Unearth Gold in Landfills of Digital Data

Dr. Dorian Selz & Cesare Allavena

squirro

How to benefit from the largely untouched 80% of data

In the 2001 report "3D Data Management" [1], Douglas Laney of Gartner defined the three commonly accepted dimensions of Big Data that are Volume, Velocity and Variety. Volume is simply the amount of data available. Companies generate more and more digital information through customer interactions, transactional information, and social media just to name a few. They acknowledge the value of this data and therefore the volume of information available for analysis increase accordingly. Velocity is the ability of an enterprise to have access and utilise data in a fast way and to distribute it at an equally high speed. This not only requires physical bandwidth it requires architectures that balance data latency with data requirements. Finally this paper focuses on Variety or the range of data types and sources.

The variety of sources points to numerous types of data in various formats with different definitions and structures. While these issues have been addressed manifold – e.g. Master Data Management, one issue has received less attention: Putting this data in context.

This white paper looks at the enterprise information space and different data types. We outline strategies to combine data sets, referred to as Context Intelligence [2] to drive visibility and more informed decision-making. Additionally, customer vignettes discuss applications of use case and value generation. The paper concludes with a number of suggested action items to jump-start the analysis of the largely untouched 80% of data.

The Enterprise Information Space

Simplifying to some degree the information space in an enterprise may by split between structured and unstructured data. While the next chapters deal in more detail with the difference between the two, we assume the following split:

- Structured data: Mainly numbers lending themselves to calculation
- Unstructured data: Mainly text lending itself to full-text search

Experts estimate the split between structured and unstructured data of all data available in an enterprise to be roughly 20% / 80% [3]. This includes a wide array of elements amongst which: databases, XML data, enterprise systems (Enterprise Resource Planning (ERP), Business Intelligence (BI), Customer Relationship Management (CRM), spreadsheets, documents, call notes, email, chat, transcripts, etc.

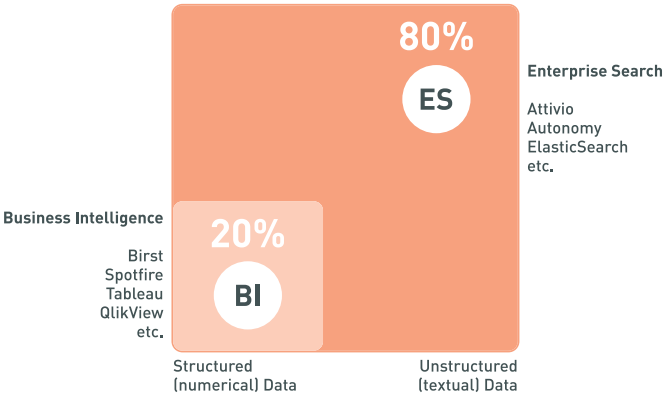

Figure 1 – The Enterprise Information Space

Both areas saw a tremendous growth of solutions allowing you to access and analyse this data. In the structured realm these are mostly Business Intelligence solutions, while in the unstructured space a number of full-text search solutions targeted at the enterprise appeared over the last years. Let's now look at both areas in a bit more detail.

Structured Data

According to PCMag.com structured data is: *"Data that resides in fixed fields within a record or file. Relational databases and spreadsheets are examples of structured data. Although data in XML files are not fixed in location like traditional database records, they are nevertheless structured, because the data are tagged and can be accurately identified."* [4]

Structure in data is critical because it is the attribute that makes it efficiently usable. It makes the information much more accessible, easier to deal with, to search, to compound and analyse; simplified: It lends itself to calculation. This is the main reason why early forays into business intelligence nearly entirely focused on this kind of data.

The comparative ease of use of structured data and the analysis of it has been paramount to the fast evolution of descriptive analytics. Descriptive analytics focuses on the past and analyses data to determine the probable outcome of an event or its likelihood to occur.

Structured data is also relevant for interpretative analytics but with certain limits as it covers mostly numerical data. It can help interpret what is happening for example to your sales

figure or to the current status of your supplies or stock.

How can your business benefit from structured data?

There are numerous enterprise systems that enable you to effectively work with structured data. We focus here on BI tools to demonstrate how structured information can be dealt with, visualized, and how important it is for businesses to employ such tools for the tracking of Key Performance Indicators (KPI) like sales, or expenses or inventory levels.

A BI tool will typically enable you to track over time multiple KPIs, compare them and derive correlations and insights.

As an apparel manufacturer you may look at sales of a specific item such as wool hats over time and the outdoor temperature. This is essentially the descriptive analytics that BI tools offer, which can be much more complex and multivariate but essentially follows that principle.

BI tools also offer interpretative analytics. In fact if you are able to map out various KPI's you can start determining correlations between them. In the above example you could observe that every time the temperature drops below a certain level the sales spike, which while this may seam obvious structured data enables you to confirm your intuition.

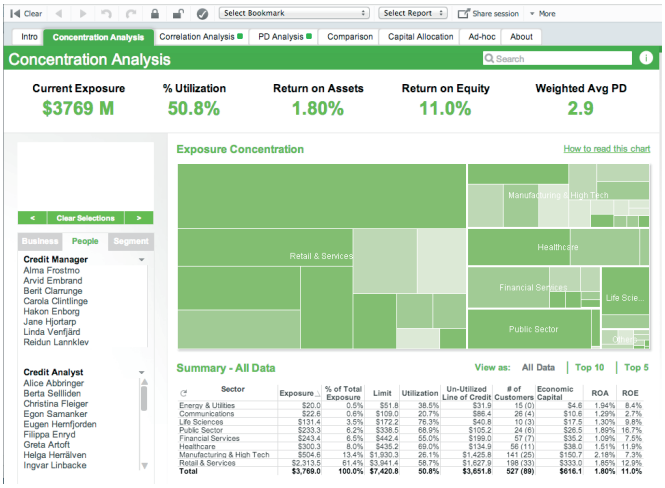

Figure 2 – Example of structured Data in QlikView Risk Management Dashboard

Case Study: Crisis Monitoring and Risk Mitigation

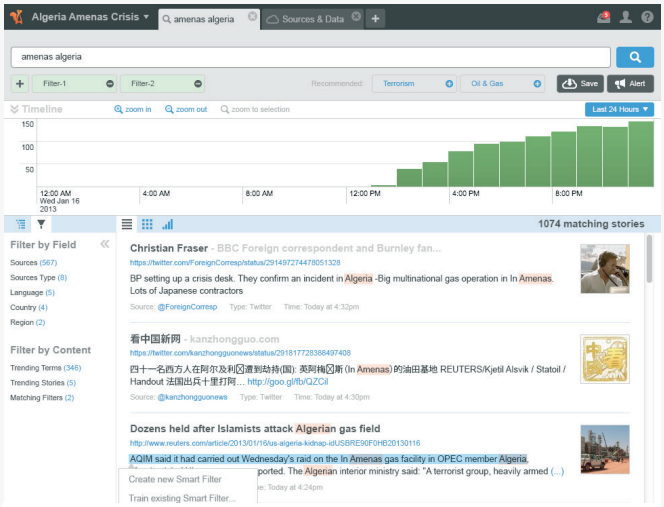

Figure 3 – Solution Interface of Crisis Monitoring Application

Company: A large multi-national corporation providing power and energy solutions[6]. The company is operating worldwide with activities also in politically fraught geographies. The company is unfortunately regularly exposed to critical situations in these areas. The corporate communications team is tasked tracking these situations around the globe. For this, the team scans multiple internal and external channels such as press and social media in multiple countries and languages.

Challenge: Current solutions ranging from media monitoring to social media monitoring fail at detecting crises events. While they keep track of what is been said in general terms about the company they disappoint when used to identify key events as they happen and follow key stakeholders. Instead they return too many so called false positives – results seemingly matching the criteria but actually are just noise.

Solution: In the event of a crisis it is critical for the team to track all involved stakeholders and key players such as local subsidiaries including employees, employee relatives, journalists, bloggers, etc. The core request is for a platform, which allows the team in the event of a crisis to quickly identify and monitor key players and key developments with limited manpower. Squirro is such a platform.

Context generation: In the crisis monitoring application a number of internal structured sources, especially CRM data, are used to create a frame of reference. A number of external data sources ranging from newswires, to social media, general web search, specialized community sites and in some countries also chat data are continuously mapped into this information space. The mapping rests on enriched data streams drawing on a number of techniques from data comprehension and enrichment, entity detection, sentiment analysis, natural language processing, and pattern classification. Figure 3 shows the interface trending information on a developing event.

Results: The variable setup allows the company in the event of a crisis to quickly setup a crisis specific monitoring and leverage this across the organization. The solution allows the company to include all type of external news sources ranging from traditional newswires, to social media and in some countries even the inclusion of chat services. The system in place lets the company include the local subsidiaries and other stakeholders to contribute to this monitoring approach. A federated system setup guarantees local relevance and availability while allowing corporate headquarters to stay in control as the situation evolves. The company realizes substantial risk mitigation by staying on top of developments.

Businesses are more and more concerned with both predictive and prescriptive analytics and this is where structured data and the tools that enable its use start reaching their limits.

In our wool hat example if we wish to start predicting with a certain level of confidence the evolution of sales over time, we need to have the capacity to treat much more data than just the structured, especially when we know that 80% of all data available is of the unstructured kind.

Unstructured Data

Unstructured data refers to information that is not organized in a pre-defined manner or does not have a pre-defined data model according to Wikipedia [5]. Unstructured data is typically text heavy, but it includes also the content of audio files, images, videos, or facts. In this paper we focus on unstructured data of textual format.

Unstructured data is by nature irregular and therefore harder for traditional BI software to deal with. There are nevertheless a variety of techniques that enable a better understanding of unstructured data such as data mining, text analytics, entity extraction and many more. All of those enable users to find patterns in that information and in a way structure the unstructured.

Unstructured data contains much more information than structured data, but because of its complexity and because of the need of complex algorithms to process it, its analysis has not been at centre-stage for businesses until the past decade.

Unstructured data is particularly relevant for descriptive and interpretative reporting and also to a degree for predictive reporting.

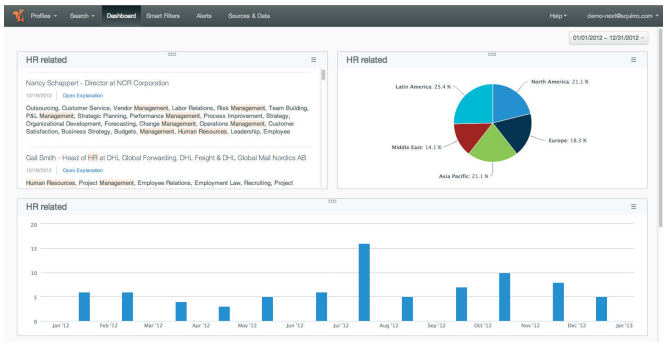

Figure 4 – Example of Squirro's unstructured Data Analytics Dashboard

How can your business benefit from unstructured data?

Unstructured data is found in e-mail communication, social media feeds, news article, reports, csv, pdf, doc, text, etc. Any un-tagged text-based document is essentially unstructured (e-mails for example can also be qualified as multi-structured information because if taken as a whole they contain structured information such as author or time-stamp and unstructured information such as the text body or a picture file attachment) so we can easily understand the volume of information that it represents.

As opposed to structured data that reveals what is happening to specific KPIs, unstructured data can tell users why events occur by extracting meaning and giving context to this data.

Unstructured data analytics tools reveal business insights that would not be available from structured data, therefore providing much more elaborate descriptive and interpretative reporting solutions.

The "why" behind the data enables to better describe it, but most importantly it allows you a much more granular interpretation of its evolution.

Let's get back to our wool hat example. With structured data we understood and observed a correlation between temperature and sales. By including a layer of contextual intelligence we can analyse unstructured data such as fashion blogs and extract meaning from them, map them and determine if the content of those blogs has an impact on sales. You are now not only analysing numbers but also text, which can contain complex information that reveal much more about elaborate events.

In our example with this analysis we could determine that for a specific period not only the cold weather impacted sales, but the fact that such and such colour was in fashion according to blogs, and this affected sales of a specific model.

Unstructured data ads layers of complexity to analytics and modern tools today enable the user to understand and work with these new layers in a very efficient and simple way.

The understanding that users derive from unstructured data analytics tools is so much more precise that it sets a very strong foundation for predictive analytics.

Context Intelligence

Anne Lapkin of Gartner [7] described in a paper in 2010 how enterprises may use context information to improve business processes and enhance performance. In the paper she describes a situation whereby enterprises make use of 'context aware' information to drive insights and business value and continues to state: *"Gartner believes that by 2020 context enriched services will be used throughout the Global 2000 companies to increase productivity and business effectiveness."*

A key to deliver on the promise is to connect structured with unstructured data. Structured data may often be used to serve up background information and information on the underly-

ing business process while the unstructured content delivers the actual context in which such a conversation or transaction takes place *(See use case vignettes)*.

Context Intelligence therefore combines both structured and unstructured data to deliver key insights that impact businesses' top and bottom-line.

What makes a Context Intelligence solution effective for business?

Context and time are the two elements that make a Context Intelligence tool effective.

Context is to be understood as the broad framework in which a certain piece of information or data is being created. Context encompasses both structured and unstructured data. The capability to understand that framework and model is paramount to the understanding of the information of interest. For example when formulating a search query within an enterprise environment for a customer, if you only know the name of the company, the result will lack precision. On the other hand if you can formulate that query with more context, such as the topic of interest, a specific division or product-line, a particular employee or better yet a combination of those elements, the information you'll retrieve will be more accurate and relevant.

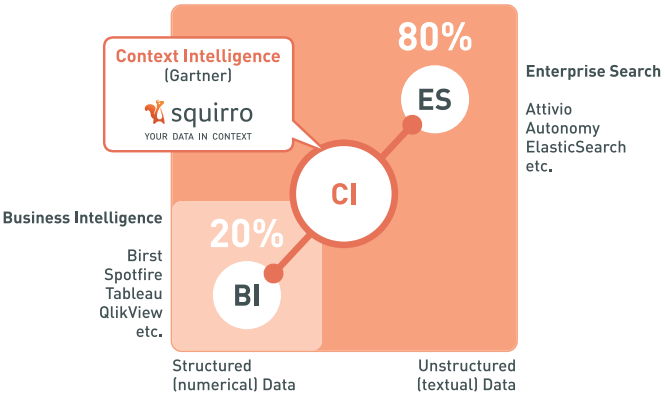

Figure 5 – Context Intelligence: The combination structured and unstructured Data deliver information based on the actual information needs of a user in the context of her work.

Case Study: Productivity Gains through Chat and Email Analysis

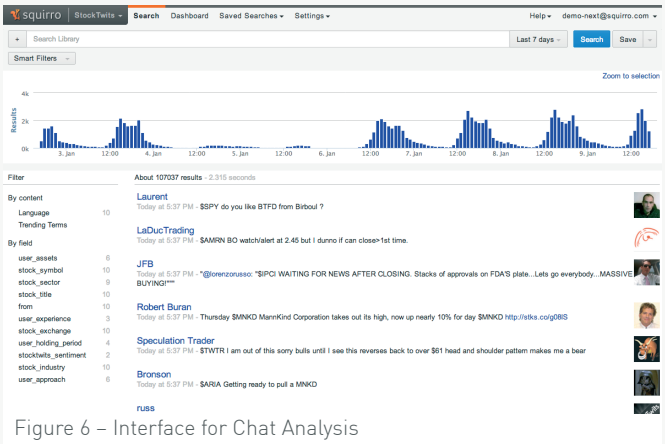

Figure 6 – Interface for Chat Analysis

Company: A global financial firm with large foreign exchange trading (Forex) activities [9]. The bank is reputed to have a best-of-class Forex-trading platform and wants to extend this industry leadership to the monitoring and analysis of the forex trades. Because of regulatory demands, the fast paced nature of the Forex business and legacy systems, an actual trade may be passed between up to 12 different systems (systems for trading, ledger, front-, middle-, back office, etc.).

Challenge: The bank lacks at any one point in time a 360-degree view on the condition of its Forex operations. Last year it started to implement a solution combining the structured data into one single stream for analysis in a BI platform. However exceptions and escalations often are best traced in the chat and emails [10] accompanying a trade. These types of information are typical cases of unstructured information and have so far not been incorporated in any analysis. However, trading exceptions and escalations – e.g. when a trade did not close at the projected price because of market movements – are sources of disputes and of considerable costs. The bank sought to close the loop by incorporating the analysis of chat and email data to get a true 360-degree view of their Forex operations.

Solution: To capture the structured elements of any Forex transaction (trade, log entries, transaction times, etc.) the bank deploys a state of the art bank specific ETL-framework [11] to deliver the necessary data input to a BI solution for analysis. Squirro provides the missing piece – analysis of unstructured data. Loading and analysing chat and email, automatically analysing the data on a number of dimensions such as exception, escalation, and compliance issues, the results are pushed to the BI layer for in depth exploration. The in-built alerting function of Squirro allows triggering events programmatically when certain escalation conditions are met.

Context generation: The analysis of the chat and email data streams is based on state-of-the-art techniques such as density based spatial clustering, topic evolution analysis, and cross-document reference and concept detection. The resulting patterns are normalized and pushed to the BI platform.

Results: By applying context intelligence techniques the bank is able to generate a full 360-degree view of the health of their Forex operations. They are able to catch escalation scenarios substantially earlier saving the bank massively while at the same time increasing client satisfaction. Return on investment calculations show cost savings of → $1 million with ROI figures of 25-30% per annum.

Time is the second building block of a state-of-the-art Context Intelligence tool. The moment when a piece of data is created and the moment when it is consumed or the frequency at which it is consumed reveal essential patterns that can be used for analytics. For example looking at a company's historical financial reports can indicate trends that may influence how the stock price may evolve over time, but the publication of a quarterly financial report will have immediate impact on the share price. The analysis of both is necessary even if they serve a different purpose and yield different insights, this is why the element of time is so important.

As detailed in Squirro's Smart Filter Technology white paper [8], the combination of those elements solves the 'too-much data' problem in a novel way by focusing on the context of your information request. It positively impacts your top and bottom-line and delivers ROI within 4-5 months by allowing users to reduce time invested in re-searching already found information by up to 90%.

Action Points

The combination of structured and unstructured data adds real value to your business. It increases your top-line by enabling users to uncover new business opportunities and positively impacts your bottom-line by uncovering trends that may affect it.

Additionally the integration of both structured and unstructured data in your decision-making process will make it more efficient and reduce risks attached to it.

Enabling these types of insights derived from your data takes time and involves a substantial effort of both the business and technology divisions of your company, but as we have shown it yields great benefits.

In order to achieve a high level of confidence, unified information access (UIA) needs to be at the centre of your data strategy.

Here are action points that will help you build the right environment to generate the return on investment you seek from your data. The focus here is on data types and management and not on architecture.

Identify your needs

Many businesses get into data analytics without clearly understanding their needs and their objectives.

The first step in any data project is to understand which data you have available; what you wish to do with this data (instant reporting, finding key insights in your data, discover operational patterns...) which department or business unit is the most apt and ready to roll-out your plan and finally who are the different stakeholders involved in such a project (CTO, CIO, data scientists, users...).

The appropriate data structure

Bring together all the necessary data sources, both internal and external and establish data governance.

At a minimum your data governance should cover data quality, data management and risk management surrounding data handling.

This will allow you to develop a structure with clean data, with a clear overview of how to work with it and an understanding of potential risks.

Data access

Free your data. Data yields the best value when an organization can generate insights across all former data silos.

Having access to different data layers keeps users engaged, therefore creating a virtuous circle by which they have a better understanding of what they can get from their data and return more valuable intelligence.

Integrate external sources and combine structured and unstructured data

You can derive a tremendous amount of learnings from your internal data, but no matter how developed your structure and your data access are, adding external sources will complete the picture.

There are a great variety of external sources, such as news articles, reports, social media and industry specific data sources.

Incorporating these sources in your strategy will allow you not only to derive insights from your own experience, but also to leverage data sources that span across entire industries.

This level of data combination and integration will allow you to easily move from descriptive to interpretative to predictive and even to prescriptive reporting.

Experiment, measure, learn, repeat

Adapt your data project to your evolution. Businesses are in a constant state of development and your data strategy needs to be as well.

Data growth is exponential, therefore your needs are bound to evolve. Be agile, experiment with new technologies and new sources. Measure their benefits and the business values. Apply to your data strategy your data-driven decision-making process.

Conclusions

We have explained the difference between structured and unstructured data and have shown how they differ. The paper makes a forceful point about combining both structured and unstructured data into a single information space to derive insights. Gartner refers to insights derived from such an analysis as Context Intelligence.

As a business you need to understand that your data (both structured and unstructured) contains information that you cannot afford not to use anymore. There are today number of techniques and solutions available, which allow you to combine data and derive insights. These solutions are available at reasonable cost to your business, yielding high net returns of such projects as the two customer vignettes have shown.

About the Authors

Dr. Dorian Selz is co-founder of Squirro and co-invented Squirro's pattern detection technology.
He holds a PhD in information systems form the University of St.Gallen.

Cesare Allavena is marketing director at Squirro. He previously worked for Mediacom, OMD and the Economist.
He holds a communication degree from UCLA.

1 Gartner: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
2 Anne Lapkin; Context Enhanced Performance: What, Why and How?; 28 September 2010, https://www.gartner.com/doc/1441514
3 InformationWeek: http://www.informationweek.com/software/information-management/structure-models-and-meaning/d/d-id/1030187?
4 PCMAG.COM: http://www.pcmag.com/encyclopedia/term/52162/structured-data
5 Wikipedia http://en.wikipedia.org/wiki/Unstructured
6 Client confidentiality doesn't permit disclosing the name of the company.
7 Anne Lapkin; Context Enhanced Performance: What, Why and How?; 28 September 2010, https://www.gartner.com/doc/1441514
8 Squirro's Smart Filter Technology, http://info.squirro.com/digitalfingerprint
9 Client confidentiality doesn't permit disclosing the name of the company.
10 Intra-company and in-company communication
11 ETL: Extract-Transform-Load

About Squirro

Squirro is the leader in Context Intelligence, combining structured and unstructured data to provide the 'Why' behind the data. Squirro brings the relevant context from the sea of information directly to your regular workplace.

What Business Intelligence systems did for numbers, Squirro does for content: make unstructured data usable. ‚So What?' - because achieving this reduces searching time by 90% and allows for better, more effective decision-making.

The highly skilled Swiss team of search experts has been working together for over 10 years to create a precise software engineering solution that delivers a real-time, self-learning embedded 360° context radar.

Squirro: Your data in Context.

Contact

Squirro by Nektoon AG
Badenerstrasse 120
8004 Zurich
Switzerland

Tel CH: +41 44 586 98 98
Tel US: +1 650 353 93 68
Email: info@squirro.com
Web: www.squirro.com