I

Unstructured Textual Data in the Organization

Most organizations have two kinds of data: structured data and unstructured textual data. Because structured data preceded unstructured data in the workplace, unstructured data is often best understood in contrast to structured data. Structured data is data that is represented by numbers, tables, rows, columns, attributes, and so forth. Most IT professionals have spent the better part of their professional lives with structured data. As its name implies, structured data is usually disciplined, well behaved, predictable, and repeatable.

Structured data usually is generated as by-products of doing a transaction. A check is cashed, an ATM activity is done, an insurance claim is made, a production run is completed, a car is sold—these are typical transactions that generate a lot of structured data about the activity that has been done.

Structured data is made up of data types that are repeated continually. The same types of data are found in almost every transaction. The only things that differ from one transaction to the next are the values that the data types take. In addition to repeatability and predictability, the essence of the structured environment is numeric data. Although there is text in the structured environment, most text serves the purpose of identifying or describing some numeric data. The numeric data in the structured environment makes up the heart of the data that is found there and is heavily used for analytical purposes.

Unstructured Textual Data

The other major category of data found in the corporation is unstructured data. There are several forms—textual unstructured data and nontextual unstructured data, which includes images, colors, sounds, and shapes. This book is about textual unstructured data, which presents enough challenges on its own to fill a book (or even more than a book!).

Unstructured textual data is textual data found in emails, reports, documents, medical records, and spreadsheets. There is no format, structure, or repeatability to unstructured textual data. There is no one sitting on your shoulder telling you what to do when you write an email. You can write anything you want, however you want, and use any language you want. In addition, there are other forms of text that occur well outside the email environs, such as contracts, warranties, spreadsheets, telephone books, advertisements, marketing materials, annual reports, and many more forms of textual information that are the fabric of the organization. In short, unstructured textual data occurs almost everywhere and represents both a challenge and an opportunity to the organization that wants to use it for decision-making purposes.

It is true that many forms of unstructured data are not text-based. There are X-rays showing bones and breaks, real-estate listings with pictures, engineering change control documents mapping the structural changes made to complex edifices, MRIs that show detailed aspects of the human body, and scientific photos that help mankind unlock the secrets of the universe. But the most basic form of unstructured data is in the form of text. The focus in this book is on text, which presents its own set of challenges.

The purpose of this chapter is to provide an overview of unstructured textual data and the environment in which it sits. Unstructured textual data is so pervasive, is so ubiquitous, and has so many variations that it is hard to classify. In one place, unstructured textual data has one set of characteristics. In another place, unstructured textual data has a completely different set of characteristics. Because of this topsy-turvy, unpredictable nature, it is extremely difficult to generalize how to approach unstructured textual data.

Following are some examples of the nonuniform characteristics of unstructured textual data:

- **Emails**—Emails are usually short relative to other documents. Emails usually contain a combination of business-related and nonbusiness-related information. There are usually a lot of emails, many of which are informal. Emails can be in any language (English, Spanish, German). In fact, emails can be in more than one language at the same time. Emails are normally identified by an email address and the time the email was sent.

- **Medical records**—Medical records can become quite voluminous and are full of medical jargon and terminology. Medical records are often quite large, some containing volumes of text. The reason why most medical records are text-based is that doctors prefer it that way. Most doctors prefer to write notes in text when dealing with a patient. In most environments, there are almost always fewer medical records than there are emails. (Although in the medical environment, there still are plenty of medical records.) Medical records are somewhat formalized, in the sense that doctors have a certain protocol that is used when writing a medical record. These records usually contain only information relevant to health and medicine.
- **Contracts**—Contracts are usually full of legal jargon and might contain business-related jargon as well. The text found in contracts usually has nothing but information relating to the contract. Contracts by and large vary greatly in size—some contracts are short, and some contracts are lengthy. Contracts are written in a legal style, according to the conventions of law and lawyers. There sometimes are a fair number of contracts in an organization, but never the number of contracts that there are emails. Contracts are almost always in a single language–for example, English, Spanish, or French.

And these are but a few of the types of unstructured textual data that exist in the organization. Each type of unstructured textual data has its own peculiarities and its own characteristics.

Unstructured Textual Data and Organizational Functions

To start to understand unstructured textual data, consider that there are different kinds of unstructured textual data associated with different functions within the organization. Table 1-1 shows some of the corporate functions and the unstructured textual data that is typical of those departments.

Table 1-1 Corporate Functions—Unstructured

Corporate Function	Unstructured Data Types
Accounting	Spreadsheets, notes, Word documents, audit trails, account descriptions
Call center	Conversations, notes, replies
Engineering	Bill of material, engineering changes, production archives, design specs
Finance	Spreadsheets, notes, annual reports
Human Resources	Emails, letters, hiring offers, termination documentation, evaluations, job specifications, employee manuals, holidays, policies
Legal	Agreements, amendments, proposals, contracts, meeting notes, telephone transcripts, patents, trademarks, nondisclosure

continues

Table 1-1 Corporate Functions—Unstructured (continued)

Corporate Function	Unstructured Data Types
Marketing	Ads, spreadsheets, targets, accounts, forecasts, webinars, seminars, conferences, booth notes, feedback, customer contact notes
Operations	Manufacturing runs, defective products, reservations, claims processing, precious goods store, delivery notes, scheduling notes
Sales	Sales leads, sales calls, sales meetings, sales forecasts, spreadsheets, performance evaluations, customer meetings
Shipping	Delivery directions, fragile specifications, cooling temperature specifications, time of delivery specifications, speed of delivery specifications, tracking

The corporate function of accounting typically has spreadsheets, Word documents, audit reports, and audit trails associated with its activities. Call centers typically have recorded or transcribed conversations, replies, follow-up activities, and other notes associated with their activities. The engineering department typically has unstructured textual data associated with the bill of materials, engineering changes that have been made, production archives, and design specifications.

Each of these different forms of unstructured textual data has its own set of characteristics.

Unstructured textual data is not just endemic to different departments of the organization. Unstructured textual data appears in different forms and in different measures in different industries. Some industries have a lot of unstructured textual data while others have little. Table 1-2 shows unstructured textual data by industry and emphasizes the differences.

Table 1-2 Industries and Unstructured Data

Corporate Type	Type of Transaction Processing	Amount of Unstructured Data
Banking, finance	Heavy transaction (tx) processing	Light, scattered
Construction	Light tx processing	Light, scattered
Government	Tx processing	Heavy concentrations
Healthcare	Tx processing	In the fabric
Information processing	Heavy tx processing	Heavy concentrations
Insurance	Heavy tx processing	Heavy concentrations
Manufacturing	Tx processing	Heavy concentrations
Medicine	Light tx processing	In the fabric
Mining	Light tx processing	Heavy concentrations
Pharmaceuticals	Tx processing	Heavy concentrations
Real estate	Light tx processing	Light, scattered
Retailing	Heavy tx processing	Some pockets

Corporate Type	Type of Transaction Processing	Amount of Unstructured Data
Scientific	Light tx processing	Lots of heavy concentrations
Transportation	Heavy tx processing	Light, scattered
Utilities	Tx processing	Light, scattered

Table 1-2 shows that different industries have different mixes of transaction processing data and unstructured textual data. For example, banks are rich with transaction processing, checks, ATM activities, and other banking activities. Whereas banks certainly do have unstructured textual data, they do not have (relatively speaking) nearly as much unstructured textual data as they have structured data. On the other hand, medical environments are rich in unstructured textual data. The unstructured data is so ingrained in healthcare that it is part of the fabric of medicine and healthcare. Doctors write and take notes in a textual fashion, hospitals take notes textually, and so on. The world of medicine is rich in text, and certainly transactions occur in the medical environment. Patients get billed on a regular basis, for example. But—relatively speaking—the medical environment is heavily ingrained with textual data.

Unstructured Data and Its Characteristics

One of the perplexing aspects of unstructured textual data is that the characteristics associated with the different forms of unstructured textual data are mixed across the many different forms of the data. There is little uniformity among the different forms of unstructured textual data.

Table 1-3 on unstructured textual data and characteristics points out the extreme lack of uniformity of the characteristics. Table 1-3 is complex and deserves an explanation. The columns are as follows:

- Direct business relevance refers to the unstructured textual data. Unstructured textual data that is directly business relevant would include legal documents, customer credit reports, insurance claims, and airline reservations complaints. Indirect business relevance of unstructured textual data might include human resources, employee evaluations, and some email.
- Formal/informal refers to the way the unstructured textual data is written. Informal unstructured textual data would include email and letters. Formal unstructured textual data would include contracts and quarterly reports. (Note: Only Warren Buffet can get away with an informal annual report, and even then there is a section that is formalized.)

Table 1-3 Unstructured Data and Its Characteristics

Corporate Data Type	Direct Business Relevance	Formal/ Informal	Jargon	Keys	Number of Documents	Repeated Data	Retention	Typical Storage	Updated	Volume
Advertising	High	Informal	Limited	Usually not	Few	Almost never	Long cycle	Electronic	No	Very limited
Company manuals	High	Formal/ informal	Large	Limited	Few	Very small amount	Long cycle	Paper, electronic	Yes	Very limited
Contracts	Very high	Formal/ informal	Large	Limited	Moderate	Almost never	Very long cycle	Paper	Occasionally	Very limited
Customer call feedback	High	Informal	Limited	Very infrequent	Moderate	Almost never	Limited	Electronic	Never	Can be voluminous
Customer complaints	High	Informal	Limited	Very infrequent	Moderate	Almost never	Limited	Electronic	No	Small to moderate
Descriptive text	High	Formal/ informal	Large	Usually not	Large	Almost always	Long cycle	Electronic	Yes	Moderate to large
Email	Mixed	Informal	Some	Specialized	Very large	Almost never	Varies	Electronic	No	Can be very large
Employee files	Low	Mixed	Limited	Almost never	Small	Very limited	Long cycle	Paper, electronic	No	Relatively small
Engineering bill of matls	Very high	Formal/ informal	High	Yes	Moderate	Yes	Varies	Electronic	Yes	Varies
Employee evaluations	Low	Mixed	Low	Almost never	Small	No	Long cycle	Paper, electronic	No	Small to moderate
Internet	Mixed	Informal	Mixed	Seldom	Small	No	Varies	Electronic	Infrequently	Moderate
Legal docs	High	Formal/ informal	Large	Very few	Moderate	None	Very long cycle	Paper, electronic	Occasionally	Usually very small
Marketing materials	High	Informal	Some	Very few	Small	Almost never	Varies	Paper, electronic	Never	Usually very small
Memos	Usually high	Informal	Some	Very few	Small	Almost never	Varies	Paper, electronic	Never	Very small

Corporate Data Type	Direct Business Relevance	Formal/ Informal	Jargon	Keys	Number of Documents	Repeated Data	Retention	Typical Storage	Updated	Volume
Order information	High	Formal/ informal	Some	Lots	Moderate	Sometimes	Not long	Paper, electronic	Yes	Small to moderate
Presentations	High	Informal	Some	Infrequently	Small	Almost never	Short	Paper, electronic	Sometimes	Small
Quarterly report	High	Formal/ informal	Some	Infrequently	Small	Almost never	Very long cycle	Paper, electronic	Never	Very small
Reference tables	High	Formal/ informal	Large	Lots	Small	Lots	Varies	Electronic	Yes	Small
Reports	High	Informal	Large	Infrequently	Small	Sometimes	Varies	Paper, electronic	Sometimes	Small
Shipments	High	Formal/ informal	Some	Lots	Moderate	Sometimes	Short	Paper, electronic	Sometimes	Small to moderate
Spreadsheets	Usually high	Formal/ informal	Some	Frequently	Small	Yes	Short	Electronic	Yes	Small to moderate
Telephone transcripts	Varies	Informal	Limited	Occasionally	Moderate	No	Varies	Electronic	No	Small to moderate
Warranties	High	Informal	Some	Yes	Moderate	Seldom	Varies	Paper, electronic	Never	Small to moderate
Word docs	Varies	Informal	Varies	Varies	Moderate	Seldom	Varies	Electronic	Yes	Small to moderate

- Jargon refers to the type of language used. Some unstructured textual data tends to have a lot of jargon, and others tend to have very little. Contracts have a lot of legal jargon. Scientific research tends to have a lot of scientific jargon. Emails usually have a little jargon. Employee human resources files typically have little jargon.
- Keys refer to the ability of the analyst to tie the unstructured textual data to some singularly identifiable entity, such as a customer, product, or salesperson. Emails have keys such as the email address. The engineering bill of materials has lots of keys, referring to a product. Employee evaluations might refer to only an employee number. Telephone transcripts typically are identified only by a time and telephone number.
- Number of documents refers to the general number of documents on hand. There are usually a lot of email documents and only a relatively small number of contracts on hand (although there are exceptions to this).
- Repeated data refers to the repetition of data within the unstructured textual data. Phone books are full of repetitive data. Emails—as a rule—are not filled with repetitive data.
- Retention refers to how long unstructured textual data is kept. Internet text is not usually kept around for a long time, but quarterly reports are kept for a very long time.
- Typical storage media refers to the media on which the unstructured textual data is stored. Transcripted telephone conversations are stored almost exclusively electronically. Emails are usually stored electronically; although, they are occasionally printed out. Many forms of unstructured textual data are stored in both paper and electronic format.
- Update refers to whether the unstructured textual data can be changed when originally created. (Note: Update does not refer to adding unstructured textual data to an existing body of data. It refers to the changing of textual data when created.) Emails are almost never updated. In fact, in some cases it is illegal to update emails. On the other hand, ordinary Word documents are updated all the time.
- Volume of data refers to the total volume of data that is associated with a type of unstructured textual data. Typically, there is a large volume of data associated with email. But there would be a small volume of data for advertising, for example.

When you look down any of the columns, you see that there is little or no rhyme-or-reason to the characteristics. One form of unstructured textual data has one set of characteristics, and the next form of unstructured textual data has a completely different set of characteristics. It is this complete lack of characteristic pattern coupled with the complexity of language that causes the difficulty with the automated usage of data.

Updating Structured and Unstructured Data

One of the major differences between the structured and unstructured environments is that in the structured environment data is updated on a regular basis. Whenever your ATM is used, whenever a deposit is made, whenever a check is cashed, your bank account is updated. And it is at these times that structured data is commonly measured, or some record is created.

But unstructured data—for the most part—doesn't change after creation. After a contract is created and signed, it might be amended, but the original contract cannot have wording changed. After an email is written and sent, it is not changed. After a magazine article is written and published, it is not changed. In case after case, unstructured data is not changed when published. This difference between structured data and unstructured data creates many operating differences in the environments.

The Challenges of Unstructured Textual Data and Analytical Processing

Traditionally, analytic processing is used for business analysis of structured data. Structured data is particularly amenable to analytic processing because structured data is

- **Shaped by transactions**—Transactions are predictable in that the same types of data appear over and over again.
- **Shaped by numbers**—Transactions have a lot of numerical data, and in many cases, it is the numerical data that is of primary interest. Furthermore, numeric information can be easily manipulated—addition, subtraction, multiplication.

For these reasons, analytical processing is a natural partner with structured data. However, unstructured textual data has none of these characteristics. To do analytical processing against unstructured textual data, it is necessary to overcome or address several obstacles. Some of these challenges follow:

- **Physically accessing unstructured textual data**—Unstructured textual data is stored in a wide variety of formats. Some formats are easy to read and other formats are difficult to read. However, unstructured textual data must be accessible wherever it hides, whether or not it is easily accessible.
- **Terminology**—Terminology is a real issue. Three people call the same thing something different. If analytical processing is to be done against unstructured textual data, there must be a rationalization of terminology, or the analyst cannot recognize when the text illustrates the same thing.

- **Languages**—What if analysis must be done against unstructured textual data that is in English, French, and Japanese?
- **Volume of data**—The sheer volume of data that can be collected can defeat the efforts of the analyst who wants to see what is contained in a body of text. The sheer volume of unstructured textual data, coupled with the fact that analysis usually has to look at every word, makes the resources required for some kinds of unstructured analysis daunting.
- **Differing priorities of unstructured textual data**—Some unstructured textual data is "hot," meaning that the unstructured textual data needs to be treated with a great deal of sensitivity. Other unstructured textual data is merely normal, meaning that nothing is sensitive or urgent about it. The "hot" unstructured textual data usually has real and dire circumstances surrounding it, such that it must be handled in a completely different manner than other unstructured textual data.
- **Searchability of unstructured textual data**—It is one thing to do a simple search of unstructured textual data. In a simple search, some data—an argument—is passed onto the search engine, and a search is made on the basis of the argument. For example, a search is made on the term felony. In a simple search, the term felony is used, and everywhere there is a reference to felony, a hit to an unstructured document is made. But a simple search is crude. It does not find references to crime, arson, murder, embezzlement, vehicular homicide, and such, even though these crimes are types of felonies. A more sophisticated search—an indirect search— finds references to these types of felonies.
- **The economics of the infrastructure**—A cost of the infrastructure is required to support the unstructured textual environment. There is the cost of software, storage, and hardware. In addition, the cost of the personnel is required on an ongoing basis to make the unstructured environment operate on a daily basis.
- **Security**—Some unstructured textual data is not secure. Other unstructured textual data requires a careful consideration for how the unstructured textual data is to be handled and who has access to it. As the organization moves forward with the unstructured environment, it must not be assumed that all unstructured textual data is open and available to anyone who has access to the unstructured textual data.

This is just the short list of challenges that await the organization that attempts to come to grips with the unstructured textual data environment.

The Opportunities of Unstructured Textual Data

From the preceding discussion, it is obvious that there is a cost—in time, money, and manpower—to get a handle on the unstructured textual data environment. Why would an organization want to come to grips with the unstructured environment?

A world of promise and opportunity in information is buried in the unstructured textual data environment. Organizations have a tremendous opportunity at making better decisions—more timely, more accurate, more informed decisions—when they incorporate unstructured information into the decision-making process. That is the reason why organizations need to come to grips with the unstructured textual data environment. Stated differently, organizations that look only at their structured data—usually transaction-based data—miss an entire class of information that waits to be used for the decision-making process.

Organizations that base all their decisions on structured data use only a portion of the corporate information on which to base decisions. It is like a manager making decisions solely on revenues for this month. Although this month's revenues are an interesting figure and are certainly important, a lot of other types of information are factored in as well. There are monthly expenses, revenue figures for next year, the projections for next year, the size of the customer base, and new product announcements that need to be considered.

So exactly what kinds of useful information can be gained from looking at unstructured textual data? Some examples of useful information that might be buried in unstructured textual data include the following:

- **Customer feedback**—Do customers like or hate a new product? A new service?
- **Contractual commitments**—How many contracts have represented productive agreements? Unproductive agreements? How have corporations performed over many different contracts?
- **Warranties**—Is there a pattern to the warranties that are presented to a company?
- **Medical information**—What medical conditions correlate with other medical conditions? If a subpopulation of subjects is created, how does that change the correlations to medical conditions?
- **Compliance**—How has the corporation met reporting obligations?
- **Security**—Are employees saying and doing things that are not proper according to the guidelines for corporate conduct?
- **Marketing "buzz"**—What is said in the customer community about new products? New services? The company and its activities?

- **Competition**—What is known about new products? New promotions? New services?
- **Human resources**—Is the company actually living up to its obligations for hiring practices? To opportunity? To dismissals?

But perhaps the biggest promise of unstructured textual data lies in its ability to be combined with structured data. As a simple example of combining structured data and unstructured textual data, consider emails. How important are emails for the creation of the complete picture for a customer? The answer is that communications with the customer are important.

Consider the following simple example. An organization has a lot of demographic information about its customers. Suppose there is a Mrs. Jones who is a customer. The company knows that Mrs. Jones

- Is a mother of three
- Is a college graduate (Tufts, 1975)
- Is a systems analyst
- Works for Merrill Lynch
- Makes $105,000 per year plus commission
- Drives a Lexus
- Vacations in the Bahamas
- Pays her bills on time
- Is married to John Jones
- Invests in mutual funds

With this information, the company thinks that it is prepared to deal with Mrs. Jones, establishing a personal and long-term relationship with her.

So how important is demographic information in understanding Mrs. Jones? Very important. How important is Mrs. Jones' demographic information in the face of having no information about communications? Not important at all.

It seems that Mrs. Jones ordered some goods a month ago. The goods were late, sent to the wrong address, and broken when they arrived. And Mrs. Jones sent a scalding email last week to the corporation. How important is it to know about communications when trying to establish a relationship with Mrs. Jones? Vitally important.

And what are emails? They are nothing but a form of unstructured textual data. There are many cases where establishing a relationship between unstructured textual data and structured data leads to business opportunities that are today unimagined.

Summary

In this chapter, we examined the totality of the world of unstructured textual data. There are many difficulties with unstructured textual data. It has many characteristics, depending on the type of unstructured textual data that is discussed. Different industries have different needs for unstructured textual data. Different functional areas in the organization have various needs for unstructured textual data.

There are many different challenges when using unstructured textual data, such as challenges of terminology, volumes of data, and the cost of the infrastructure.

With the challenges, there is the promise of opening up an entirely different world of processing and opportunity. Unstructured textual data by itself contains much useful and interesting information. But unstructured textual data coupled with structured data opens up even more opportunities for unlocking the promise of unstructured textual data.

As powerful as the opportunity for unstructured data in the organization and business is, the technical environment that holds unstructured data must be understood, because many of the opportunities for exploiting unstructured data are shaped by, or certainly influenced by, the technical environment. Chapter 2, “The Environments of Structured Data and Unstructured Data,” addresses the technology infrastructure appropriate to unstructured data.

Prentice Hall — 0132360292