbol or ticker used for a company at a given point in time. To illustrate, Texas Instruments (TXN) acquired National Semiconductor (ticker NSM at the time) in 2011. TXN had a revenue run rate of $6bn and NSM $1.6Bn. An investor wanting to perform an analysis of TXN needs to account for the significant portion of revenue from National Semiconductor pre and post-acquisition. National Semiconductor becomes the entity Texas Instruments. If the dataset keeps National Semiconductor tagged to NSM it gets even more problematic, as the ticker NSM is now assigned by the NYSE to Nationstar Mortgage Holdings Inc. Of course, this all ties back to the earlier section on identifiers and symbology. Life would be a little bit easier if the exchanges did not reuse tickers.

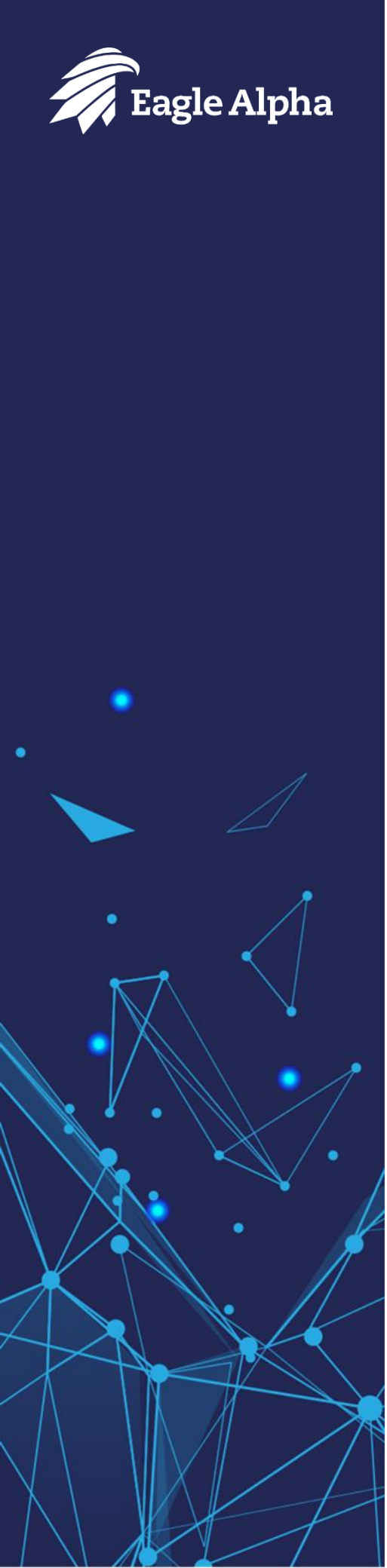
Of the technology solutions providers and alternative data vendors we have spoken to for this report there is universal agreement that advanced NLP, ML and a good technology stack helps resolve the NER conundrum, but it only gives you 80 to 85% of the answer. The other 15-20% requires “brute force” and human input. For the human element analysts are required who have industry and domain knowledge. These analysts are also needed on an ongoing basis in the maintenance of the NER file. Another aspect in the construction of a NER master file is that identifying larger companies is easier than smaller companies. Using NLP on unstructured web data in a NER model requires lots of data and smaller companies are intuitively going to have less public data. One last related note is that private companies are harder to get information on and are harder to map in the entity file. Mapping private companies is still needed, however, and these entities also need to be tagged with some sort of an identifier. Private company data can be as useful as public company data to the buyside and it can also be used by corporates and private equity.

Eagle Alpha

Section 3

Alternative Data Vendors

This section discusses the experience and learnings of some selected alternative data vendors.



Q&A With A Selected Alternative Data Vendors

In this section we collate firsthand knowledge of four alternative data vendors and their experience with ticker mapping, how they approached the task and how they resolved any issues that arose during the mapping process. We have asked the vendors a series of questions relative to building an entity master file and mapping to tickers.

Data Vendor #1 - 1010 Data

Founded in 2000, 1010data's core business centers around its industry-leading cloud-based analytics platform. More than 850 of the largest Retail & CPG companies rely on 1010data's technology to power their own business, transforming large data into insight.

Data Vendor #2 - Yodlee

Yodlee is a market leader in de-identified consumer transactional spend data with a history dating back to 2011 and includes bank, credit and debit card data providing a user centric near-real time 360 view of consumer/economic activity.

Data Vendor #3 - LinkUp

LinkUp delivers aggregated labor market data from company websites in the US and worldwide. The dataset is of high quality meaning it has no duplicates, no job pollution, no expired listing, and is updated daily. The dataset has a history to 2006.

Data Vendor #4 – Owlin

The Owlin News NLP analytics platform provides finance professionals with intelligence on all events relevant to their portfolio, business or peer environment — enabling them to monitor risk, ESG, opportunities and trends from 3 million sources across multiple languages in near real-time.

Question # 1
Did you enter the market with raw untickerized data? How quickly did you begin to address the buyside requirements on this and how long did it take to develop a mapped dataset? What resources were applied and what is the ongoing resource commitment?

1010 Data: No, we entered the market with merchant/company tickers. We built up the mapped ticker list over time. When we entered the market we had more limited tickers built, under 100 companies. Through time, we enhanced our tagging and processing capabilities and now have thousands of tickers built out, creating more on a continuous cycle.

Yodlee: We started this activity three years ago, looking at 300 tickers based on market cap. We looked for listed, and delisted dates etc. We looked at a number of available services but there was no standard information and nothing fit our need or desired template format. We realized the need for continuously monitoring and adding tickers to our tracker to derive the most value.

This is an ongoing process. We have chosen to keep enriching extra dimensions for the years we currently have. For these tickers, we maintain various standardized meta-data fields. SIC, NAICS, GICS, ISIN, SIDOL, exchange on which the ticker is/was traded are some of the examples of how we have enhanced this database. Our updating process is a combination of semi-automated and manual processing.

LinkUp: We began selling our data into the capital markets in about 2014/2015 after having sold our data to corporations and human capital management companies for about 4 or 5 years prior to that. Until that point, we hadn't had the need to map our data to entities or tickers but as soon as we started talking to asset management firms, it became clear that mapping would be a critical aspect of our data solutions.

We partnered with a third-party firm and essentially outsourced the mapping of our data to them because they had a lot of experience in working with and selling market data within the capital markets. Through a combination of people and technology over the course of about 6 months, we used the company URLs that we had in our dataset to map to entities and tickers.

Owlin: We entered the market with raw untickerized data. We developed a library of entities with our own internal reference IDs and matched these to entity names corporate website URL.
-Next step was to map this database to company's internal IDs.
-We currently support ISIN, tickers, LEI, and any other (custom) IDs
-We are building an authoritative reference master, which means that we can easily extend our coverage and easily increase the reference data libraries, e.g. with CUSIPs, and so on.
-We have automated extraction methods, and a data analyst team for curation and QA, to ensure proper matching and accuracy.

Question # 2
Building a named entity recognition (NER) database is the most challenging aspect of tickerization. How did you go about this, did you construct a NER master file yourself or use a third-party source?

1010 Data: *We built our own ticker mapping* (entity recognition) *system. We wanted full control of the process and used our strong data science resources coupled with our in-house analytics platform to do so efficiently.*

Yodlee: We constructed a master file from multiple sources such as the Securities and Exchange Commission website. We have used some NLP, incorporating a Fuzzy match algorithm to get an approximation of the ticker. As this is structured data, we have not leveraged NER in our work.

LinkUp: We essentially built everything in-house because, at that time about 5 or 6 years ago, there weren't really any viable options for alternative data providers to map their data easily. It was really the earliest stages in the evolution of the alternative data industry, so we had to build things from the ground up using Refinitiv's PermID and OpenPerm framework. As our product has evolved over the years, we've added additional 3rd-party data and are now working with a number of firms such as FactSet and S&P, in addition to Refinitiv, to provide additional reference data to make it easier for firms to use our data.

Owlin: We started by building a base library of entities that contains the essential information to handle entity related data in an agnostic way, regardless of whether we're working structured data like pricing or unstructured data like blog articles.
-Once an entity has been setup it can be enriched with reference data from 3rd party providers like IHS Markit where we have integrated both CDS pricing and RED in our platform.
-Each entity in the library contains legal names, tickers, subsidiaries as well as short names or how the general audience calls it in the street.
-This means that once an entity has been setup in the library, we are able to match datapoints using long names and identifiers but also short names or brand names.
-The rationale behind building our own reference entity dataset is that it is core to our future data offerings in order to match (inhouse build) data and signals with the right entity as well as being able to facilitate 3rd party data on our platform.

Question # 3
When applying the NER database to ID's such as ISIN/CUSIP/SEDOL and then tickers how do you solve the point in time aspects of ID's and tickers (M&A, Parent-child, survivorship bias etc) rather than just having a current ticker?

1010 Data*:* We maintain an acquisition/divestiture mapping table that is available to all clients. We track, monitor and update for corporate actions for the tickers we cover. Additionally, as we provide point in time data, the history of 'deprecated/changed' tickers can be tracked.

Yodlee: We check our data constantly, using a combination of SEC investigation of filings, press releases and alerts. There is no perfect source for this, but we double check across various sources.

We maintain the temporal aspect on which all of this information has been tracked so if a ticker is delisted or relisted, we can track it. All the events that affect ticker-merchant mapping are tracked with the corresponding date. We can flex up or down on resources as it's directly proportional to the frequency of the data being published.

LinkUp: After the initial mapping exercise in the earliest version of our raw product for the capital markets, we realized pretty quickly that we did need to create a point in time mapping file so firms could better evaluate and integrate our data. That took about another 6 months to complete. We had a huge advantage with our data in that it went back to 2007 which is great for backtesting but it created kind of a nightmare for building our point in time files.

In the past two years or so, we have established strategic partnerships with both FactSet and Refinitiv in conjunction with making our data available on the OpenFactSet platform and Refinitiv's QAD platform. Both firms have been phenomenal to work with, and as part of that process, we have been working really closely with their teams to strengthen even further the quality and thoroughness of our mappings and point in time data. We've also been working with the team at ThinkDataWorks around a number of initiatives to keep improving our data products and making it better and easier for clients to use our job market data.

Owlin: For listed names we have a variety of corporate action feeds to first ensure that the right identifiers get updated with the right date stamp.
-On top of that our analyst team double checks the impact of a corporate action to ensure that the data for the related companies in the hierarchy are still up to date.

Question # 4
Is there any technical aspects or data structure that you think a vendor should have in place at the outset to satisfy the buyside and also help map the data to tickers?

Yodlee: We have a monthly file that we can provide to clients. There isn't typically a lot of change on a daily or weekly level, but monthly update is sufficient. We have shared the information as a flat file, but we want to develop a database where a client can query it. This way, we have a central server so everyone who needs it has the latest view.

LinkUp: Because of everything that has taken place over the past 5 years or so, it really is mandatory that data providers deliver their data to buyside firms in a form that allows their clients to easily evaluate and integrate their data. The specifics around what that looks like are entirely dependent on the data, the use cases, and the types of firms that are like to find value in the data, but at the highest level, it is just not the case that a data vendor can ask or expect buyers to shoulder the load around mapping data to entities and tickers.

Owlin: Flexibility to allow mapping to various IDs: tickers, LEI, CUSIP, ISIN, but also more basic IDs such as Company website URL.
-Providing service to buyside buyer and support with mapping internal IDs with vendor landscape.

Question # 5
Can the providers of IDs, tickers and entity databases do anything to help make the process easier? Is there a bottleneck anywhere that these providers could fix, potentially?

1010 Data: Not from our perspective.

Yodlee: Not from our perspective. None of the providers gives a guarantee on the quality of the information, so we know we have to double check everything. We are constantly scouring for updates and don't know if others do the same. It is not something any services we considered have provided.

LinkUp: I'm not sure, from an economic standpoint, whether it would be a good business or not, but there is absolutely no doubt in my mind that there is a tremendous opportunity in the market for someone to solve the massive set of challenges around mapping alternative data and then

organizing and managing the licensing of basically all available reference data. As much as things have advanced in recent years, there is still a huge need in the industry for an independent, centralized service bureau to reduce or eliminate the complexities around mapping and efficiently orchestrate the licensing arrangements between everyone in the market - market data firms, data vendors, data buyers, and all of the other 3rd party firms that are working with and building products around alternative data.

Owlin: Providing new means for improving accuracy and reducing false positives while mapping legal entity to correct ID, especially when names are ambiguous.

Question # 6
In hindsight, or if you had an opportunity for a "do-over", what would be the one thing you wish you could change or have had in place before you started your tickerization project

1010 Data: Aligning tickers better across datasets, which we have subsequently solved. For example, in our Credit/Debit card data, we created a ticker for Home Depot as HD. For our Visits dataset (off-line only dataset), we also had the ticker as HD. We should have made the ticker HD_OFFLINE for the Visits dataset, so the ticker would be explicit in what it represents.

Our credit/debit dataset now has three Home Depot tickers, HD, HD_ONLINE and HD_OFFLINE. Our Visits dataset now has one Home Depot ticker, HD_OFFLINE. You can match up the online/offline breaks.

Yodlee: Initially, we had a limited set of merchants of interest. We are tracking now over 800 listed and delisted names. In hindsight, we wish we had started a larger set of merchants with a complete set of information and then drilled down into subsegments.

LinkUp: There's a ton we'd do differently knowing what we know today about mapping and reference data. It's really obvious, but it would have been so much easier, 5 years ago, if things were where they are today. And of course, that's also true about where things are today compared to where they'll be 5 years from now. But that's exactly what makes it so much fun to be in an industry that's growing and evolving as quickly as the alternative data space. So in that regard, I guess I wouldn't change a thing.

Eagle Alpha

Section 4

Technology Solutions Providers

A matrix of technology solutions providers that provide different degrees of ticker mapping services.

The matrix below outlines the capabilities of the third-party technology solution providers (SP) that Eagle Alpha has engaged with. Capabilities vary by provider and we can provide more detailed information on request.

Figure 6: Technology Solution Providers

	SP #1	SP #2	SP #3	SP #4	SP #5
Build Entity File (NER)	✓	✓	✓	✓	✗
Can address all types of alternative dataset	✓	✓	✗	✗	✗
Perform entity PIT	✓	✓	✓	✓	✗
Map to ticker	✓	✓	✓	✓	✓
Map to ticker PIT	✓	✓	✗	✗	✗
Private company NER	Limited	Limited	Limited	Limited	✗
Map private companies to an identifier	✓	✓	✓	✓	✓
Mainly English language	✓	✓	✓	✗	✓
Non-English	Limited	Limited	Limited	✓	✗

Source: Eagle Alpha.
Named Entity Recognition (NER) – Point in Time (PIT).

SP #1 – Technology solutions used to convert unstructured data into a structed data. This data is overlaid with a robust dataset of corporate events for a point in time NER.

SP #2 – Technology vendor with financial service and corporate customer base. Provides point in time NER for quants, data analytics and alt data vendors.

SP #3 – Vendor uses technology and trademark/patent database to map entities to and NER. Not all alternative dataset are suitable for this solution.

SP #4 – Technology solution using structured and unstructured data to map entities.

SP #5 – Vendor applies technology stack to map symbology and ticker where a dataset is already mapped to entities.

Eagle Alpha

Section 5

About Contributors To This Report

A brief outline of the main contributors to this report

About The Contributors To This Report

About Eagle Alpha

Established in 2012, Eagle Alpha is the pioneer connecting the universe of alternative data. First adopted by alpha-seeking hedge funds over 10 years ago, alternative data is now being sought for use in the wider asset management space, as well as the private equity and corporate verticals.

In parallel, there is an explosive increase in the supply of alternative datasets, as many corporates are looking to monetize their exhaust data and new technologies enable the emergence of new alternative data vendors.

Eagle Alpha was one of the first companies to recognize the value from these new data sources and has been investing in educating and connecting alternative data vendors and buyers since 2012, in the process building trusted relationships with both sides of this market.

As of October 1st, 2020, Eagle Alpha partners with over 1,398 data vendors and hundreds of data buyers across the asset managers, private equity and corporates.

Eagle Alpha's solutions mirror the user journey of our customers. There are three steps in the vendor user journey: discovery phase, productization and go-to-market. There are three steps in the buyer user journey: data strategy, discovery and prioritization, delivery and insights.

Our unique breadth of datasets, knowledge of the industry and customer relationships have cemented Eagle Alpha as the global leader and strategic partner in the data space.

Eagle Alpha partners with industry leaders to continue to shape the industry:

1. J.P. Morgan, lead sponsor of our data conferences.
2. FISD, member of this association to create standards for the industry.
3. Lowenstein Sandler, partner with this US law firm.

About The Contributors To This Report

About IEX

IEX Cloud, a financial data delivery platform and subsidiary of IEX, delivers a modern API and powerful tools that enable developers to build financial applications with ease.

IEX was founded in 2012 and started by creating a new stock exchange – IEX Exchange – which is built to work for all market participants. The story behind the founding of IEX was chronicled in Michael Lewis' 2014 book Flash Boys. IEX Cloud, which is separate from the Exchange, was launched in 2019 with the mission of making financial data more accessible.

Tim Baker is Head of IEX Cloud.

Eagle Alpha

CONTACT US

Emmett Kilduff

Founder & CEO
Eagle Alpha

emmett.kilduff@eaglealpha.com

Emmett worked in Investment Banking with Morgan Stanley and Credit Suisse. Morgan Stanley was the first investment bank to work with big data within its research department.

Brendan Furlong

Senior Analyst
Eagle Alpha

brendan.furlong@eaglealpha.com

Brendan supports the Data Analytics and Data Strategy team at Eagle Alpha. Prior to joining Eagle Alpha Brendan spent almost seventeen years working in New York on both the buy-side and sell-side as an equity analyst.