An Information Provider's Wish List for a Next Generation Big Data End-to-End Information System

Mona M. Vernon
Thomson Reuters
22 Thomson Place
Boston, MA 02210, USA
mona.vernon@thomsonreuters.com

Brian Ulicny
Thomson Reuters
22 Thomson Place
Boston, MA 02210, USA
brian.ulicny@thomsonreuters.com

Dan Bennett
Thomson Reuters
610 Opperman Drive
Eagan, MN 55123, USA
dan.bennett@thomsonreuters.com

ABSTRACT

As the leading source of intelligent information, Thomson Reuters delivers must-have insight to the world's financial and risk, legal, tax and accounting, intellectual property, science and media professionals, supported by the world's most trusted news organization. In this paper we describe a recent initiative at Thomson Reuters to establish Big Data infrastructure to store most of this data, and to apply open standards for linking the contents of this data. We describe our wish list of technology not currently available from open source and commercial solutions for a next generation big document and data system.

Categories and Subject Descriptors

D.4.3 **[Software]**: Operating Systems - File Systems Management

General Terms

Management, Performance, Design, Data Management, Open Data

Keywords

Big Data; Open Data; Enterprise Information Management; Innovation; Data Management

1. INTRODUCTION

As the leading source of intelligent information, Thomson Reuters delivers actionable information to the world's financial and risk, legal, tax and accounting, intellectual property, science and media professionals, supported by the world's most trusted news organization. Since Thomson Reuters customers are professionals, they rely on precise trusted information, which in most cases is enhanced or curated by human experts. Recently, the company set an enterprise strategy with a focus on new enterprise opportunities in addition to those within each business unit. To align the data management strategy to the business strategy, Thomson Reuters launched an initiative, called the Big Data Initiative for the purpose of this paper, to store all Thomson Reuters data, and adopt open standards for linking the contents of this data. This initiative has, for scope, Thomson Reuters' massive and varied data sets and repeatable data management processes for keeping the resulting "Data Lake" and its derivative products up to date. In this paper, we describe some of the technologies employed to ingest, store and link the various data sets across Thomson Reuters in this initiative. We also discuss the challenges with the current state and the future technologies potentially useful to address these challenges.

2. Thomson Reuters' Data Management: Some Historical Context

In 2009, the idea of a "Content Marketplace" (CM) [8] started to develop on the heels of the Thomson Reuters merger, which presented the opportunity to think about not just combining Thomson Corporation and Reuters Group information assets, but combining them in a way that made them completely enabled and interoperable with each other.

Thomson Reuters acquires, editorially curates and delivers a vast sea of business information, organized around entity categories, such as:

- Organizations
- Financial Instruments
- Quotes
- Estimates
- Patents & Trademarks
- Briefs, Dockets & Case Law
- People, etc.

Symbology is a term used to relate public and well-known identifiers to these entities either directly or via inference. We actively use symbology for entity instance coverage analysis, quality checks, and visualizations of entities in the Content Marketplace. A proprietary infrastructure enables quick and efficient navigation, linkage and accurate error resolution as well as generation reports across the range of our families of entities.

A consistent, robust, accurately and reliably updated entity aggregation protocol is an enterprise-wide core requirement that enables Thomson Reuters products to provide accurate and intuitive content navigation. Thomson Reuters currently provides both deep and broad content across a large spectrum of information at considerable cost. To find and use this information, there must be a working mechanism that can easily search, navigate and successfully find and retrieve any and all desired information or content. This functionality far exceeds typical full text searching and enables customers to understand and interpret the complex relationships of organizations in today's financial and professional landscape.

Without fast and accurate content delivery, the customer assumes that content is either missing or that the connections to the content

7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA.

are in error. Entity aggregation makes productization of symbolic resolution services possible. It is critical to deployment efficiency that these mechanisms be a global shared service across the enterprise, available to all systems, including both content and products. By providing a shared global service, products can fully leverage our data in the generation of its workflows, displays, etc., to our end customers, who will be delighted to have the same data available to synchronize with their internal reference sources. Other applications can be developed to manage data quality and customer queries more efficiently.

The Content Marketplace is Thomson Reuters' Information Architecture strategy and approach to federated master data management prior to the Big Data Initiative.

The Content Marketplace was conceived as a means of standardizing the otherwise diverse and fragmented methods for organizing curated content and delivering it from master databases to product platforms. The Content Marketplace enables otherwise disparate databases, both internal and external to Thomson Reuters, to distribute interoperable and easily commingled information.

The Content Marketplace is built on a uniform set of policies, a common language of content and a standardized set of interfaces. At the core of connected content is a well-constructed and managed entity model. Each entity type and content set has a clear and single owner. Each entity instance is uniquely identified, and all content that is published references the mastered entities. Every published data item is registered in a metadata registry with a clear definition and can be traced back to its origin. The distribution is standardized through a finite number of approved delivery mechanisms or Strategic Data Interfaces (SDI). A Strategic Data Interface is an XML-based publishing standard for content that is published to the Content Marketplace. Each Thomson Reuters internal content publisher implements one or more SDIs that comply with the policies of the Content Marketplace and derive from base publishing schemas defined by the Marketplace. The full definition of each SDI is mastered in the Data Item Registry (DIR), a metadata store. Thomson Reuters internal content producers are able to use the metadata registered in DIR combined with the schema structures to derive their own data models to ingest content on an ongoing basis. Policy and governance were used to ensure that products used the content as intended. This marketplace is built in a federated way; each content group delivers their piece of the content puzzle based on their domain specialization. This poses a benefit where domain specialization can remain distributed and a challenge since all parties need to contribute to obtain the benefit to the whole in such a federated approach.

In addition to Content Marketplace, prior to the current Big Data technology stack, Thomson Reuters deployed distributed systems for searching and retrieving information but not distributed computing systems for transforming or analyzing data. Novus is Thomson Reuters' distributed, cloud-like search architecture, patented in 2006 [1]. The Novus architecture provides a single platform for supporting online services from each of the four Thomson Reuters market groups. More than thirty applications use the Novus architecture.

The proprietary Novus system is a distributed search architecture that uses thousands of Linux servers each running custom software. Each search server is responsible for part of the overall content index, which fits in system memory so it can be accessed extremely quickly. When a search is executed, it hits thousands of machines at once. The results are sent back to a controller, which sorts them, aggregates, ranks and sends that information back to the requesting application. This provides sub-second search performance.

The application then decides whether it wants to pull the documents identified in the search. The content stores typically aren't actually touched until a document is requested. The content itself is stored using hundreds of Oracle RAC databases, typically with four nodes per cluster. Each cluster holds a subset of the total content.

While systems like Novus and the Content Marketplace, Thomson Reuters made storage and retrieval of its content distributed, they did not provide a Big Data infrastructure for performing computations on the data.

3. A Big Data Initiative

Thomson Reuters' Big Data initiative is an enterprise level infrastructure (Figure 1) that aims to gather, aggregate, connect and disseminate Thomson Reuters large and varied content in addition to allowing useful linking with customer, external partner and open web content. The Big Data initiative enables automated content collection, leverages commodity hardware storage, and increasingly cheaper processing. By enabling distributed, scalable computation alongside the data, this infrastructure provides tools to enable new insights into the data through machine learning, predictive analytics, visualization and natural language querying. This initiative is a response to changing customer needs and expectations with respect to integrating multiple content sources, including customers' own proprietary, third party, open and social media data. The ultimate goal of the Big Data initiative is facility building analytics and visualization on top of diverse data sources.

The contents of the Content Marketplace and other datastores, such as Novus, will be regularly ingested into the Data Lake as shown in Figure 1 so that analytics at scale can be done on the most up to date information. This information will be supplemented with non-Thomson Reuters data that is also regularly updated. For example, maintenance records on patents that are supplied by patent agencies will be ingested as it becomes available.

The Data Innovation Lab at Thomson Reuters was recently stood up to take advantage of the opportunities posed by ingesting this content into a Big Data infrastructure. An initial focus of the Lab has been doing Big Data analytics involving several data sets from Thomson Reuters and open data to produce interesting analytics in the domain of intellectual property. These large scale analytics, which are called here as a shorthand patent analytics, can be produced by combining patent information with information that straddles data sets associated with distinct business units such as company financial information, legal proceedings, academic research and open data. Combining patent information with additional cross-business unit information is a good way to demonstrate what is now possible given a common, scalable computing environment in which analytics can be readily performed. As an example, various indicators of the risk that a patent will be litigated can be correlated across disparate data sets and used to train and test a machine learning model of litigation risk, validating and extending smaller scale academic research in this area.

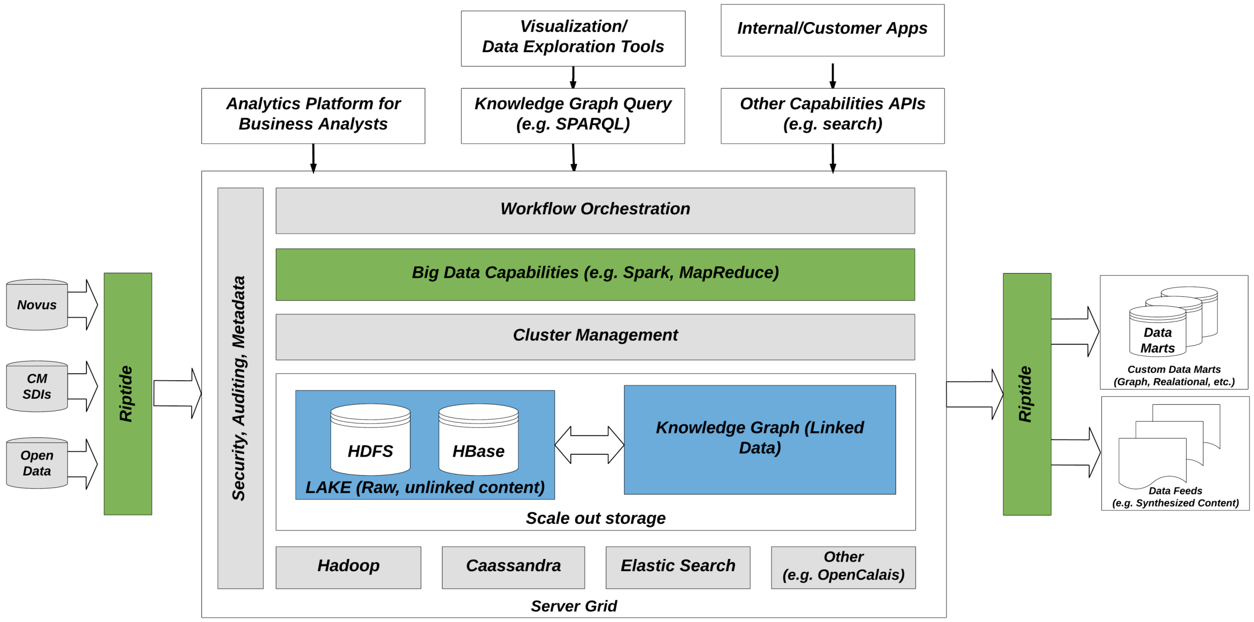
Visualization/
Data Exploration Tools
Internal/Customer Apps
Analytics Platform for Business Analysts
Knowledge Graph Query (e.g. SPARQL)
Other Capabilities APIs (e.g. search)
Workflow Orchestration
Big Data Capabilities (e.g. Spark, MapReduce)
Cluster Management
Security, Auditing, Metadata
Novus
CM SDIs
Open Data
Riptide
HDFS
HBase
LAKE (Raw, unlinked content)
Knowledge Graph (Linked Data)
Scale out storage
Hadoop
Caassandra
Elastic Search
Other (e.g. OpenCalais)
Server Grid
Riptide
Data Marts
Custom Data Marts (Graph, Realational, etc.)
Data Feeds (e.g. Synthesized Content)

Figure 1 One Possible Future View of the Big Data Initiative Architecture

With this Big Data initiative, Thomson Reuters is extending the master data management and service oriented architectures approach that have been in place in order to meet four objectives:

1) A desire across Thomson Reuters to use new Big Data technologies to solve existing and new problems.

2) Overcoming difficulty in accessing data from across the entire organization. Data from around Thomson Reuters can still be siloed in stores dedicated to the products that deliver it.

3) The fact that applying Big Data technologies requires a lot of perspiration before the innovation can happen. Thomson Reuters wants to provide a sandbox that business units can jump into and use.

4) The challenge of emerging open data standards. The rise of open data with emerging de facto standards such as opencorporates.org, schema.org, in addition to commercial standards such as OMG's Financial Industry Business Ontology (FIBO) poses an opportunity: What data does Thomson Reuters have that can be opened up to drive customer innovation? How do we deliver to our customers the ability to leverage open data and link to many types of external data and internal data accurately?

The Big Data initiative does not throw away the work done in the Content Marketplace and similar efforts for content management. Rather, it uses this content as input to a Big Data computing environment. Within this Big Data initiative, Thomson Reuters is building four key capabilities:

1) A Data Lake of mostly diverse content available in Hadoop, with automatic ingest from Novus, Content Marketplace and other potentially external sources as necessary, including open data. Business units can freely access the content and schedule their own jobs to run on the content as it becomes available.

2) A Linked Data store, the Knowledge Graph, which represents our knowledge of the interconnectedness of entities within the content.

3) Data Marts that take slices of data from the Lake and Graph to solve particular business problems. Neither the Lake nor the Knowledge Graph can be optimized for every possible use case. The Data Marts solve for this.

4) Metadata as a Service (MDaaS), a web application and associated APIs to support open data.

The Lake is (currently) a Hadoop-based consolidated data store of content with compute alongside store of Thomson Reuters content that enables Map/Reduce and SPARK processing. The Lake automatically ingests Novus, Content Marketplace and other content of interest via a landing strip using a scheduler for periodic or on-availability jobs. An ingestion system based on the Apache Kafka [2] messaging framework is used to load data into the Lake.

Above the Lake is a job scheduling capability, allowing developers to configure jobs on a per content type basis: for example, the scheduler may be programmed to run a certain set of jobs every time a new patent is available. It is envisioned that there will be a library of pre-existing jobs for entity extraction, mapping strings of text to entities in the entity master lists. Every loaded data item is assigned a URI for downstream representation in the Knowledge Graph. Currently, no customer data or PII is being ingested into the Lake. Big Data infrastructures raise many familiar and novel privacy concerns that Thomson Reuters is committed to handling responsibly. [4].

It is important that processes and technologies for ingesting content into the Lake are implemented to ensure the repeatability of the content ingestion. In order for this to be successful, it must be the case that all the content in the Lake is kept up to date in a reliable way.

A typical workflow for a user of the Lake might go something like this: Using a search interface to Novus, or other repository, someone interested in processing data using the Big Data infrastructure could use GUIDs associated with a particular entity of interest to identify content collections that contain the relevant GUID. Once the collection is identified, sample XML can be examined. XML elements of interest can be looked up in entity masters to understand what they mean. Then, the data can be ingested into the Lake. Once there, a Hive query can be used to extract attributes of interest by means of XPath specification of the elements of the XML. The extracted and columnized data can then be used as the basis for visualizations, joins, downloaded or used for further analysis.

4. OPEN LINKED DATA & KNOWLEDGE GRAPHS

The Knowledge Graph is a highly scalable link data store, using REST standards for input and output, built on Cassandra and Elastic Search over a cluster of servers. The Knowledge Graph represents relationships between known entities extracted from the content of the Lake as well as those represented within Thomson Reuters' authority databases for people and organizations initially. The Knowledge Graph enables content retrieval by URI, Search, and pub/sub mechanisms. The ambition here is to capture the knowledge encoded about entities in our content organized around core entity types like company and represented in the form of a graph, rather than as tables or XML. While the initial Knowledge Graph may be derived from more or less static authoritative data that the company maintains on key entities, it is hoped that the rest of the business units will augment these relationships with additional relationships, building out a much more robust graph of all the known entities within the scope of Thomson Reuters data.

The problem of mapping text sequences with entities in a master entity list in order to populate such a Knowledge Graph is extremely common with Thomson Reuters data. Normalizing references and identifying the specific entity they denote in context has many subtleties. In addition to internally developed solutions, Thomson Reuters investigates new tools for concording text references to known entities [5].

With Metadata as a Service, the motivation is that customers would prefer to use a 3rd party master data as long as it is under a sufficiently open license. Thomson Reuters maintains entity masters for people, organizations, instruments and a variety of other entity types. If these master entity lists are exposed as open data [3], then customers can link their data with Thomson Reuters content via these authorities. Tools for managing this data will be provided, including Web services for search and retrieval as well as via integration with our named entity recognition software, called OpenCalais [6].

Graphs that we build externally to the central Knowledge Graph can either be explicitly joined to the Knowledge Graph by augmenting the current data store and joining on the identifiers of the entities (e.g. persons and organizations) already in the Knowledge Graph. Alternatively, two or more graphs, including the central Knowledge Graph can be federated for the purpose of querying, using the search federation functionality of SPARQL 1.1. Alternatively, semantic web virtualization techniques can be used to query data stores that are not even triples based, using the Relational to RDF Markup Language (R2RML) to mediate the translation of SPARQL to SQL and back again.

One of the use cases that the Data Innovation Lab is pursuing here involves extracting large-scale graphs from the large datasets we are ingesting into the Lake. We want to be able to do graph-based analytics on these representations of relationships between entities. Centrality metrics (beyond simple degree centrality) and path-based queries are of particular interest to us.

In the Data Innovation Lab, we have been having difficulty getting graphs of a size that it is easy to extract from our data into a graph database that is responsive in reasonable time. For example, just to represent the citation relationships between patented inventions (where a single invention might contain multiple patent documents, filed in different countries or jurisdictions), using the data in the Derwent Patent Citation Index [7] going back only to the late 1960s, we extracted 180 million simple citation relations. Path-based queries such as "select all the patented inventions that are at least one citation hop and no more than ten citation hops from an invention assigned to Google" are difficult to ramp up to the size of the graph we need.

5. ENVISIONED FUTURE STATE

While it is valuable to be able to perform computations easily at scale in the Big Data architecture we have sketched and to extract relationships from that data into a Knowledge Graph, that can be searched, queried, and analyzed alongside other graph-based representations of data, joined on common identifiers, some additional capabilities are desired.

INFORMATION INTEGRATION: A major issue with combining datasets for analytical purposes is identifying the same entity across content. For internal data that has strictly followed our entity management protocols, this is not an issue; each instance of an entity type has a unique identifier, and all the XML elements that use that identifier are well-documented. However, when not even all our internal and diverse data strictly conforms to our protocols, external data cannot be expected to conform to it. Internally, not all data conforms to Content Marketplace protocols because these protocols are a top down initiative that requires significant upfront investment that not all business units are willing to make for all content sets without the future applications that would benefit from embracing the Content Marketplace protocol. One benefit of the Data Lake is that it can potentially ease business groups into that making the necessary upfront investment to conform to best practices by showing some of the benefits earlier.

The result of non-conformant information is that considerable effort must be spent on identifying the same entity in order to integrate information prior to analysis. This problem cannot be reduced to merely some kind of fuzzy string matching. Entity resolution, or concordance, can involve understanding the part-whole structure of entities and changes that it undergoes across time and space. As an example from the patent analytics domain, non-practicing patent entities (sometimes disparaged as "patent trolls") often wish to obscure their patent portfolios. This is done in various ways, including using a multitude of subsidiary or affiliated organizations as assignees of inventions as well as deliberate misspelling of names on documentation, so identifying the owner of a patent can require considerable effort in understanding the relationships of assignees to known organizations. Additionally, reassignments for patents from one owner to another are sometimes but not always registered with patent offices. In other cases, the transfer of patents can be inferred from merger and acquisition deals. Again, mapping the parties in mergers and acquisition deals to the patents they hold is non-trivial, and to be accurately used in predictive models, these relationships must be represented in time, in addition to representing the most current owner. Further, the ultimate parent organization of a patent assignee organization must be reflected in the organization entity model in order to be able to accurately represent the patent portfolios of entire companies, and not just subsidiaries of companies.

ACCESSIBILITY: Currently, data is being ingested into the Lake, but it is difficult to know what data has been ingested without consulting ingestion schedules.

The value of data is directly proportional to the degree to which it is accessible. As this data grows in size, in type, in dimensionality and in complexity, accessibility becomes of paramount importance. As our Data Lake becomes the data store and computing platform for pan Thomson Reuters as well as third party content, it can fulfill its full potential if and only if it is accessible. Accessibility means a very low barrier of entry: allowing product designers, innovators and non-technical users to explore and navigate the store without requiring them to be familiar with Big Data tools or in-depth understanding of data schemas and information models. Search and navigation support for identifying the data necessary to solve a particular analytical problem is lacking.

The Lake needs a data catalogue, which is being built, that will enable finding the dataset that is needed for a particular purpose. In the mean time, Thomson Reuters' previous master data management technologies can be used to find the names of the relevant data sets that one wants to locate in the Lake.

CURATION: Currently, data is ingested into the Lake and then abstracted into the Knowledge Graph only after a considerable amount of editorial curation occurs upstream of the Lake. It would be preferable if the data entered the Lake in its raw state, and Big Data techniques were used to automate, at least in part, the editorial curation. Such automated tasks are done currently, but outside of the Lake. Such tasks as entity normalization and concordance shouldn't take place out of the Lake; by performing these operations within the Lake itself, we can take advantage of the distributed computing environment in order to make use of common, synchronized data, rather than slightly different slices of the same data used for different purposes in different repositories and updated at different times.

One part of the curation process that would be valuable would be the capability to train document processors to extract information from PDF documents into structured tables. A great deal of the information that Thomson Reuters curates and sells originates in PDF documents. PDF contains instructions for how to render a page, but the order in which the content format is specified is not fixed and need not correspond to reading, or logical, order. The logical structure of PDF documents must be reconstructed in order to extract the data. Big Data tools that applied machine learning to the problem of extracting data from PDF documents at scale are need.

Since most of the content form is in XML, best in class XML manipulation tools, especially around profiling and fact extraction.

PROVENANCE TRACKING: Ultimately, we would like to be able to track the provenance of information all the way back to the first time that we encounter it. This includes all content types including structured databases, documents, video files etc. In the short term, it is critical that we can distinguish authoritative information sources in the Data Lake and Knowledge Graph from derived information and understand the processes that created the derived information. These requirements bear some relation to the requirements on privacy handling that currently attract a lot of attention.

In addition, managing analytic artifacts in the Lake is an underexplored issue right now. Currently, anyone can write new data into the Lake using a Map/Reduce job or similar process. There are no best practices in place for representing the provenance of the resulting artifact within the current Lake infrastructure. Nor is there a way to make the content of the data represented to users since the Big Data infrastructure bypasses the master data management techniques of the previous master entity-based content management infrastructure.

RIGHTS AND PERMISSIONS: An important need is the ability and to express and enforce what can be done to content. Thomson Reuters manages multitudes of varied content sets with different rights and licensing terms. Some content is fully owned by Thomson Reuters, other content sets are licensed for redistribution under specific terms. There may be other future considerations that required a flexible and scalable technology solution to expression and enforcement of rights and permissions such as adding novel types of external contents.

KNOWLEDGE GRAPH: The Lake contains many, many relationships between entities that can be identified and categorized by Thomson Reuters' named entity recognition engines. It is not clear whether all of these relationships should be represented in the Knowledge Graph or only some. Representing temporal information within the Graph is also a challenge. Knowledge Graphs can grow very large, and we have been having trouble identifying technologies that can enable representing, querying and performing analytics on graphs of the size that we can extract from just one subset of content in the Lake.

Not all relationships need to be made explicit in the Knowledge Graph. Some are inferable on the basis of ontologies representing the conceptual structure of the Knowledge Graph. Determining which relationships to represent explicitly and which to infer is yet to be decided, since the whole issue of scalable inference on graphs of the size we are constructing is very much a subject of ongoing research right now.

DATA VISUALIZATION: Current methods for visualizing document sets and the connections between them are lacking. Visualizations require a great deal of effort to set up, such that exploring datasets for those outside of the group that normally curates it is difficult. Such visualizations allowing one to inspect datasets and their relations should change when the content and connections between them changes. Existing exploration tools for massive data sets are still primitive.

6. CONCLUSION

In this paper, we have sketched Thomson Reuters' recent efforts to extend its distributed content storage and retrieval for the vast, highly curated datasets that are its stock in trade to include Big Data analytical capabilities. Thomson Reuters has implemented a Big Data infrastructure that moves computational capabilities at scale alongside its data. Sitting on top of this distributed, scalable data store are a Knowledge Graph and Metadata services that enable the joining of Thomson Reuters data with external data by means of semantic web technologies. While the capabilities enhanced by these developments are a great improvement over the previous environment, there are many capabilities that the Big Data infrastructure lacks that have become apparent even in the short period that this effort has been underway in the patent analytic work that we have begun and other projects. The intent of this paper is to motivate the development of information systems, or components of such systems to achieve the envisioned future state.

7. REFERENCES

[1] Bluhm, M. “Distributed search methods, architectures, systems, and software”. Patent CA2523591 C.

[2] Apache Kafka project. http://kafka.apache.org

[3] “Open Data Institute and Thomson Reuters, 2014, Creating Value with Identifiers in an Open Data World, Retrieved from http://thomsonreuters.com/site/data-identifiers/"

[4] The MIT Big Data Privacy Workshop Summary Report, Retrieved from : http://web.mit.edu/bigdata-priv/#sthash.QcZsYwLu.dpuf

[5] Stonebraker, Michael, et al. "Data Curation at Scale: The Data Tamer System." *CIDR*. 2013.

[6] Open Calais. http://opencalais.org

[7] Derwent World Patents Index. http://thomsonreuters.com/derwent-world-patents-index/

[8] “Content Marketplace”, Peter Marney https://www.youtube.com/watch?v=UximifFsx0w