<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dashboards and Scorecards</td>
<td>4</td>
</tr>
<tr>
<td>Hugh J. Watson</td>
<td></td>
</tr>
<tr>
<td>Integrating Structured and Unstructured Data Using Text Tagging and Annotation</td>
<td>8</td>
</tr>
<tr>
<td>Sreekumar Sukumaran and Ashish Sureka</td>
<td></td>
</tr>
<tr>
<td>Data Quality for Enterprise Risk Management</td>
<td>18</td>
</tr>
<tr>
<td>Vikram Ramchandra and Sreedhar Srikant</td>
<td></td>
</tr>
<tr>
<td>Experts’ Perspective: Mobilizing Business Intelligence</td>
<td>24</td>
</tr>
<tr>
<td>Howard Dresner and James Thomas</td>
<td></td>
</tr>
<tr>
<td>Beyond the Dashboard: Making Better Decisions</td>
<td>30</td>
</tr>
<tr>
<td>Gabriel Fuchs</td>
<td></td>
</tr>
<tr>
<td>Operationalizing Business Intelligence— Turning Insight into Action</td>
<td>35</td>
</tr>
<tr>
<td>Bennett Indart</td>
<td></td>
</tr>
<tr>
<td>Closing the Business Process Management Execution Gap</td>
<td>39</td>
</tr>
<tr>
<td>Russell Keziere</td>
<td></td>
</tr>
<tr>
<td>Increasing BI Adoption: An Enterprise Approach</td>
<td>44</td>
</tr>
<tr>
<td>Anupam Manglik</td>
<td></td>
</tr>
<tr>
<td>BI Bookshelf</td>
<td>53</td>
</tr>
<tr>
<td>BI StatShots</td>
<td>55</td>
</tr>
</tbody>
</table>
Abstract
The popular practice of handling structured and unstructured data as distinct information entities often results in decision management failure. Studies show that the majority of data resides in the unstructured format in many organizations, so an infrastructure that meaningfully integrates and manages structured and unstructured data would act as a complete data warehouse (CDW)—the fundamental backbone for true enterprise business intelligence. This approach would also enable data owners to treat data as data rather than segregating it as structured or unstructured.

We present the role of text tagging and annotation techniques as a preprocessing step toward the integration of unstructured and structured data. The nature of unstructured data makes it hard to search, retrieve, and analyze; directly integrating it with structured data is non-trivial. However, structure and semantic information can be added to unstructured text content using natural language processing techniques; combining this data with structured data is thus more efficient.

Text tagging and annotation are also popular natural language processing techniques. We present a high-level architecture of a generic system that uses text tagging as a preprocessing step to integrate structured and unstructured data. We illustrate the benefits of tagging natural-language text with two real-world applications.
Introduction

The amount of data stored in an enterprise is growing quickly. The ability to access and analyze data sources for intelligent decision making is a key element to an organization's success. Enterprises evolve and transform over time, resulting in a heterogeneous world of information where data is distributed across a variety of sources and systems. Data stored in different systems, locations, formats, and schemas poses a challenge to integration and usability.

Operational data typically stored in relational database management systems (RDBMS) and data in a data warehouse are examples of structured data, and several mature tools exist to enable data integration across structured sources. Moreover, a variety of tools are available that extract, transform, and load (ETL) data from disparate databases into a single data store for reporting and analysis.

Enterprises are increasingly interested in accessing unstructured data and integrating it with structured data. Unstructured data consists of freeform text such as word processing documents, e-mail, Web pages, and text files, as well as sources that contain natural language text. Although unstructured data also includes audio and video streams as well as images, the focus of this article is on freeform text.

Unlike structured data, the nature of unstructured data makes it more difficult to query, search, and extract, complicating integration with other data sources. Data stored in a structured format is inherently record-oriented; it is typically stored with a predefined schema, which makes it easy to query, analyze, and integrate with other structured data sources. Unlike structured data, however, the nature of unstructured data makes it more difficult to query, search, and extract, complicating integration with other data sources.

Building BI and performance-management solutions that rely solely on the structured data is equivalent to making a decision based on limited and incomplete information.

Regardless of the complexity in manipulating and integrating unstructured content, there is a strong need to build tools and techniques for managing such data. Some 80 percent of the data residing in an enterprise is in unstructured format (Knox et al., 2005). Building BI and performance-management solutions that rely solely on the structured data (which constitutes a small percentage of the organization’s data) is equivalent to making a decision based on limited and incomplete information.

The information hidden or stored in unstructured data can play a critical role in making decisions, understanding and complying with regulations, and conducting other business functions. Integrating data stored in both structured and unstructured formats can add significant value to an organization. Such integrated data will define the CDW infrastructure so an organization can derive a single version of truth.

Text tagging and annotation is a popular technique based on natural language processing and machine learning and is an important component of a document processing and information extraction system. Text tagging and annotation consists of analyzing freeform text and identifying terms (for example, proper nouns and numerical expressions) corresponding to domain-specific entities.
The text annotator input is freeform text; the output is a set of named annotations for text sections.

Text annotation is also referred to as named entity (NE) extraction, and in earlier days the named entity extraction technique was used to identify common entities such as persons, locations, organizations, dates, and monetary amounts from newswire text. Named entity detection has been the subject of research for more than a decade, and has been incorporated into both open-source and commercial systems. Current named entity detection systems offer a good degree of accuracy and are widely used in diverse domains, with applications in text mining, information extraction, and natural language processing.

In this article, we demonstrate the value of text tagging and annotation as a preprocessing step toward integrating structured and unstructured data. Text annotation is used to add semantic information or structure to unstructured data originating from such sources as e-mail, text files, Web pages, and scanned, handwritten notes. Meaningful information is added to large amounts of text, which can then be integrated with structured data for further analysis. With two real-world examples, we illustrate the efficacy of the text tagging technique in the context of structured/unstructured data integration.

In the next section, we provide a brief introduction to text tagging and annotation and offer a high-level overview of popular underlying techniques behind text tagging. We follow that discussion with a generic and high-level architectural diagram of a system that makes use of text tagging as a preprocessing step toward integrating structured and unstructured data.

**Figure 1.** ETL, text tagging, and annotation are used to build the complete data warehouse (phase 1)
Text Tagging and Annotation

Text tagging and annotation, also called named entity extraction, forms an important component of many language-processing tasks, including text mining, information extraction, and information retrieval.

Named entity extraction consists of identifying the names of entities in freeform or unstructured text. Among the common types of entities are proper nouns, names, products, organizations, locations, e-mail addresses, vehicle data, times and dates, and numerical data such as measurements, percentages, and monetary values.

Domain-specific entities are included as well. Named entity extraction has applications in diverse domains, such as detecting chemical and protein names from medical literature; gaining market intelligence by detecting personal names, locations, organization names, and product names in newswire text; finding names of weapons, facilities, and terrorist organizations for military and defense purposes; or building a semantic search application to overcome the limitation of regular keyword-based search engines.

Several approaches and techniques have been developed to perform named entity extraction, from manually developing a set of rules and using a dictionary or a list lookup from pre-existing databases to linguistic analysis and machine learning.

Generic and High-Level Architecture Diagram

The process of gathering intelligence from structured and unstructured data sources occurs in two phases. In phase 1, structure is added to unstructured data using named entity extraction, then the results are integrated with the structured data. The output from phase 1, the CDW, acts as one of the inputs to phase 2, in which a BI application is built atop the single version of truth represented by the CDW.

As shown in Figure 1, data within an enterprise can come from traditional transactional sources such as an RDBMS, legacy systems, and repositories of enterprise applications, and from unstructured data sources such as file systems, document and content management systems, and mail systems.
To build an effective decision-support backbone, this data must be moved into the CDW. An ETL process executes the required formatting, cleansing, and modification before moving data from transactional systems to the CDW. In the case of unstructured data sources, the tagging and annotation platform extracts information based on domain ontology into an XML database. As in Figure 1, extraction of data from an XML database into the CDW is accomplished with an ETL tool. This materializes the unified data creation into the CDW—the foundation for the organization’s decision-support and BI needs.

The CDW offers an all-encompassing view of an organization’s data assets for building BI and decision-support applications. As illustrated in Figure 2, the CDW enables the building of reliable BI and decision-support applications based on holistic, all-inclusive enterprise data. With the CDW, applications such as corporate performance management can produce highly reliable results.

**Market Intelligence from Daily News Feeds**

Access to accurate and real-time information about competitors and the market is crucial. Decision makers must absorb and analyze a huge amount of information generated every day. To remain competitive, an organization needs to be aware of continuously changing market trends, competitor policies, product launches, mergers and acquisitions, and management changes, among other information published in daily newspapers, magazines, newsletters, and Web sites. The daily news articles and reports are freeform text—a vast collection of unstructured

---

**Information Sources**

- News feeds
- Articles and reports
- Newsgroups and forums
- Blogs

**Advantages**

- Make effective and informed decisions
- Strategy planning

**Figure 3. Gaining market intelligence from news feeds**

With the CDW, applications such as corporate performance management can produce highly reliable results.
data that is difficult and time-consuming to review and analyze. The quality of a decision is directly related to the quality of its information inputs, so it is important to analyze as much quality information as possible in a limited amount of time.

Figure 3 illustrates a scenario in which an executive gains market and competitive intelligence by browsing information from sources such as daily news feeds, forums, blogs, articles, and reports. It may be straightforward for the human mind to discern from the news headline “Oracle Acquires Innobase” that the story concerns an acquisition and to recall that Oracle and Innobase are software database companies. However, it is non-trivial to invoke a query directly on this natural language text headline to decipher the meaning behind it. How can a text annotation tool add structure to unstructured data and thus make it more amenable to query and search?

Consider the company acquisition news example in Figure 4. The news text is:

On November 16, 2005, IBM announced it had acquired Collation, a privately held company based in Redwood City, California, for an undisclosed amount.

The entity types present in this news text are Date, Acquiring Organization, Acquired Organization, Place, and Amount. As shown in Figure 4, a text annotator identifies the entities and tags them. The output can be in the form of an XML document or a database table. Tagging important named entities makes it easy to carry out an entity link and relationship analysis. The XML tags or the database table schema has to be predefined by the user and is domain-specific. The text annotator tool also needs to be customized or programmed so it can detect specific entities. SQL queries can be easily performed on the table produced by the text tagging and annotation process.

Figure 5 illustrates the potential benefit of combing the structured and unstructured data. The answer to the question, “List companies acquired by IBM in the first quarter,” one needs to examine a large amount of unstructured content from news articles and discern the relationships between IBM, the companies it acquired, and the acquisition dates. The report of the IBM acquisition may not directly mention Q1 or Quarter One as the date or time of the event. The knowledge that the months January through March comprise the first quarter of a financial year comes from an external knowledge base or taxonomy.

How can a text annotation tool add structure to unstructured data and thus make it more amenable to query and search?

In another query, a user wishes to know the number of companies acquired in the U.S. where the financial size of the transaction exceeds a user-specified value. Very often the news article does not mention the country name since it is obvious from the state or city name. If a news report mentions the state of California or New York in a text, the user knows the acquisition occurred in the United States.

A domain ontology containing entities and their relationships is prepared by a domain expert. The ontology is combined with the output from processing unstructured content. The application end user is unaware (and unconcerned) that the information displayed was retrieved from multiple sources.

Product Performance Insights from Customer Warranty Claims Data

A quality early warning system is a BI application used to analyze large volumes of warranty claims data for diagnosing the root cause of a product’s failure. Product defects and warranty claims result in heavy costs to manufacturers. Companies see considerable value in
building a quality early warning system that, by processing warranty data, helps in the early discovery of product and system failures.

Warranty data is generally gathered when a claim form is completed by a customer and a technician. The form is entered on paper (which is scanned and imported into a database) or the information is directly entered online. Forms ask for the product code, model number, date, time, and customer ID. This information falls into the category of structured data—the information has a well-defined format and requires closed-ended answers (there are finite choices for some fields).

Usually the form also contains a comments section where a customer or technician can provide detailed information about the problem. This is the section where information is entered as natural language text or free-form text, and this unstructured data is key to diagnosing and understanding the problem.

If a manufacturer suspects a recurring problem, it sifts through claims data and manually examines the customer and technician comments for patterns or clues that can identify the cause of the problem. The large number of claim forms renders the manual process of reading all comments time-consuming and impractical. Hence, there is a strong business case for automating note analysis that gathers product performance information, and significant value in building a system that extracts information from unstructured data (the customer

**Figure 4.** Text annotation and tagging

<table>
<thead>
<tr>
<th>Date</th>
<th>Acquiring organization</th>
<th>Acquisition event</th>
<th>Acquired organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 6</td>
<td>IBM</td>
<td>Collation</td>
<td></td>
</tr>
</tbody>
</table>

On November 16, 2005, IBM announced it had acquired Collation, a privately held company based in Redwood City, California for an undisclosed amount.

<table>
<thead>
<tr>
<th>Date</th>
<th>Organization</th>
<th>Place</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 6</td>
<td>IBM</td>
<td>California</td>
<td>Undisclosed</td>
</tr>
</tbody>
</table>

On <DATE> November 16, 2005</DATE>, <ACQUIRING ORG> IBM </ACQUIRING ORG> announced it had <ACQUISITION EVENT> acquired </ACQUISITION EVENT> <ACQUIRED ORG> Collation </ACQUIRED ORG>, a privately held company based in <PLACE> Redwood City, California </PLACE> for an <AMOUNT> undisclosed </AMOUNT> amount.
and technician notes) and links it with an external knowledge base.

Figure 6 illustrates a warranty claim form received by a computer manufacturer. The form is partially structured in the sense that certain form fields have a well defined format. Other fields allow a user to enter freeform text. The user provides details (using natural language) that describe the technical difficulty or defect in the product. The text of the user’s complaint contains many domain-specific entities. For example, hard disk is an entity of type “Computer Parts” or “Components”; crashed is a “Problem” entity.

As with customer comments, a technician’s problem diagnosis is written in freeform text, from which many domain-specific entities can be extracted. The underlying technology used to tag the text can be based on a dictionary lookup, rule-based pattern matching, or machine learning; the end result is identification and marking of domain-specific entities. As shown on the right-hand side of Figure 6, the output of the text tagging and annotation process is an XML file containing the extracted entities enclosed within XML tags. The tags to be extracted are predefined by the user; the entity-recognition logic is also programmed into the text annotator.

The XML file produced by the text tagging process is in a form amenable to query, search, and integration with other structured data sources. Figure 7 illustrates an example of a pattern that can be discovered from analyzing unstructured data: a computer manufacturer receives thousands of warranty claim forms through its dealer network. All forms are imported into a central repository for further analysis.

The first step in gathering intelligence from these forms is to tag and annotate the text—this is the named entity

---

**Figure 5.** Integrating structured and unstructured data to gather market intelligence

---

List companies acquired by IBM in Q1.
List all software companies acquired in USA with a deal amount greater than $200 million.
Number of hardware companies acquired by Intel in past 6 months.
The goal is to create an emerging-issue report for the engineering and quality teams so they can take proactive measures to fix common problems and their root causes by integrating the unstructured data reported by customers and technicians with the structured data stored in an RDBMS or an ontology. The goal is to create an emerging-issue report for the engineering and quality teams so they can take proactive measures.
to fix common problems, reducing product recalls and warranty liability.

Summary
Text tagging and annotation plays a strong role in integrating structured and unstructured data. The outcome of such integration—the complete data warehouse—becomes the fundamental infrastructure for decision support and business intelligence.

Removing the barrier between structured and unstructured data significantly impacts the way companies treat and govern data. By its nature, unstructured data makes it difficult to directly extract meaningful information and combine it with structured data sources. However, structure and semantic information can be added to unstructured data through text tagging and annotation, making it suitable for integration with other data sources.

As we’ve shown, the technology can be applied to help an enterprise gather market intelligence from newswire text or gain intelligence about product performances from customer warranty claim forms. Business and technology owners of enterprisewide data management implementations realize that the complete data warehouse is the foundation for efficient and accurate business decisions.

REFERENCES

Paquet, Raymond. “Poll shows many organizations lack the foundation for successful SRM,” Gartner Research, April 2005.


Figure 7. Product performance insights from a large volume of claims data