Recorded Future –
A White Paper on Temporal Analytics

Staffan Truvé, Ph.D.
Chief Scientiest & Co-Founder, Recorded Future
truve@recordedfuture.com

Thy letters have transported me beyond
This ignorant present, and I feel now
The future in the instant. (Macbeth, Act 1 Scene 5)

Introduction

Recorded Future is bringing a new category of analytics tools to market. Unlike traditional search engines which focus on text retrieval and leaves the analysis to the user, we strive to provide tools which assist in identifying and understanding historical developments, and which can also help formulate hypotheses about and give clues to likely future events. We have decided on the term "temporal analytics" to describe the time oriented analysis tasks supported by our systems.

This white paper describes the underlying philosophy and overall system architecture of Recorded Future and its products.

Search vs. Analytics

Although the focus of Recorded Future is on temporal analytics, a comparison with traditional search engines is inevitable – since search is one important aspect of analytics.

The history of search goes back to at least 1945, when Vannevar Bush published his seminal article "As We May Think", where among other things he pointed out that:

> *The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.*

In the decades to follow, a lot of work was done on information management and text retrieval / search. With the emergence of the World Wide Web, both the need and the ability for almost everyone to use a search engine became obvious.

An explosion of search engines followed, with names such as Excite, Lycos, Infoseek, and AltaVista. All these first generation search engines really focused

on traditional text search, using various algorithms but really looking at individual documents in isolation.

Google changed that, with its public debut in 1998. Google's second generation search engine is based on ideas from an experimental search engine called BackRub. At its heart is the PageRank algorithm, and this is the core of Google's success (together with clever advertising based revenue models!). The main idea of the PageRank algorithm is to analyze links between web pages, and to *rank* a page based on the number of links pointing to it, and (recursively) the rank of the pages pointing to it. This use of *explicit link analysis* has proven to be tremendously useful and surprisingly robust (even though Google continuously have to tweak their algorithms to combat attempts to manipulate the ranking algorithm).

Recorded Future is developing a *third generation analytics engine*, which goes beyond explicit link analysis and ads *implicit link analysis*, by looking at the "invisible links" between documents that talk about the same, or related, entities and events. We do this by separating the documents and their content from what they *talk abou*t – the "canonical" entities and events (yes, this model is heavily inspired by Plato and his distinction between the real world and the world of ideas).

Documents contain references to these canonical entities and events, and we use these references to rank canonical entities and events based on the number of references to them, the credibility of the documents (or document sources) containing these references, and several other factors (for example, co-occurrence of different events and entities in the same or in related documents is also used for ranking). This ranking measure – called *momentum* – is our aggregate judgment of how interesting or important an entity or event is at a certain point in time – note that over time, the momentum measure of course changes, reflecting a dynamic world.

In addition to extracting event and entity references, Recorded Future also analyzes the "time and space dimension" of documents – references to when and where an event has taken place, or even when and where it *will* take place – since many document actually refer to events expected to take place in the future. We are also adding more components, e.g. *sentiment analyses*, which determine what attitude an author has towards his/her topic, and how strong that attitude is – the affective state of the author.

The semantic text analyses needed to extract entities, events, time, location, sentiment etc. can be seen as an example of a larger trend towards creating "the semantic web".

The time and space analysis described above is the first way in which Recorded Future can make predictions about the future – by aggregating weighted opinions about the likely timing of future events using algorithmic crowd sourcing. In addition to this, we can use statistical models to predict future happenings based on historical records of chains of events of similar kinds.

The combination of automatic event/entity/time/location extraction, implicit link analysis for novel ranking algorithms, and statistical prediction models forms the basis for Recorded Future's temporal analytics engine. Our mission is not to help our customers find documents, but to enable them to understand what is happening in the world.

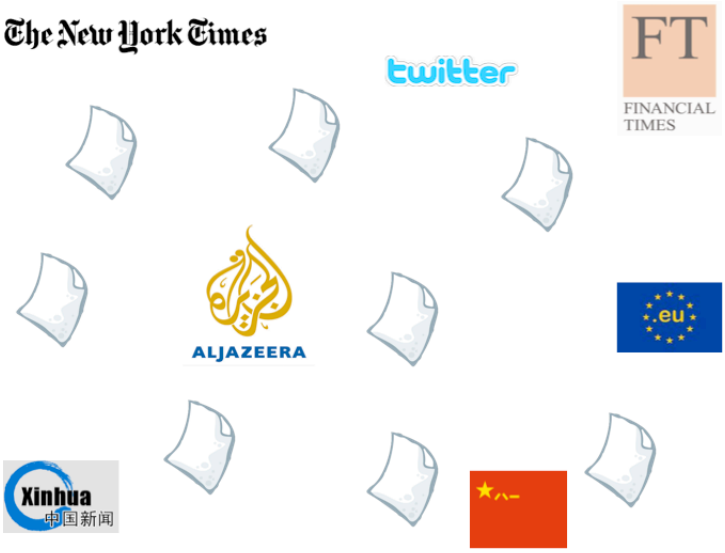
Recorded Future and Business Intelligence

There has been a long path of innovation in systems for business intelligence – trying to help decision makers in companies and organizations make better, data driven, decision. We'd like to think of these in three generations as well:

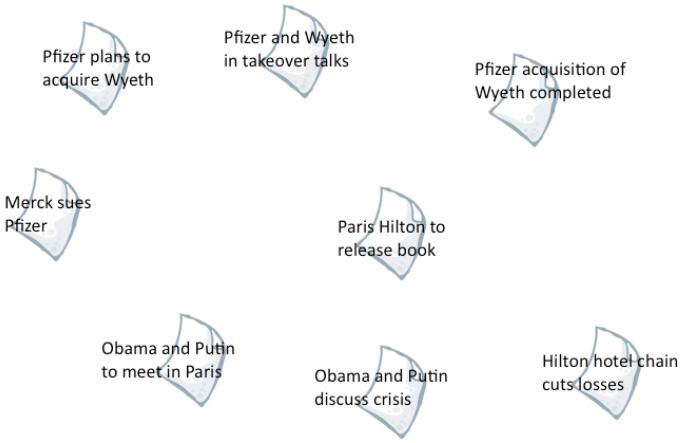
- First generation business intelligence tools (BI) were all about reporting and OLAP cubes, typically taking historical financial, sales, and manufacturing information and organizing for analysis. Very helpful – but very focused on providing a rear mirror view of the world
- Second generation business intelligence was all about real time – hooking into real time data sources as well as real time user interaction – allowing decision makers to both look at very timely data as well as adjust and interact with such views at high pace.
- Third generation business intelligence, we would like to believe, will be all about looking outside corporations and generating data and analytics for decision making based on the **world**, not just old historical enterprise data. This is Recorded Future.

Recorded Future at Work

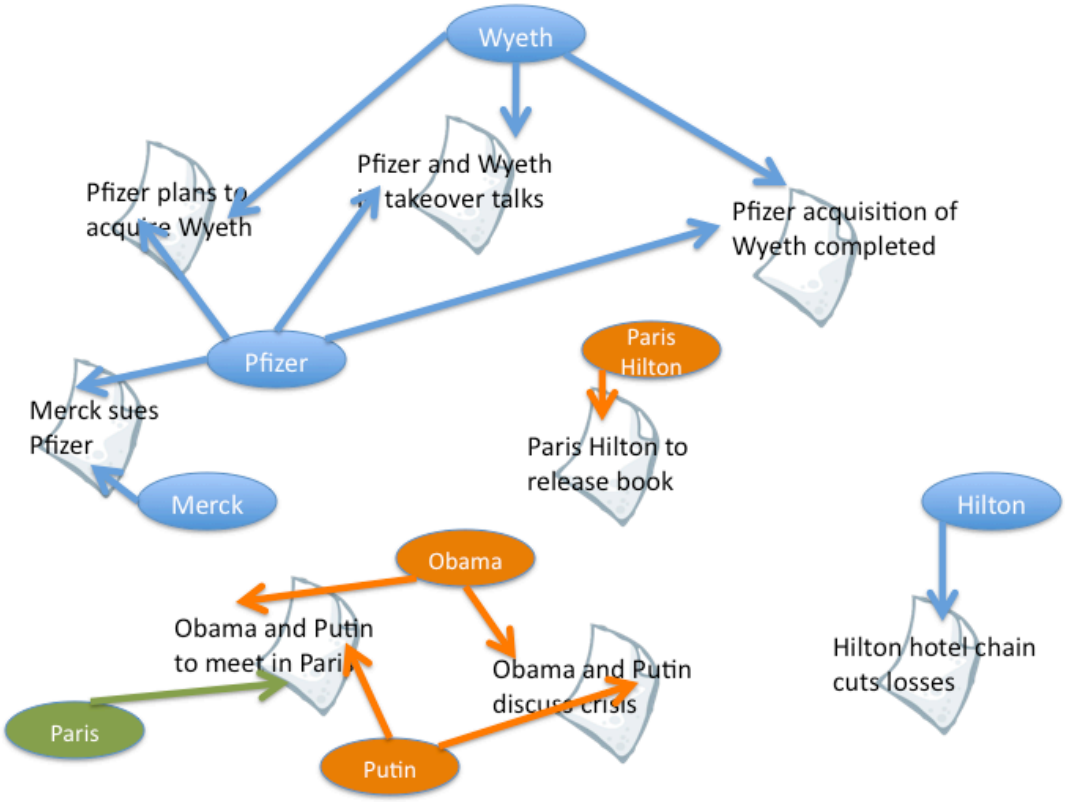
To illustrate these ideas, we'll present a simple example. Assume we have a set of different sources from the net, as illustrated in this picture:



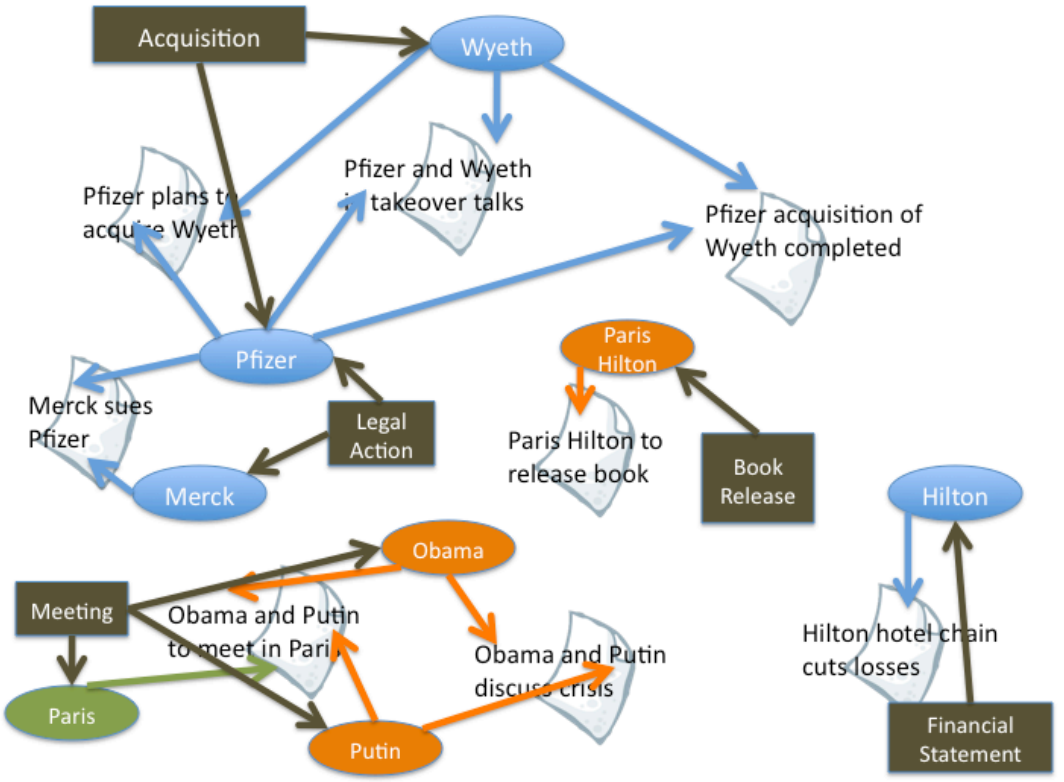
From these sources, we harvest documents, either from RSS feeds or other forms of web harvesting. An example data set might contain the following documents with short text snippets in them:



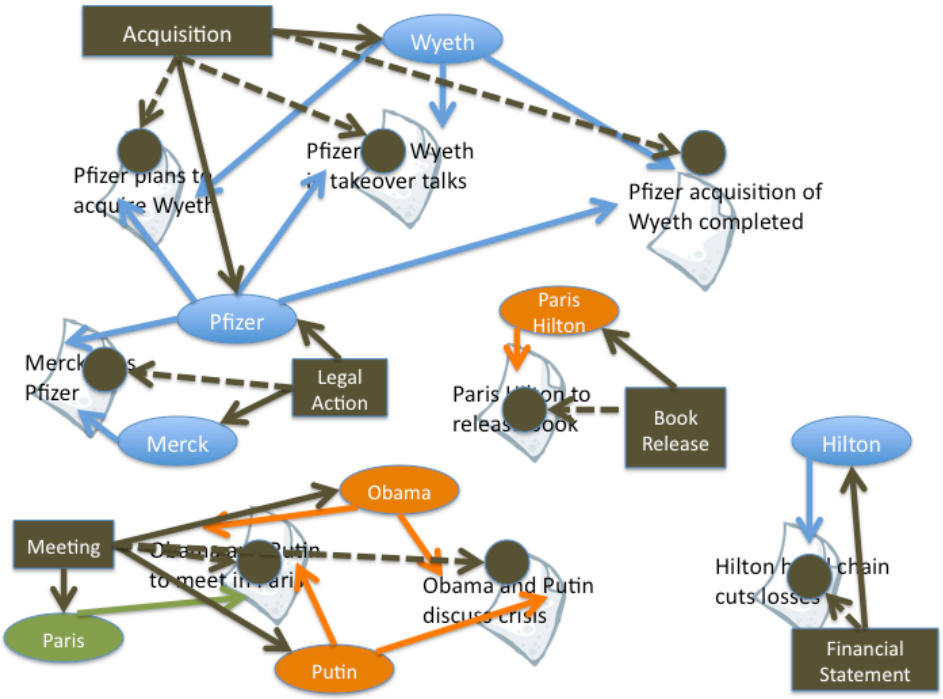
Our analysis first detects entities mentioned in the document, and decides which entity category they belong to (in this example, blue for Companies, Orange for Persons, and green for Cities):



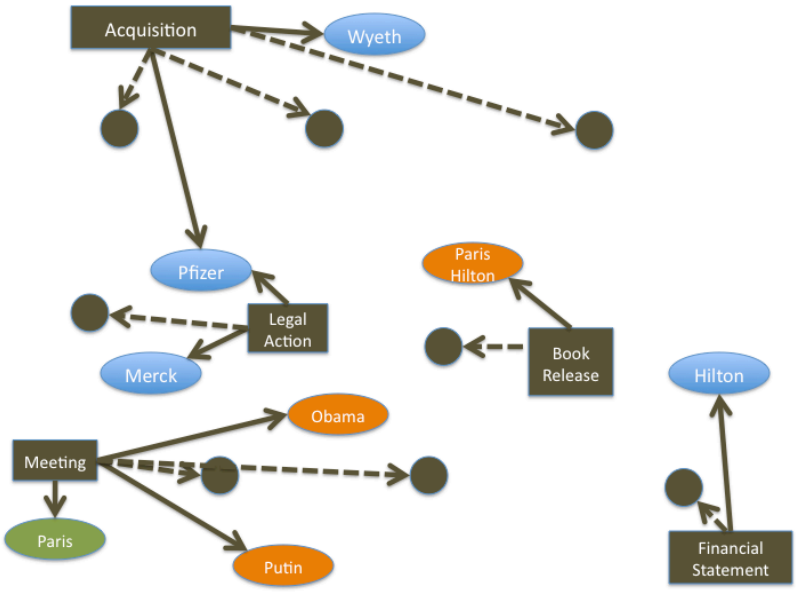
Next, events involving these entities are detected; in this example five different kinds of events:



These are the canonical events; we now add event references / instances derived from the different documents (and the same for entity instances, but for the sake of graphical clarity these are not included in these pictures):

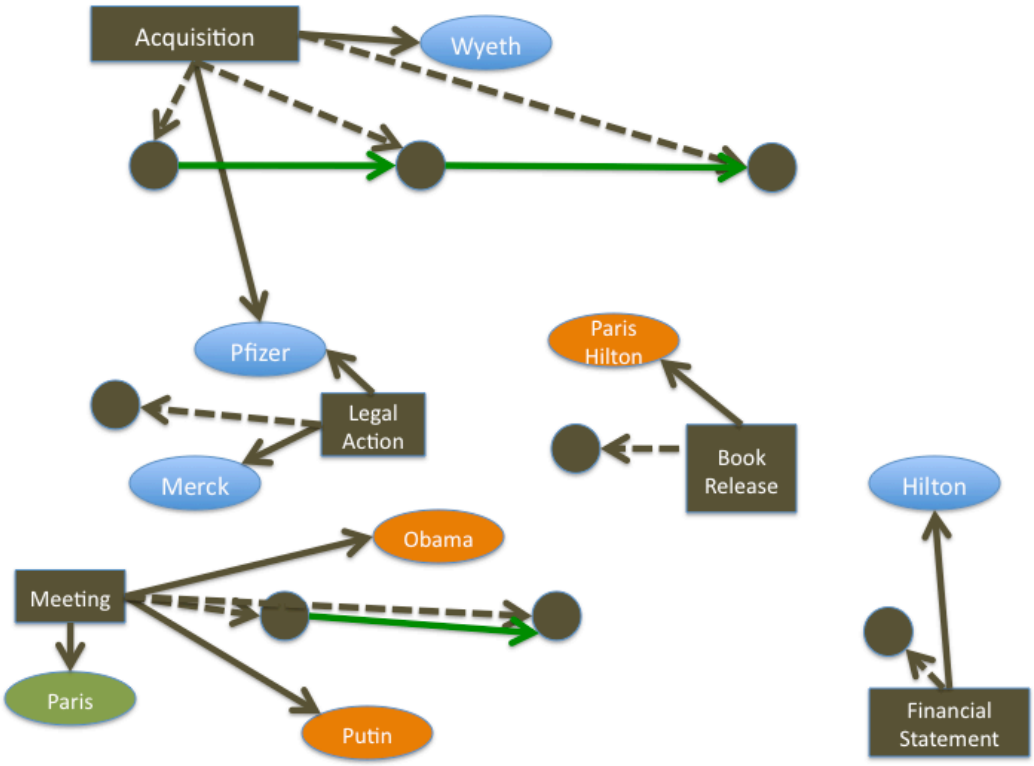


Once this analysis is completed, we can actually dispose[1] of the original texts, since we have completed the transition from the text to the data domain:



[1] We do keep references to the original documents, but we do not store any copy of the actual text.

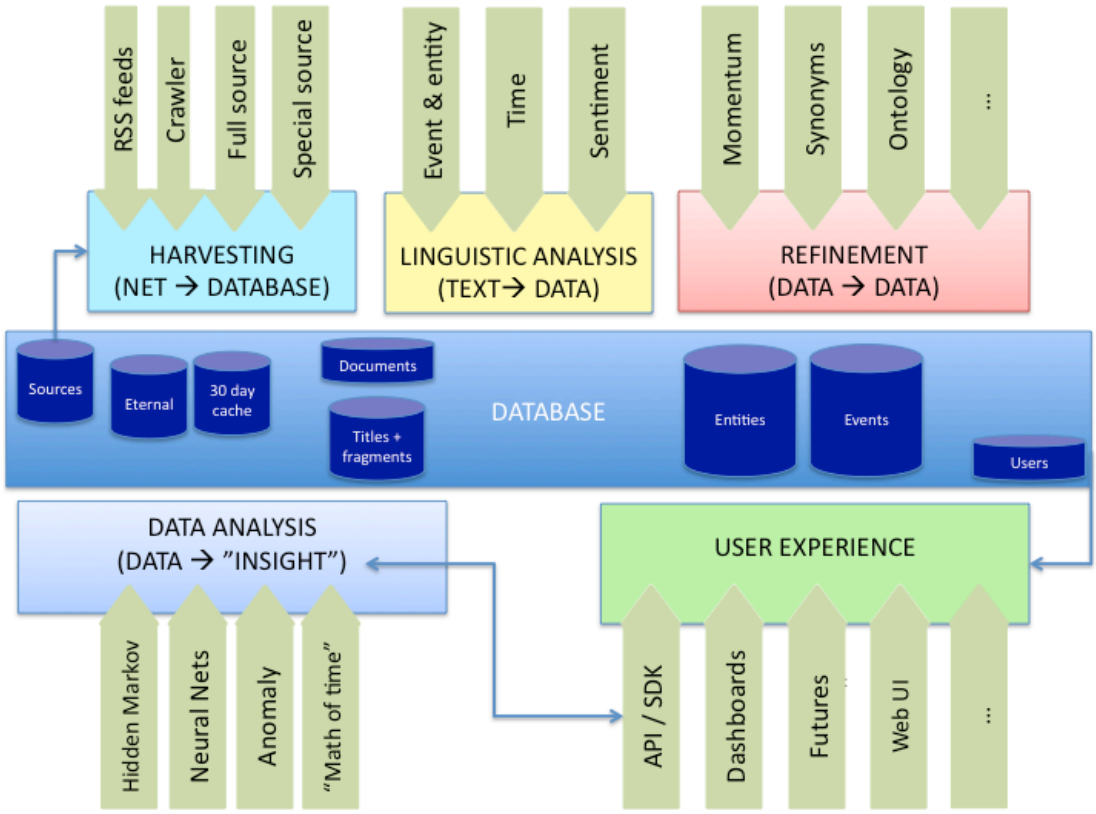
Since we have information about time, we can add more relations to our database, e.g. about which event instances precede others, as indicated by green arrows in this picture:



This completes the transition from the text domain of documents to the "idea world" of canonical events and entities, references to/instances of these, and relationships between these instances. Once this vital step is taken, all kinds of analysis can be used to further enrich the data set, and allow both algorithmic models and human users to explore the data and its implications in various ways.

System Architecture

The Recorded Future system contains many components, which are summarized in the following diagram:



The system is centered round the database, which contains information about all canonical event and entities, together with information about event and entity references (sometimes also called instances), documents containing these references, and the sources from which these documents were obtained.

There are five major blocks of system components working with this database:

- Harvesting – in which text documents are retrieved from various sources on the net and stored in the database (temporarily for analysis, longer term only if permitted by terms of use and IPR legislation).

- Linguistic analysis – in which the retrieved texts are analyzed to detect event and entity instances, time and location, text sentiment etc. This is the step that takes us from the text domain to the data domain. This is also the only language dependent component of the system; as we are adding support for multiple languages new modules are introduced here. We are using industry leading linguistics platforms for some of the underlying analyses, and combine them with our own analysis tools when necessary.

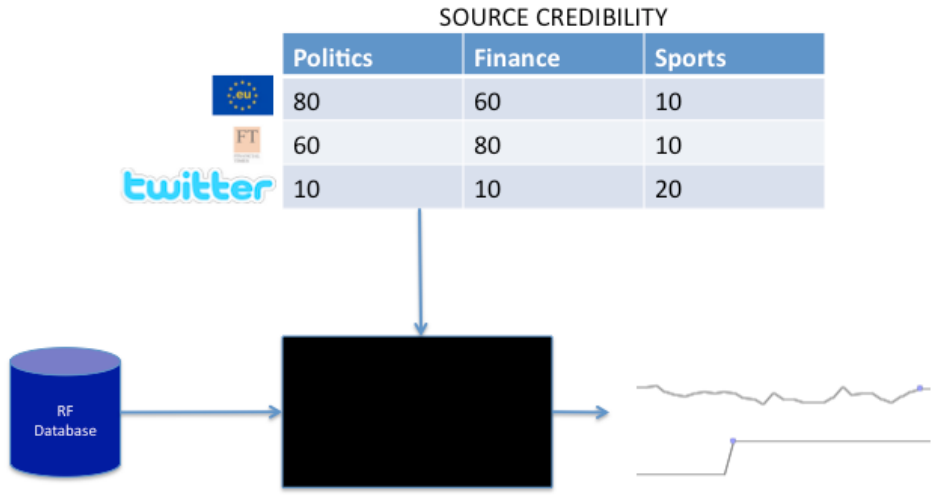
- Refinement – in which data is analyzed to obtain more information; this includes calculating the momentum of entities, events, documents and

even sources (see next section), calculation of sentiment, synonym detection, and ontology analysis.

- Data analysis – in which different statistical and AI based models are applied to the data to detect anomalies in the data and to generate predictions about the future, based either on actual statements in the texts or other models for generalizing trends or hypothesizing from previous examples.
- User experience – the different user interfaces to the system, including the web interface, overview dashboard, alert mechanisms, and the API for interfacing to other systems.

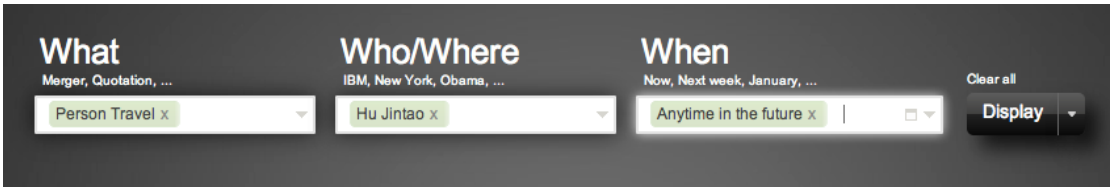
Momentum

To find relevant information in the sea of data produced by our system, we need some relevance measure. To this end, we have developed "momentum" – a relevance measure for events and entities which takes into account the flow of information about an entity/event, the credibility of the sources from which that information is obtained, the co-occurrence with other events and entities, and so on. Momentum is for example used to present results in most relevant order, and can also be utilized to find similarities between different events and entities.

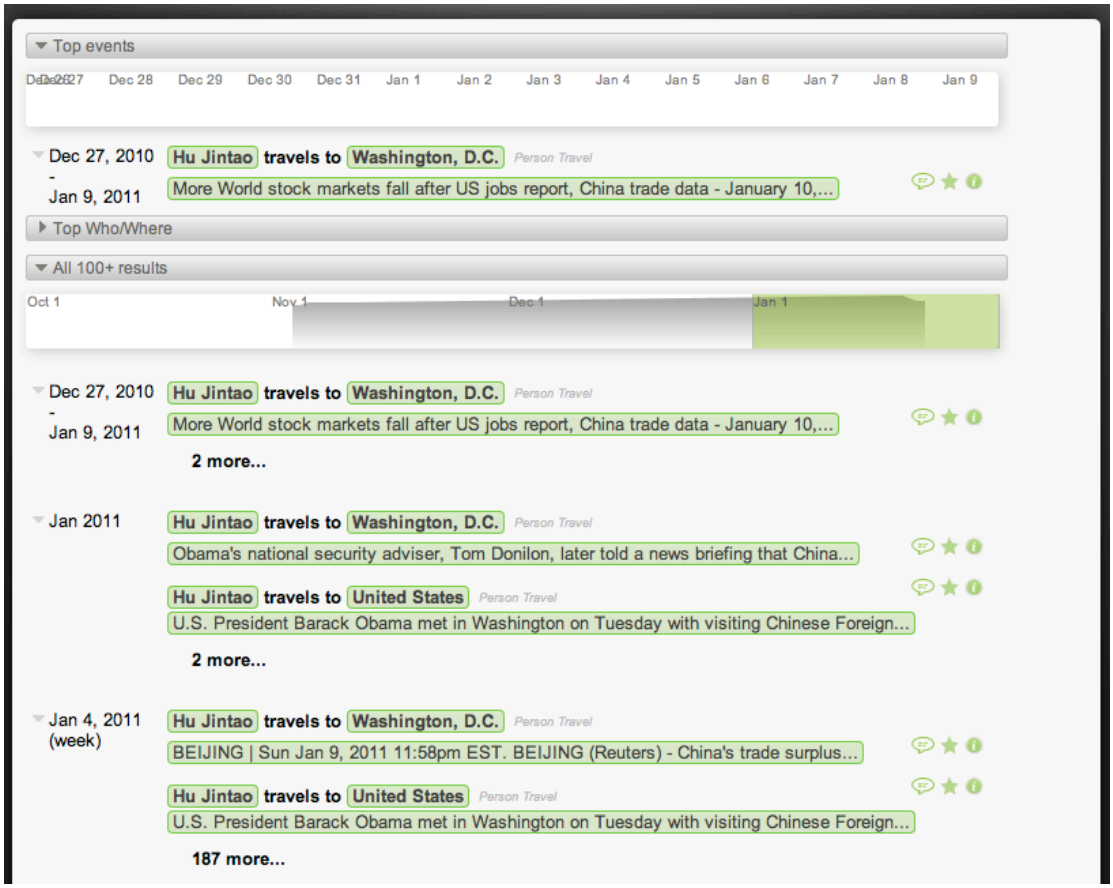


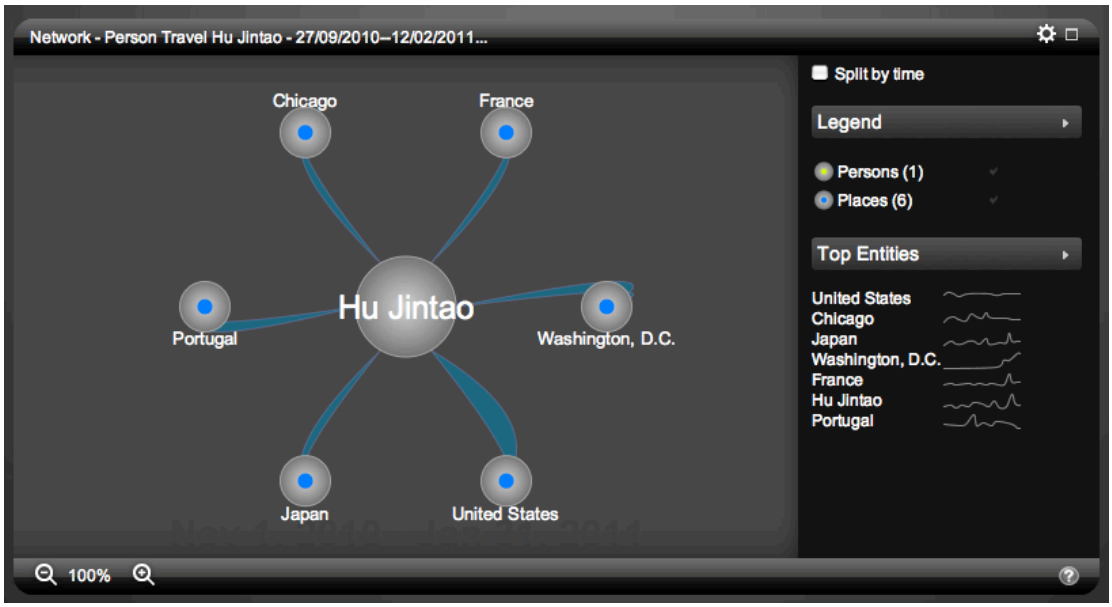
User Experience

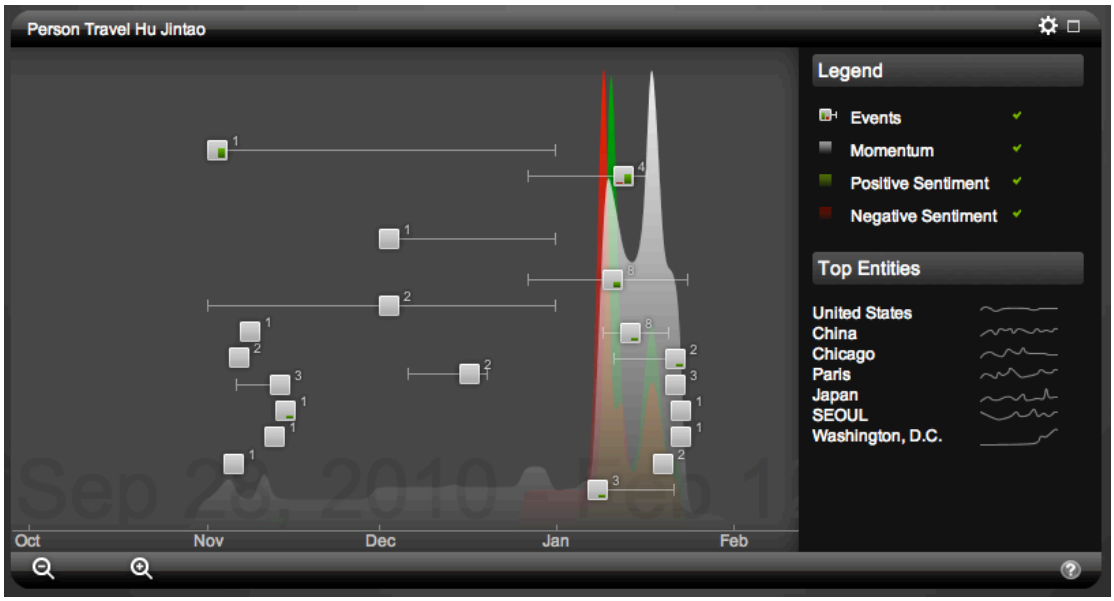
End users interact with Recorded Future through a series of rich user experiences. The analytics query interface allows users to specify events (such as "Person Travel"), entities (such as "Hu Jintao") and time intervals (such as "2009" or "Anytime in the Future"):



The results can then be analyzed in several different views (details, charts, timelines):





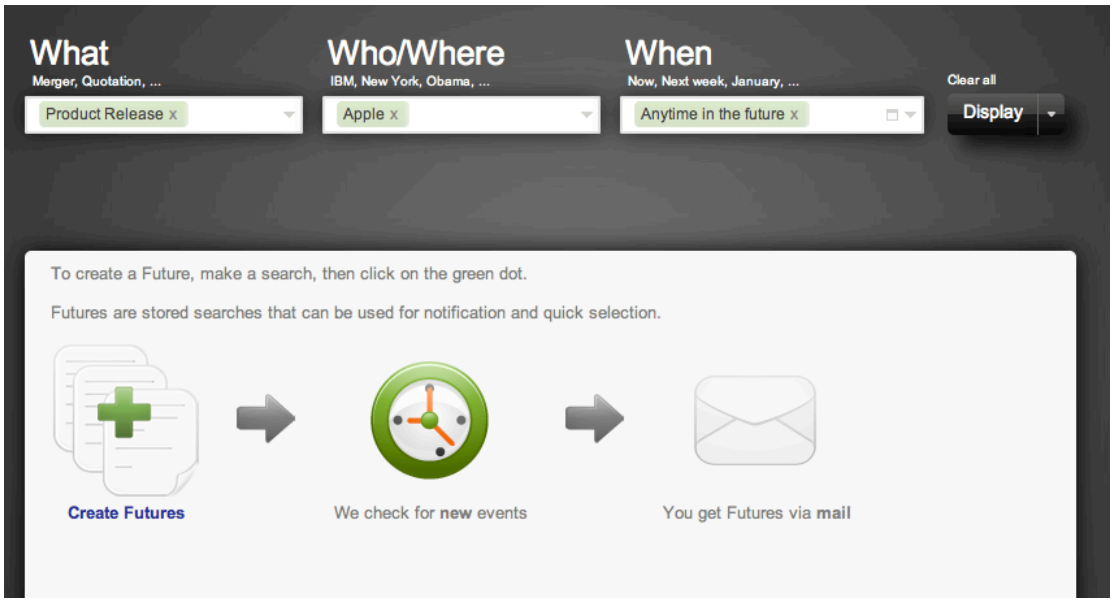


Videos showing the use of the system are available at:
http://www.youtube.com/recordedfuture

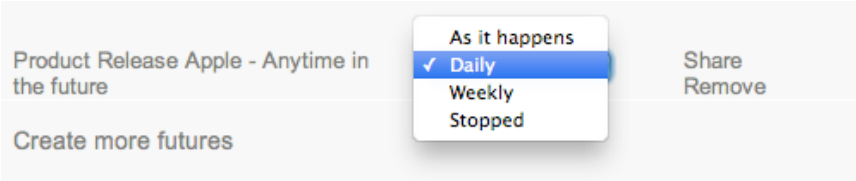
Finally, end users can easily subscribe to email alerts (called Futures) corresponding to interesting queries. Live visualizations with up-to-date data from Recorded Future can also be embedded in blogs, etc.

Futures

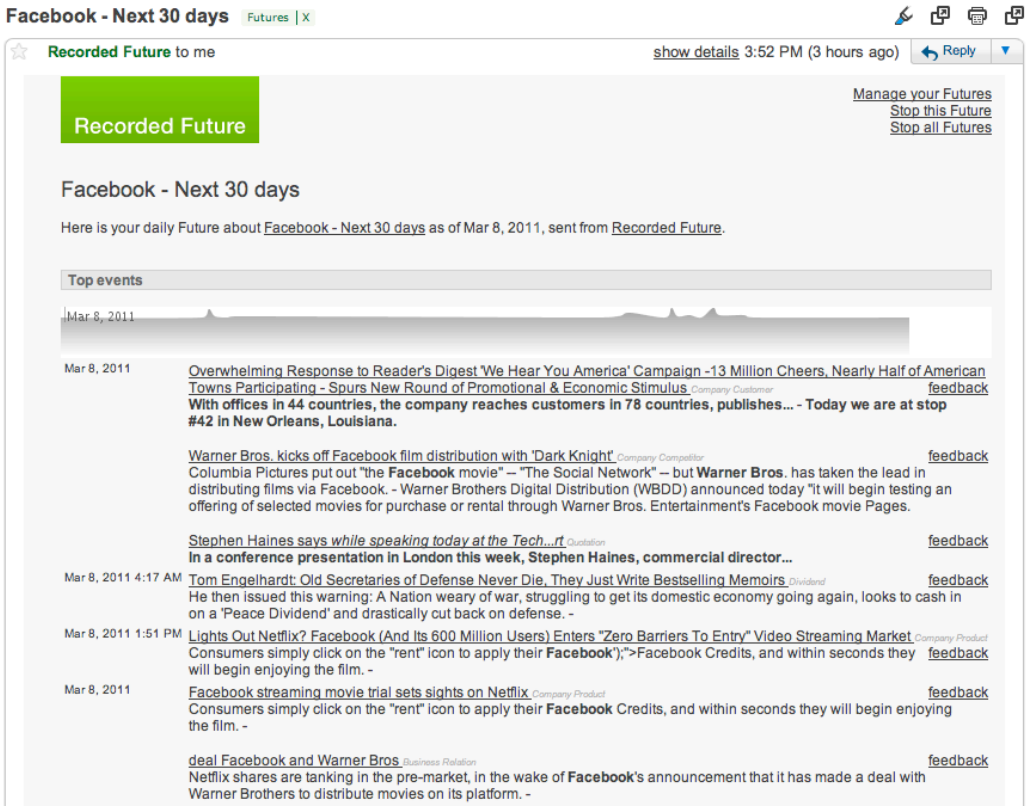
Futures are a way of storing analytic questions and having Recorded Future monitor them with respect to the continuous flow of data from the world. Any query in Recorded Future can be turned into a Future at the click of a green button:



When a Future is defined, the frequency of updates can be specified (and of course changed later), and the Future can also be shared with others:



Futures are then delivered as they are detected by Recorded Future, in a rich email format which works well on both large and small screen devices:



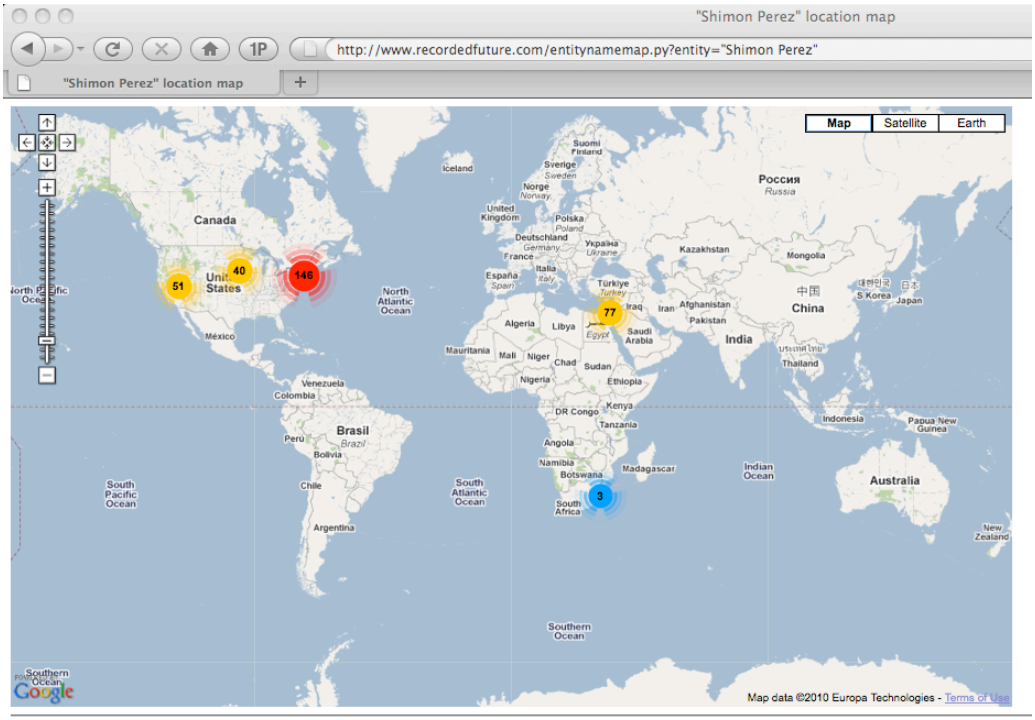
API

Developers can access Recorded Future data and analytics through a web services API (documentation available http://code.google.com/p/recordedfuture). Queries to the system are expressed using json (http://json.org/) and results are provided as json or csv text. The API can be used to interface Recorded Future with statistics software such as R (http://www.r-project.org/) or visualization software such as Spotfire (http://spotfire.tibco.com/), as well as proprietary analytics applications.

Examples of applications of the Recorded Future API include:

- Algorithmic trading – using the Recorded Future data stream to enhance automated trading/risk decision making, e.g. by monitoring momentum and sentiment development of companies in a portfolio.
- Media monitoring – building new applications that monitor social as well as traditional media coverage of a company, industry sector, organization, or country.
- Dashboards – using the Recorded Future data stream to display novel, externally oriented, indicators of the world, like the following very simple example:

- Geographical information accessed through the API can easily be used to present results in 3rd party applications such as Google Maps and Google Earth:



A Final Word

Recorded Future brings a paradigm shift to analytics, by focusing on time as an essential aspect of the analyst's work. Sophisticated linguistic and statistical analyses combined with innovative user interfaces and a powerful API brings new opportunities to both human analysts and developers of 3rd party analytics systems. We continuously develop all these aspects of our system to bring new tools into the analysts' hands - the future has only just begun!

"Thus, what enables the wise sovereign and the good general to strike and conquer, and achieve things beyond the reach of ordinary men, is foreknowledge."

(from The Art of War by Sun Tzu, Section 13)

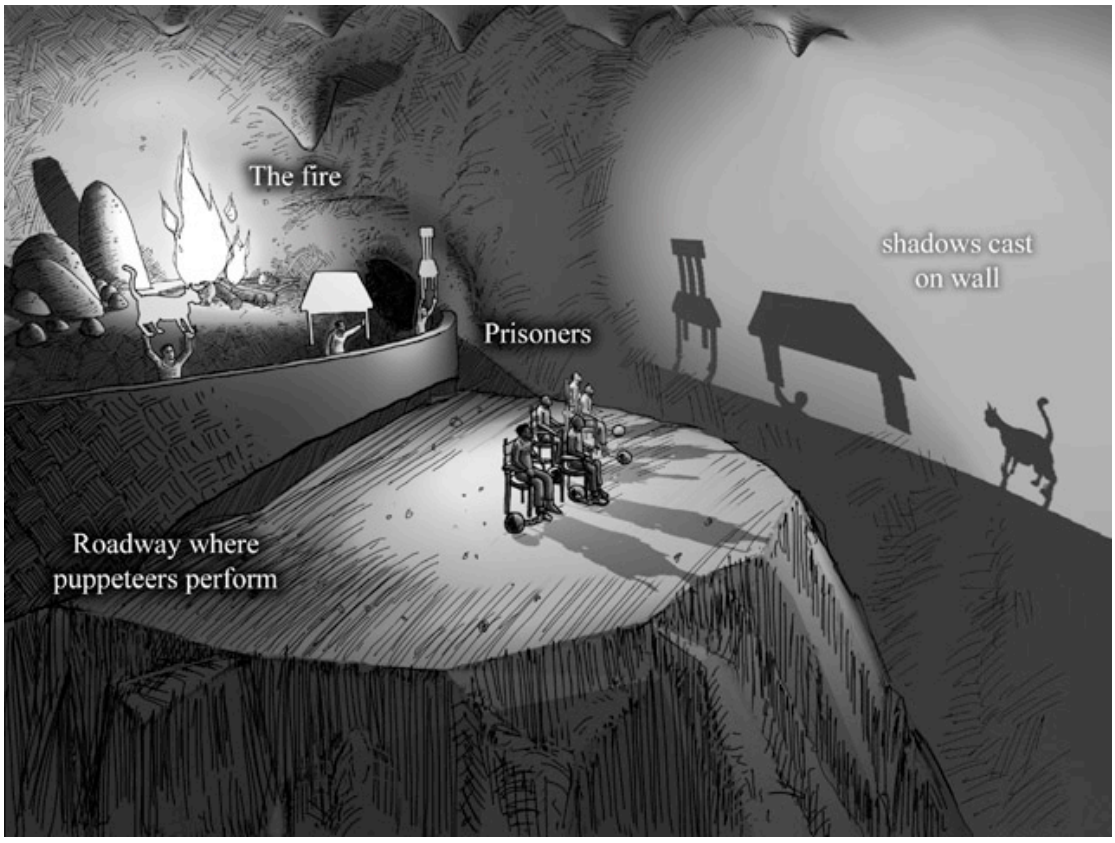
WHITE PAPER ADDENDUM

Plato, the Cave, and Recorded Future

Staffan Truvé, Ph.D.

To understand the philosophy behind Recorded Future, it is helpful to consider the famous "cave allegory" by Plato:

Plato imagines a group of people who have lived chained in a cave all of their lives, facing a blank wall. The people watch shadows projected on the wall by things passing in front of a fire behind them, and begin to ascribe forms to these shadows. According to Plato, the shadows are as close as the prisoners get to seeing reality. He then explains how the philosopher is like a prisoner who is freed from the cave and comes to understand that the shadows on the wall are not constitutive of reality at all, as he can perceive the true form of reality rather than the mere shadows seen by the prisoners. (en.wikipedia.org/wiki/Allegory_of_the_Cave)



(image from www.thatmarcusfamily.org/philosophy/Course_Websites/Phil_Math/Photos/Cave.jpg)

What we read in newspapers, blogs etc. is not unlike the shadows on the wall of the cave – we get reports about events in the real world, and attempt to use that information to get an idea about what is really happening. As good analysts, we naturally consult several sources, and weigh together the information obtained from them – always keeping in mind that some sources are more credible than others, and thus should be given higher weight. We call the evidence we get from

different reports "event instances", and the real world events they report on we refer to as "canonical events".

A canonical event, in our system, is a representation of a particular happening in the real world. For example, assume we read the following statement in the New York Times:

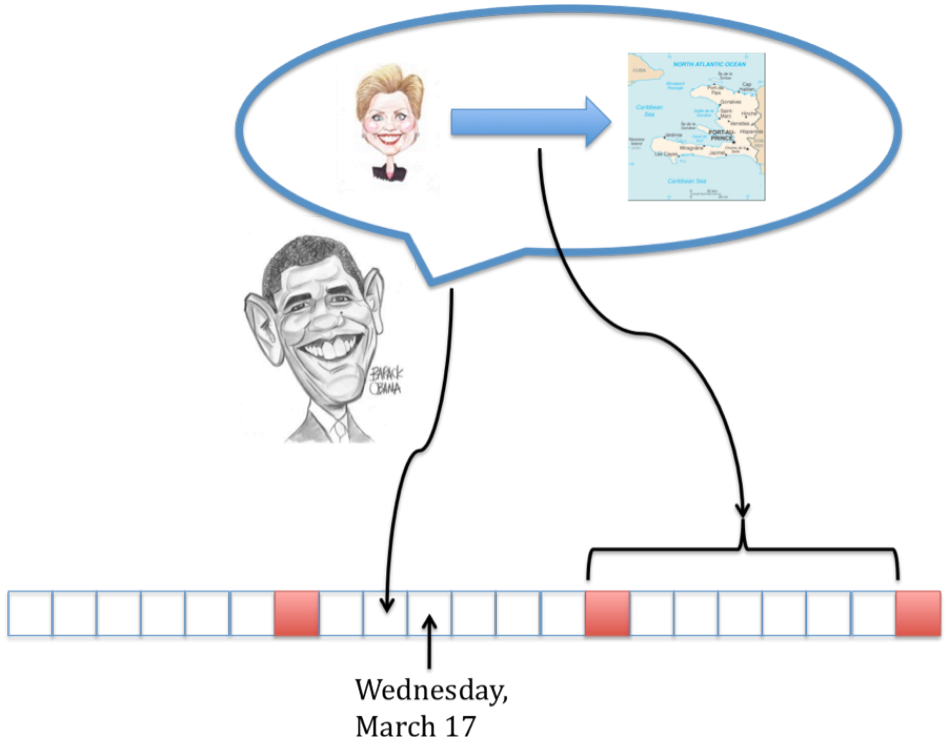
"Barack Obama said yesterday that Hillary Clinton will be travelling to Haiti next week"

This statement describes two events: a canonical "Quotation" event and a canonical "Person Travel" event.

The quotation event refers to a canonical entity, "Barack Obama", and a statement "Hillary Clinton will be travelling to Haiti next week". It has an associated time, "yesterday".

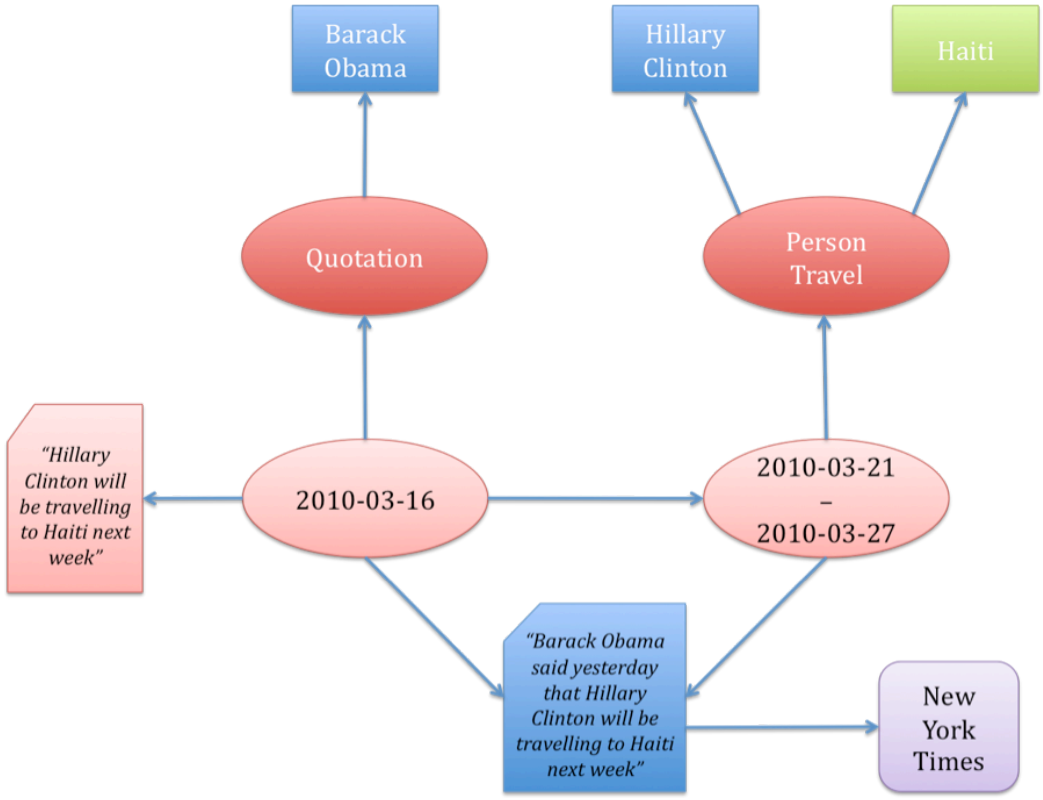
The "Person Travel" event includes references to two canonical entities, "Hillary Clinton" and "Haiti", and has an associated time "next week".

Note that "yesterday" and "next week" are relative times, and to place them on an absolute time axis we need to know when the entire statement was uttered. Let us assume that the statement was uttered on Wednesday, March 17th. Then we might represent the statement pictorially in the following way[2]:



[2] Note that "next week" is culturally dependant – in the US, weeks begin on Sundays whereas in many other countries they begin on Mondays!

In our system, this statement will be represented in the following way:



We have three canonical entities: Barack Obama and Hillary Clinton, which are Person entities [blue rectangles], and Haiti, a Location entity [green rectangle].

There are two canonical events [red ovals] – "Quotation by Barack Obama" and "Person Travel of Hillary Clinton to Haiti".

Furthermore, there are instances of these events [pink ovals], which are tagged by the time or time interval during which they are expected to have occurred or will occur.

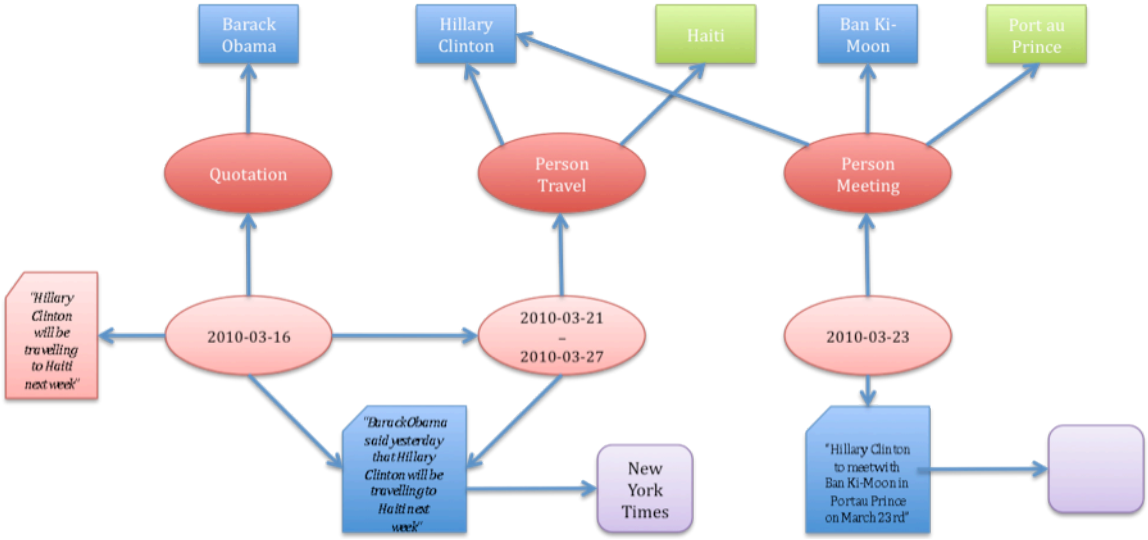
The Quotation instance also has a reference to the text of the quote and to the instance of the event referenced in the quote.

Finally, both instances refer to the text fragment representing the original statement, and the fragment refers to its source – the New York Times.

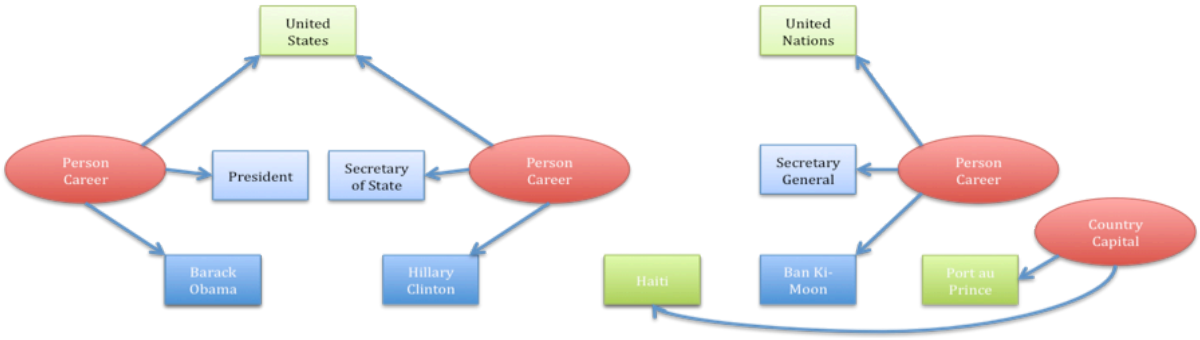
Multiple text documents, retrieved from different sources, can of course be used to gather evidence of the same canonical event, i.e., to provide different instances of the canonical event. Several different canonical events – and instances – will also refer to the same entities. To extend our example, let's add the statement:

"Hillary Clinton to meet with Ban Ki-Moon in Port au Prince on March 23rd"

The representation of our "world knowledge" will then be updated to:

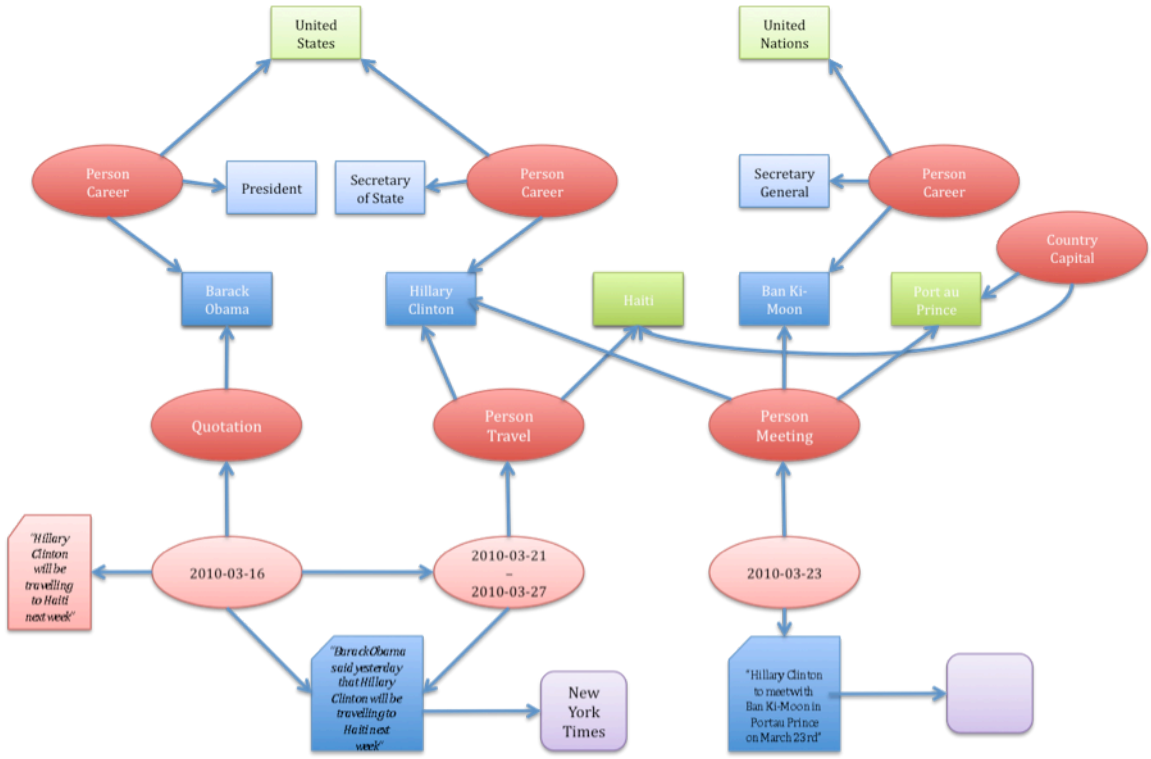


Is this all we know? Not really! Recorded Future also maintains an ontology[3], with additional information about canonical entities and their relationships. In this particular example, the following information can be found in our database:



[3] **Ontology** is the philosophical study of the nature of being, existence or reality in general, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.
(http://en.wikipedia.org/wiki/Ontology)

Combining the information derived from analyzed text and the ontology gives us the following picture for this minimal example. In the real Recorded Future database, there are millions of event instances. This should give you an idea about how the richness of Recorded Future data can help you in analyzing events in the real world!



Additional reading on our blogs:

Company Updates:
http://blog.recordedfuture.com

Government & Intelligence examples:
http://www.AnalysisIntelligence.com

Finance & Statistics examples:
http://www.PredictiveSignals.com