# TURNING WEB X.0 DATA INTO COMPETITIVE ADVANTAGE

Tyrone Grandison, Daniel Gruhl
IBM Almaden Research Center
650 Harry Road, San Jose,
California 95120, USA.
{tyroneg, dgruhl}@us.ibm.com

**ABSTRACT**
Over the last decade, companies have been slowly realizing that the World Wide Web represents both a pivotal new source of information on their customers and game-changing technology that will augment their current business operations. In response to this recognition, companies are showing interest in technology that tells them what is happening online and leverages it to improve their deliverable or value. This paper presents a project at IBM Research, commissioned by the British Broadcasting Corporation (BBC), to leverage deep Web data to help them stay in touch with their client base. This project also demonstrates the emergence of Web X.0 data as a critical (freely available) resource that can be captured by anyone, harnessed for any particular application domain and then packaged and sold to corporations and governments.

**KEY WORDS**
Web X.0, Deep Web, Text Analytics

## 1. Introduction

The World Wide Web (WWW) is fast becoming a rich source for information on people's tastes, dispositions, interactions. The last two generations of humanity have made the Internet an integral part of their everyday lives; from blogging to online shopping to product reviews to social networks.

As this information is in the public domain, it represents a rich source of context on people, their interactions and their habits. The opportunity to leverage this is very tempting for most corporations. For others, this paradigm shift represents an indication of where their business should be going.

## 2. The Project

In late 2007, the Switch division of the British Broadcasting Corporation commissioned the Intelligent Information Systems Group at IBM Almaden Research Center to help them create a new platform for creating music charts, based on online buzz.

### 2.1 Objective and Target Audience

BBC Switch's mission is to recapture the lost teen audience and engage with it across all platforms – Web, TV, Radio and mobile. Music is viewed as one of the keys to this market. Unfortunately, the music charts do not necessarily reflect current teens' interests. BBC Switch wanted to produce an engaging, interactive and immediate alternative to the traditional sales-based, weekly chart.

The idea of the Sound Index [1] was to capture the buzz around music from the major social networking and music sites on the Web and find the artists and tracks that were the most popular online, right now. Which artists and tracks were people talking about, writing about, downloading and listening to? Is the buzz on an artist or track positive or negative? What demographic segment was feeling artist X or track Y?

Aware of IBM's research in this area, the BBC asked IBM to deliver the pilot. IBM in turn asked its business partner, NovaRising, to design and develop the web site to render the data that was provided by IBM's semantic data extraction and analysis tools. IBM provided overall programme management and systems integration to apply the tools to obtain the data required from MySpace, Bebo, Yahoo, MusicBrainz, Google Groups, iTunes and LastFM.

### 2.2 Rationale and Innovation

The Sound Index employed both a bottom-up and a top-down approach. The team simultaneously examined and designed the visualisation requirements for the teen market and also studied the data provider characteristics. We then built both the user interface and data acquisition and processing components in parallel.

The technology powering the data-driven components of the Sound Index is IBM's MONGOOSE (MONitoring Global Online Opinions via Semantic Extraction) technology [2], which employed a variety of content ingestion techniques to continuously gather user comments, listens, views and other interest indicators across the data sources. This content is then processed through an analytics chain, which is comprised of advanced linguistic and language processing technology.

Comments are transliterated, de-spammed, and analysed for relevance, and listens and views are aggregated at the band and individual track level. Finally, ordering algorithms are used to generate a ranked list of bands and tracks.

The core innovations in the data extraction and synthesis technology were:

1. In analysing comments in the music domain. The team had to create novel techniques to parse Broken English. Contemporary text processing implicitly assumes some sort of well-formed, structured natural language, which is increasingly not the case online.

2. During design and implementation, it was necessary to examine and integrate multiple aspects of a phenomenon, such as what music people listen to, comment on and download. In order to integrate these pieces of information (of different modalities), we had to invent unprecedented and pioneering solutions.

3. Given that we were collecting information from many sources, with different populations, it was necessary to create a ranking algorithm that takes ordered lists as inputs and produces a composite list that strikes a balance between entities that have narrow support and broad appeal in the data sources. This ensured that the resultant lists did not have any particular source dominating, that each source's contribution to the charts was equally valued and that the system was resistant to people trying to manipulate it, through, for example, gaming attacks.

This aggregated data was then ingested, formulated and displayed in a suitable way to appeal to a teen audience. The site allowed them to generate their own specific charts by music genre & source, and geographic location. This required two innovations:

1. A Business Intelligence (BI) engine was required to model the data. Existing technologies didn't seem to work effectively in a multi-user web environment, so NovaRising developed an Online ROLAP (Relational Online Analytic Processing) BI engine in which some of the core suppositions of Business Intelligence were challenged and redefined in order to model large amounts of data.

2. The site's self-generation technology was another innovation, effectively re-building the site every 6 hours based upon the new data ingested from the MONGOOSE platform. Videos, artist history, discography and track information from a variety of ingestion mechanisms from across the Net are mashed together, providing a site which effectively self-maintains.

The Sound Index is the first instance of a consumer-focused, Business Intelligence (BI) system delivered to the masses interested in music.

### 2.3 The Hardware Setup

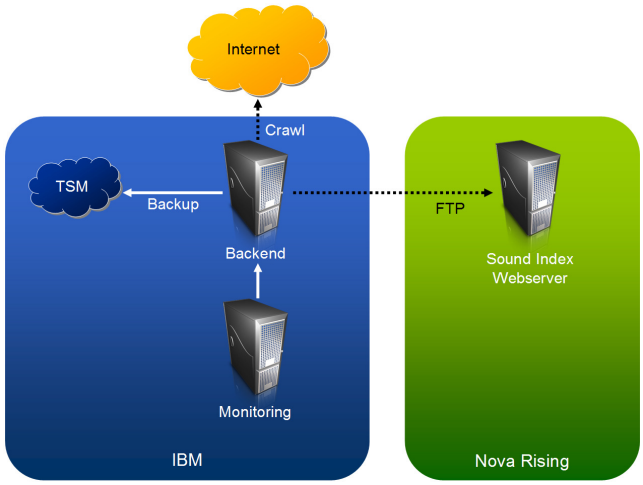


**Figure 1: Server Setup**

Figure 1 shows the hardware setup of the Sound Index system. The backend server crawls the various music web sites and processes the data. An updated listing of artists and tracks is produced ever six hours and delivered to the NovaRising web server via FTP. The IBM system also contains a second machine that uses Nagios [3] to monitor the backend server. Lastly, the backend server is being backed-up using Tivoli Storage Manager (TSM) [4].

### 2.4 The Software Setup

Figure 2 shows the software setup of the system. **Crawl & Process** retrieves data from the Sound Index partner web sites. We distinguish between two data types: *structured data* is pre-processed data, often based on logs kept by our partners, such as the top 400 songs listened to on LastFM.

*Unstructured data* is usually natural text, such as user comments from MySpace. Unstructured data requires an additional processing step called *spotting* that determines which artist or track is being talked about.

All data is formatted by post-crawl processing for import into the DB2 database with the following columns: Source, Artist, Track, Date, Count. The output files are placed into a standard directory, where they are picked up every half hour by the database importer.

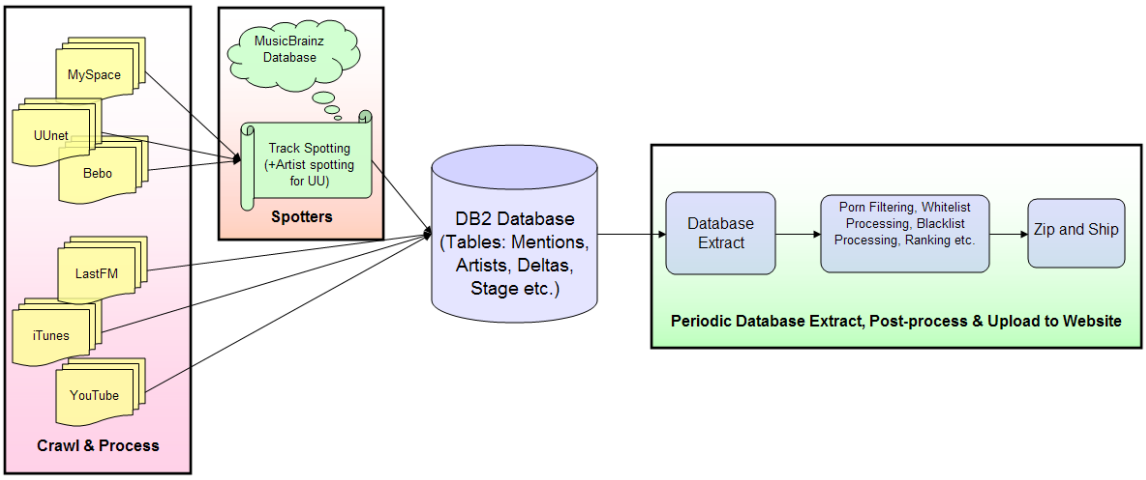


**Figure 2: The Software Setup for the Sound Index System**

Every six hours, a new Sound Index chart is produced by extracting all relevant information from the database. The profanity filter removes offensive content at this point, applying blacklist and whitelist changes in a timely manner. The final dataset is packaged using tar/gzip and made available to Nova Rising.

### 2.5 The Issues That Arose

As with all projects that break ground in a particular field, there were significant hurdles that we had to overcome.

Permission and technical help from partners

The legal position regarding copyright and intellectual property of information gathered from web sites is far from clear and the partners in this project decided to take a cautious approach, by seeking permission from all "donor" sites rather than simply collecting data without consent as a search engine would do. This inevitably delayed the project as a positive response was required from each of the partners.
Another aspect of this was that with an established relationship, it was possible to get advanced warnings of design and format changes and agreement on how hard we could hit the sites' servers. In some cases we were also able to obtain a feed of raw data.

Noise versus Freshness

One of the foremost research challenges is that of noise effects versus freshness. There is a tension between the desire for rapid and frequent updates reflecting the very cutting edge of what is hot, and minimising the influence of noise in the charts due to short term spikes. Striking a balance here poses an interesting challenge. Effects such as weekends, nights and holidays need to be weighed against events such as new album releases, celebrity gossip events and award shows. Any such system will ultimately be a compromise between being too sensitive and not reactive enough and optimising this balance is a difficult research challenge. To solve this issue, we have used a 24 hour window (that is 4 6-hour cycle periods) to smooth out some effects. Other approaches such as long (multi-month) decays have also been explored. Ultimately, it is important to have a ranking scheme that is at least somewhat resistant to "noise", while still capturing freshness. For example, one that looks for a rise in interest in diverse sources and ignores sudden spikes in a single source.

Spam and Off-topic Detection

The tremendous popularity enjoyed by websites such as MySpace and YouTube also attracts undesirable attention. Spammers and other commercial sites regularly attempt to peddle their content via these sites, by masquerading as "bands" or "users". This poses a challenge that has two distinct flavours. The first one is the ability to distinguish valid artists from those that are nearly product spam. A subject matter topical dictionary enables a first pass at this, as does a list of fairly common "spam" phrases, but the ultimate editorial adjudication at this point is subject matter expert driven. The second is the ability to filter spam and profanities.

Business Intelligence

The front-end analysis engine had to be re-built several times utilising different BI environments before it was realised that existing BI technologies couldn't cope with our requirements for a mass-user online audience. We went back to the core theory behind BI and in there lay the reasons it wouldn't work in what is effectively still a primitive online environment. We hence built our own engine which refined how searches across the data cube are undertaken and pays more credence to optimising these for web technologies.

### 2.6 Outcome of the Work

As a pilot, the Sound Index (Figure 3) has been a stunning success. Without any marketing or promotion the Sound Index went from a standing start in April to deliver 43,469 visits from 37,900 unique users in June. In that month, there were 140,383 page views at an average of 3.67 per user. Each user spent an average time of 3 minutes 40 seconds on the site (53 seconds per page). In August 2008, there were over 772,000 web page references to the Sound Index.

We found there were fewer negative comments on the various partner sites than expected, and that relatively few people actually commented at all, though the overall data volumes were so large that there was still more than enough data to create a relevant index. There was big variation between the different partner sites in terms of both which artists were popular and when they were popular. The aggregation of multiple sources was therefore vital to the production of a robust index.

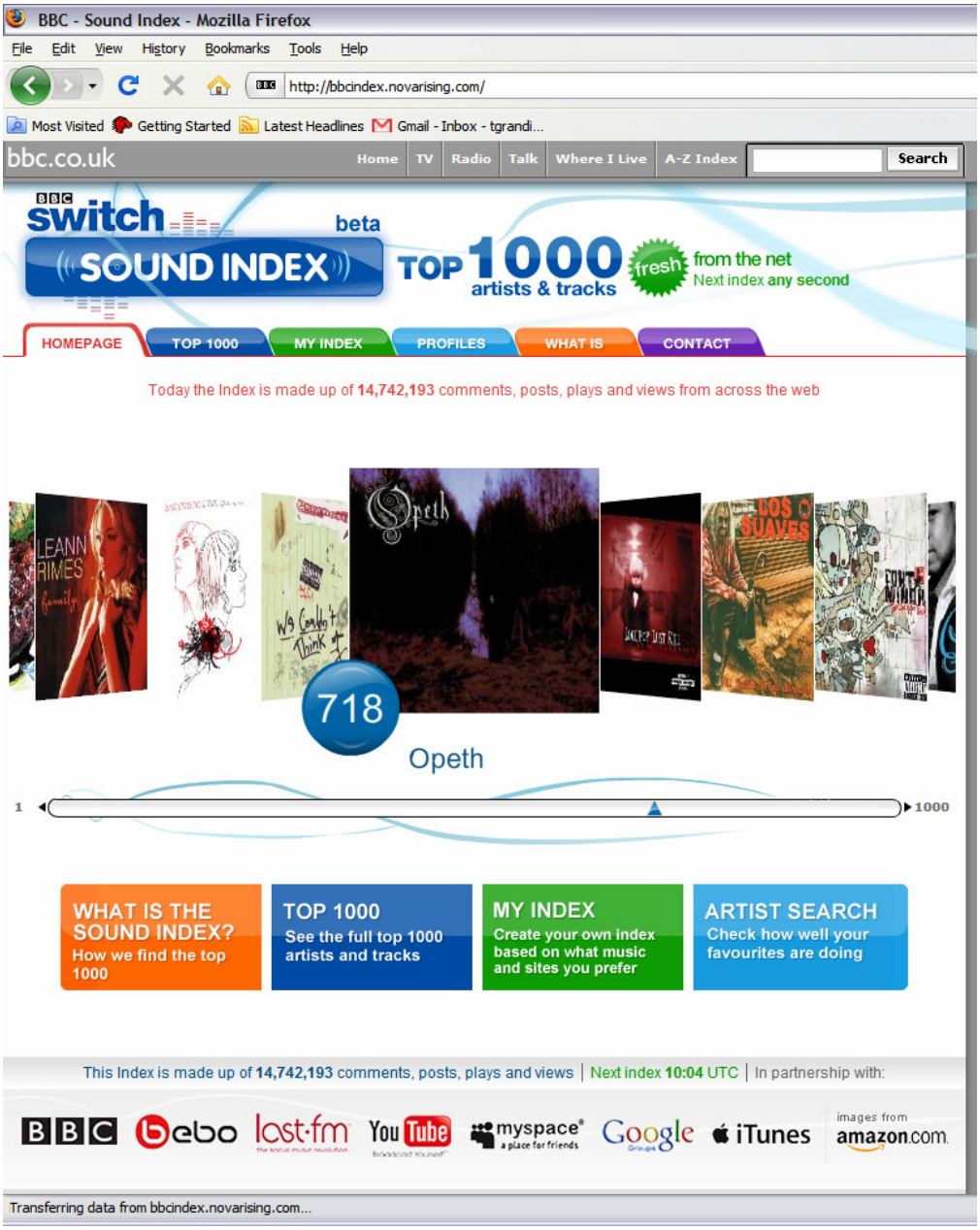


**Figure 3: Screenshot of the Sound Index (taken Dec 31, 2008)**

There was a lot of positive comment from the web and from the traditional press. The Sound Index generated a lot of debate about what constitutes popularity and how the results should be viewed. This was another of the objectives of the pilot.

The pilot has now closed and the results are under evaluation. It is anticipated that this engine can be applied beyond the music domain to serve other genres. The technology is quite obviously applicable to the television or film domain. More ground-breaking use may include chart of the news or even an index of politicians. It is expected that there will be new services which build on this success at some point in the near future.

The Sound Index is "the first definitive music chart for the internet age" [5]. It is a novel demonstration of applying ground-breaking research to keep the music industry relevant. It has been named Web 2.0 technology of the week by the UK Observer for several consecutive weeks and has been named the hottest thing in music (in March 2008) by the UK Guardian Music Monthly.

The project led to other fundamental and pioneering technology being developed for the ingestion, processing, integrating and visualization of Web data from multiple sources:

1. Holistic Disambiguation – allows Web comments like "U R 50 bad", "the guitarist killed last night", "you are sh*t", "you are the sh*t" and "pink was off the chain" to be transformed into their intended equivalents, i.e. "You are so good", "the guitarist was excellent last night", "your music is horrible", "you are great", "pink was terrific" and finds the referred entities, e.g. pink refers to Pink Floyd and not Pink the R&B singer.

2. Automated Change Detection - provides an 'early-warning' capability to allow fast detection (and resolution) of changes on the websites.

The Sound Index also serves as a model for the new era in business innovation. It demonstrates the next wave in delivering better services and products – the real-time integration of multiple, relevant online information with one's own data to drive new and significant value for, re-invigorate connection to and strengthen brand affinity to one's customer base.

## 3. Other Applications

As mentioned before, the need to collate online data of multiple modalities is an issue that is present industries. Currently, we applied the same technology and techniques to the automotive industry [6] and healthcare sector [7]. In so doing, we have demonstrated that the technology is generic enough to be applied cross-domain. However, in the process will have learned that each domain has its own set of special concerns that must be addressed.

## 4. Insight for Developing Nations

The mining of Web X.0 data is still an emerging field with industry and academic leaders just starting to explore the space. Currently, there is room for all interested parties to enter into this arena. The barrier to entry is currently quite low. The only requirements, which we have garnered from our experiences with paying customers, are 1) skilled ICT practitioners with knowledge of text analytics techniques, 2) cheap bandwidth, 3) at least a set of current personal computers.

The latter two requirements can be easily met in a developing country, such as Jamaica. The first requirement is predicated on enhancing current syllabi to include courses of programming fundamentals, UIMA (Unstructured Information Management Architecture) [8] and free text processing & analysis. As MIT and other universities have released a lot of their course materials in these areas online for free, meeting this requirement should be far less taxing than initially thought, because one does not have to hire new dedicated resources to met this need.

The most significant positive outcome from taking this course of action early would be that the developing community would be the experts and primary providers of technology, which spans multiple industries, e.g. construction, public relations, automotive, government, healthcare, telecommunications etc. In doing so, developing countries could take a very significant stride in closing the digital divide between developed and developing nations. It is hoped that this step would be the start of an ICT revolution and a new era of innovation led by the developing world.

The current posture of developing countries, like Jamaica, to be pillars of outsourced glory is not sustainable; as this is driven by the availability of cheap labor and government incentives. In targeting only outsourcing-oriented ventures, developing countries set up a system of internal competition with other developing countries, where the one with the lowest price point will get the business.

Diversification in technology pursuits is a prudent approach to take in today's economy and stepping to the forefront of new and emerging applied computer science areas like online buzz is a very viable option.

## 5. Related Work

There is a wealth of work in the industrial and academic communities that showcase the value of traditional information integration and aggregation techniques [9], where systems compare and contrast items with identical modalities, such as sales numbers from different sources. The Sound Index demonstrates how to integrate information from many different modalities (e.g., comments, passive listens, sales, hits on a website,

creation of new website, views on television, etc.), which is a solution that is required in many domains, e.g. patient preferences, drugs for certain medical conditions, cars, wine, financial products (stocks, bonds), consumer goods, cameras, computers, books, etc.

In Industry, Nielsen's BuzzMetrics technology [10] seeks to achieve a similar goal to the Sound Index, at the abstract level. Their technology examines consumer-generated media (CGM), e.g. blogs, message boards, forums, usenet newsgroups, discussions from email portals like Yahoo!, AOL and MSN, opinion and review sites and feedback & complaint sites and then analyzes, customizes and presents this data to marketers and business intelligence professionals, according to client requirements. At the time of writing, there was no publicly available technical information on BuzzMetrics. Our educated guess is that their technology relies heavily on natural language and sentiment processing, while the Sound Index relies heavily on broken English text analytics technology, techniques for the integration of information of different modalities and ranking technology.

Alexa Internet [11] is another industry instance of technology that is in the same space. They crawl web sites to produce a ranking of sites, based on traffic statistics, incoming links, etc. Their ranking methodology provides an ordered list of the sites with the most (incoming) traffic, normally filtered by geography or some other criteria. This differs from the Sound Index in that the Sound Index combines data of multiple modalities into a balanced ordered list.

From an engineering standpoint, the world of mashups mirrors our own requirements – that of a robust, reliable and repeatable means of gathering data from a variety of online sources. *ScrAPI*s (Screen-scraper + API) have been proposed as a means of mitigating the problem of unreliable or unavailable APIs from content providers [12], but even they suffer from the same issues facing traditional screen-scrapers.Clearly indicate advantages, limitations, and possible applications.

## 6. Conclusion

The Sound Index represents a first of its kind and showcases how enterprises can embrace the Internet to enhance their businesses. More importantly, it presents an opportunity to innovate and lead in an emerging field for whomever is willing to explore.

## Acknowledgements

We would like to thank the BBC, specifically Geoff Goodwin (Head of BBC Switch), for their vision, support and encouragement and Varun Bhagwan, Alfredo Alba, Jan Pieper and Anna Liu, all from IBM Almaden Research Center, for helping to make this a reality. We also would like to acknowledge our partners in IBM Global Business Services (Bill J Scott and Aidan Toase) and at NovaRising.

## References

[1] V. Bhagwan, T. Grandison, and D. Gruhl. Sound Index: Music Charts For The People, By The People. To appear in Communications of the ACM 2009.
[2] IBM Almaden Research Center, MONGOOSE Technology, http://www.almaden.ibm.com/cs/projects/iis/mongoose/
[3] Nagios, http://www.nagios.org
[4] IBM, Tivoli Storage Manager, http:// www.ibm.com/software/tivoli/products/storage-mgr/
[5] C. Salmon. Click To Download. Guardian UK, http://arts.guardian.co.uk/filmandmusic/story/0,,2274132,00.html
[6] Z. Zakharian, M. Mishra, S. Chandramohan. Cars 2.0. Masters Thesis, San Jose State University, December 2008.
[7] IBM Almaden Research Center, Health-e-Assistant, http://www.almaden.ibm.com/cs/projects/iis/hea/
[8] Unstructured Information Management Architecture (UIMA), http://www.research.ibm.com/UIMA/
[9] H. Zhu, M. D. Siegel and S. E. Madnick. Information Aggregation: A Value-added E-Service. Proceedings of the International Conference on Technology, Policy, and Innovation: Critical Infrastructures, The Netherlands, June 26-29, 2001.
[10] BuzzMetrics, http://www.nielsenbuzzmetrics.com/products
[11] Alexa, http://www.alexa.com/site/company/technology
[12] J. Musser. scrAPIs http://programmableweb.com/2006/03/21/scrapis