Keeping a Watchful Eye on Incoming Information, At Less Cost

A Process Innovation To Cue Employees to Important New Developments

Andrew Shepard--CIA, 7 June, 1995

Overview: We can achieve a much needed breakthrough in our ability to keep up with incoming information if we filter it through a corporate relational database, rather than using hours of employee time to browse the many new documents that arrive each day.

US Intelligence faces a challenge to provide closer support to key consumers, during a period of shrinking resources and general information overload. Today, much of our time is consumed by reading the mail, even though the newest information often is not the best information to help a decision maker deal with an immediate problem. Intelligence managers say we need to do a better job of exploiting all available information, including more openly available information. Clearly, we need a way to keep track of incoming data that will enable fewer people to monitor much more data with less of their time.

One way to keep an eye on critical new information, and also improve the use of all available information, is to enter key information from incoming documents directly into a corporate relational database. In this way, new information can be correlated automatically with previously recorded data, so that collectors and analysts can be notified of new developments, not just new documents. But who would enter the data? This paper suggests that the process can be largely automatic--using machines to do the initial stuffing of a corporate database, followed by corrections and additions that would occur as collectors and analysts use the database in their routine work downstream. Commercial technology is widely thought to be incapable of supporting the automatic extraction and stuffing of relational information into databases, but--as outlined in the appendix--we can boost the performance of commercial text extraction software by automatically comparing new documents against information already in our files.

The Problem: a flood of documents that obscures critical information.

Documents have served well as a means to convey and store information, until recently. Today we are flooded with documents. Most of the interdependent organizations and work processes that make up US Intelligence are designed to generate documents. The outside world is producing ever more documents. The intelligence system has become choked and, in some areas, dysfunctional. We sometimes task collectors to obtain what we already know. We sometimes settle for estimates when we could be providing customers hard data that is in our files. The effects go beyond inefficiency: document overload adversely affects the way people define their work and standards of success, as they adjust to continuing growth in the raw information they are supposed to monitor (as described in "Coping With Information Overload," 23 April, 1992.)

Running the intelligence business as a document-based process consumes a lot of labor. Much work is done by employees at NSA, CIA/DO and elsewhere to draft a textual document--a "story"--from each new set of collected facts. More work is done by editors, supervisors, and clerical workers who review the story and issue a formal document to disseminate the facts. These facts cannot help intended users elsewhere in the intelligence process, however, until the recipients do still more work to recognize that the document is apt to contain facts that matter to them. Then, there is always a probability of error as busy employees quickly read the story laid out in the document and try to capture the relevant faets. Adding to the reader's difficulty is the problem that story lines, although carefully crafted, can soon become irrelevant or even obscure the needed facts, as the context of intelligence questions changes over time.

Each day, across the Community, employees work independently to extract information from the same documents. This identical document-reading activity represents unproductive overlap. In the past, we have investigated the <u>output</u> from multiple intelligence organizations, when looking for unnecessary overlap (but have found that different organizations generally use their employees to produce different information for different purposes). The greater redundancy appears to be on the <u>input</u> side, by employees who must each spend valuable time extracting raw information from the same documents.

To tackle information overload, we need a way for employees to be confident that they are being alerted to the most important information available, and that even the information they have not seen is being constantly checked on their behalf. We need a method that is efficient in surveying a very large amount of information and presenting a small amount.

An effective approach would be to monitor incoming data information via a corporate relational database.

A corporate relational database offers a breakthrough in montioring incoming information, because it permits a much more focused selection of information for presentation to the user. A database allows information that is scattered across multiple documents, from multiple sources, reported over multiple days or years, all to be correlated with the newest tidbits that have just arrived. The result is that a user can register interest in information patterns that he or she defines as suggesting important developments, and be automatically notified whenever such patterns emerge in the database.

As proposed here, the key role of the database is to organize large amounts of raw information. The database serves as an index that can take employees directly to key information deeply embedded in the source documents. It also supports initial correlation of facts and relationships, via link displays, time-series displays, geographic displays, etc.

The potential for using databases to enhance employee productivity in this way is straightforward. But the approach has not been affordable to date, because entering facts and relationships into a database has been a very labor intensive process. This impediment is disappearing. Technology ean now support affordable extraction of information from documents by taking advantage of information already present in the institutional files (see appendix)--and therefore the exploitation of information via relational databases could be an idea whose time has come.

A closer look at tracking new information--documents vs. databases

Figure 1 shows why a shift from documents to databases can turn information overload into an intelligence advantage. On the left, the figure shows a typical analyst inbox--a list of titles of new documents that have been delivered because they include a key word that is relevant to the analyst's areas of responsibility. There are too many documents (286) for an employee to read in the time available--and each day adds to the accumulated volume of unseen but pertinent information. Data arriving during a period of absence is likely never to be used. On the right, the figure shows mail from a relational database. In this approach, some of the information from new documents has been extracted and correlated with previous reporting, to search for specified developments. The first item, for example, highlights a newly discovered transaction, which became apparent when information from one document was correlated with information from two previous documents. The inbox on the right contains only 27 highlights, based on data gathered from many documents. This approach yields a much more manageable number of items to check; is focused on developments rather than documents; ensures that older relevant information is considered; and makes the relevance readily apparent.

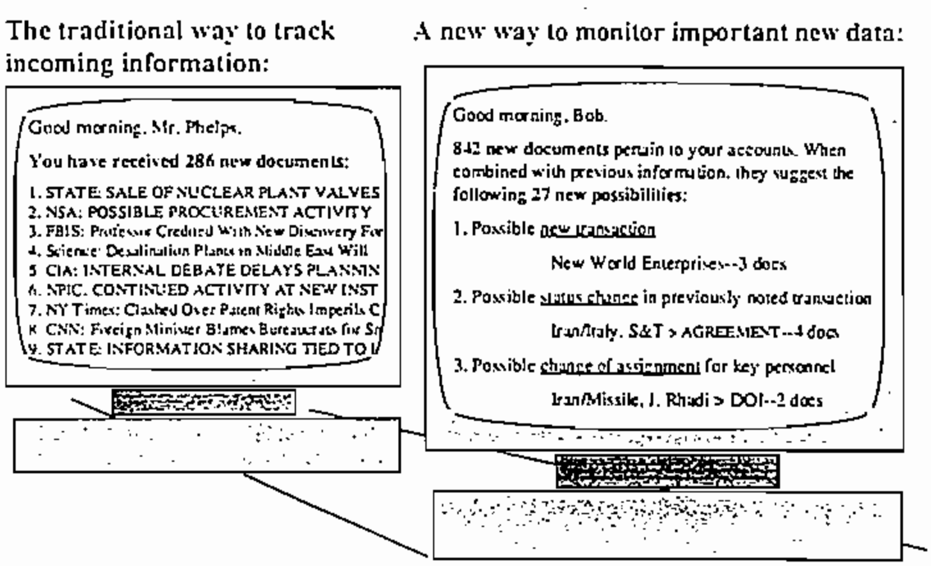


Figure 1: Old and new ways to view the daily traffic.

Figure 1 does not suggest that analysts will stop reading documents. It does suggest that analysts can stop rushing through masses of new documents, rapidly searching for data to correlate, on the chance that something important might be there. That mode is no longer feasible; there are simply too many documents and too little time to spend on each one. The challenge is to concentrate more attention on documents that are known to include important information. When this is possible, employees can take more time to read key documents carefully (see Figure 2).

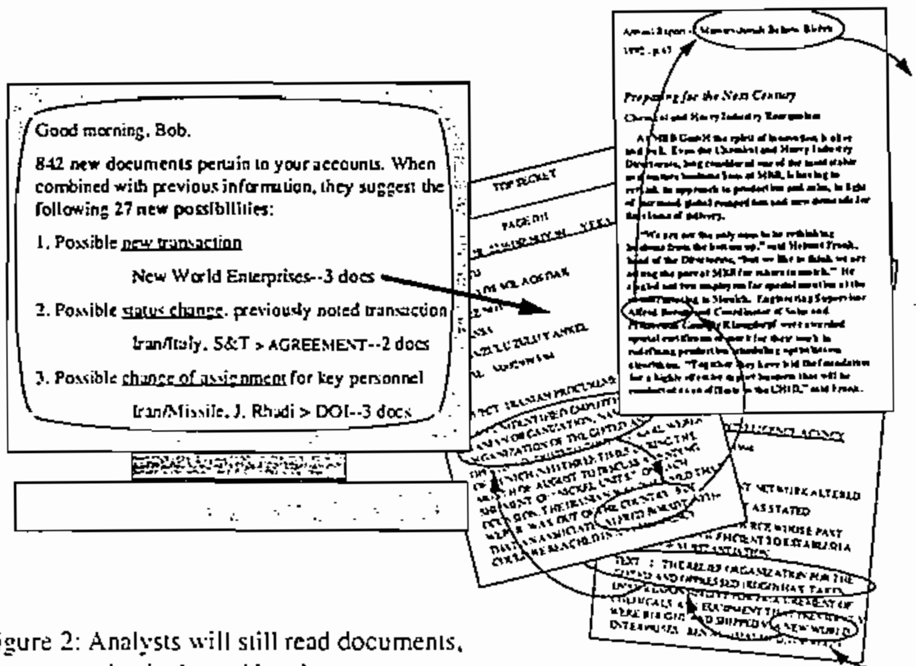


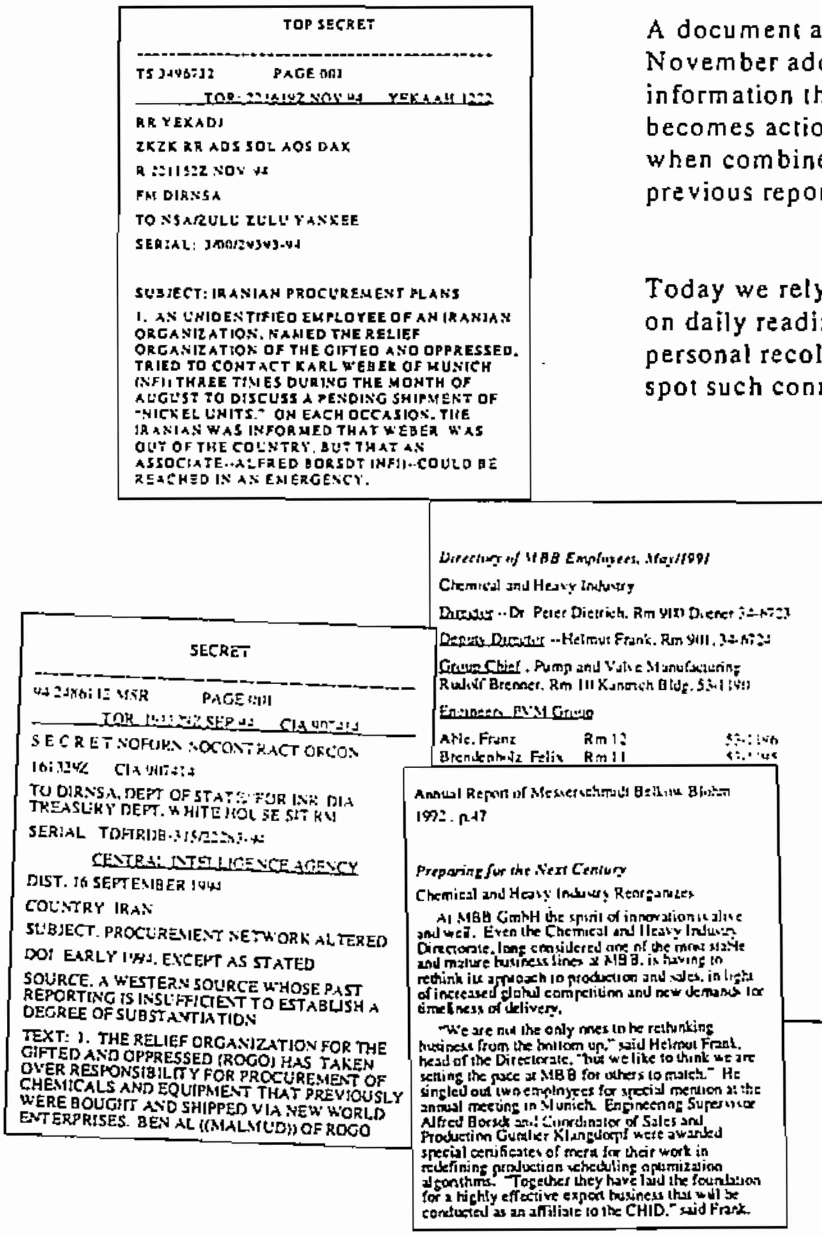
Figure 2: Analysts will still read documents, but more selectively and in a larger context.

The proposed breakthrough depends on a new approach to processing incoming information.

The proposed change in the employee's approach to tracking new data necessitates a new approach to processing information as it comes in the door. Instead of a central service distributing documents to inboxes, the new corporate process must break documents into key information elements that are entered into a relational database.

The process of breaking documents into key information elements requires some decisions about the kinds of data that are apt to be most revealing and most susceptible to automatic detection and extraction. To help illustrate the kinds of data elements that could be extracted--and the kinds of breakthroughs we could achieve in information tracking, Figure 3 presents some simulated documents. (The content is fictional but representative of the way in which key facts are often conveyed in textual prose.)

Figure 3: Simulated documents that illustrate potentially useful data.



TOP SECRET

TS 2496732 PAGE 001

TOR: 221619Z NOV 94 YEKAAH 1222

RR YEKADJ

ZKZK RR ADS SOL AOS DAX

R 221152Z NOV 94

FM DIRNSA

TO NSA/ZULU ZULU YANKEE

SERIAL: 3/00/24393-94

SUBJECT: IRANIAN PROCUREMENT PLANS

1. AN UNIDENTIFIED EMPLOYEE OF AN IRANIAN ORGANIZATION, NAMED THE RELIEF ORGANIZATION OF THE GIFTED AND OPPRESSED, TRIED TO CONTACT KARL WEBER OF MUNICH (NFI) THREE TIMES DURING THE MONTH OF AUGUST TO DISCUSS A PENDING SHIPMENT OF "NICKEL UNITS." ON EACH OCCASION, THE IRANIAN WAS INFORMED THAT WEBER WAS OUT OF THE COUNTRY, BUT THAT AN ASSOCIATE--ALFRED BORSDT (NFI)--COULD BE REACHED IN AN EMERGENCY.

Directory of MBB Employees, May1991

Chemical and Heavy Industry

Director -- Dr. Peter Dietrich, Rm 910 Diener 34-6723

Deputy Director -- Helmut Frank, Rm 901, 34-6724

Group Chief, Pump and Valve Manufacturing
Rudolf Brenner, Rm 10 Kanerich Bldg, 53-1190

Engineers, PVM Group

Ahle, Franz	Rm 12	53-1196
Brendenholz, Felix	Rm 11	53-1195

SECRET

94 2486112 MSR PAGE 001

TOR: 161329Z SEP 94 CIA 907414

S E C R E T NOFORN NOCONTRACT ORCON

161329Z CIA 907414

TO DIRNSA, DEPT OF STATE/FOR INR, DIA
TREASURY DEPT, WHITE HOUSE SIT RM

SERIAL TDFIRDB-315/22263-94

CENTRAL INTELLIGENCE AGENCY

DIST. 16 SEPTEMBER 1994

COUNTRY IRAN

SUBJECT. PROCUREMENT NETWORK ALTERED

DOI EARLY 1994, EXCEPT AS STATED

SOURCE. A WESTERN SOURCE WHOSE PAST REPORTING IS INSUFFICIENT TO ESTABLISH A DEGREE OF SUBSTANTIATION

TEXT: 1. THE RELIEF ORGANIZATION FOR THE GIFTED AND OPPRESSED (ROGO) HAS TAKEN OVER RESPONSIBILITY FOR PROCUREMENT OF CHEMICALS AND EQUIPMENT THAT PREVIOUSLY WERE BOUGHT AND SHIPPED VIA NEW WORLD ENTERPRISES. BEN AL ((MALMUD)) OF ROGO

Annual Report of Messerschmidt Bolkow Blohm
1992, p.47

Preparing for the Next Century

Chemical and Heavy Industry Reorganizes

At MBB GmbH the spirit of innovation is alive and well. Even the Chemical and Heavy Industry Directorate, long considered one of the most stable and mature business lines at MBB, is having to rethink its approach to production and sales, in light of increased global competition and new demands for timeliness of delivery.

"We are not the only ones to be rethinking business from the bottom up," said Helmut Frank, head of the Directorate, "but we like to think we are setting the pace at MBB for others to match." He singled out two employees for special mention at the annual meeting in Munich. Engineering Supervisor Alfred Borsdt and Coordinator of Sales and Production Gunther Klangdorpf were awarded special certificates of merit for their work in redefining production scheduling optimization algorithms. "Together they have laid the foundation for a highly effective export business that will be conducted as an affiliate to the CHID," said Frank.

A document arriving in November adds new information that becomes actionable when combined with previous reporting.

Today we rely heavily on daily reading and personal recollection to spot such connections.

In order to make a breakthrough in our ability to keep track of important new information as it comes in the door, we would need an automated process to integrate information from multiple sources and compare it against some definitions of "important developments." This paper proposes automatic extraction of data from each document as it arrives; storing the resultant data elements in a relational database; and regularly checking the database, automatically, for patterns that employees have defined as meaningful. Key data elements extracted from three documents are shown in Figure 4.

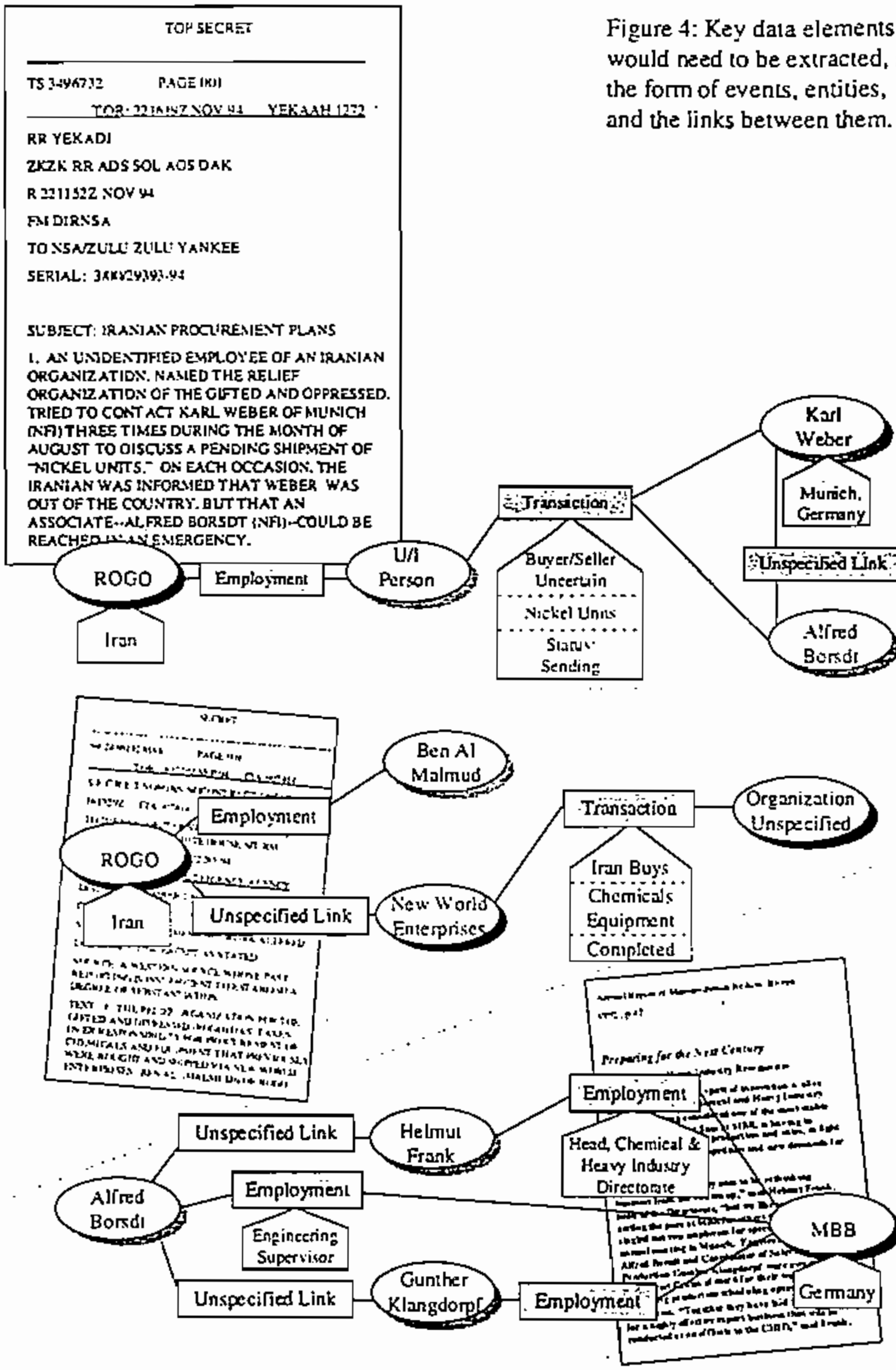


Figure 4: Key data elements would need to be extracted, in the form of events, entities, and the links between them.

The data and relationships shown here would be loaded automatically into a corporate relational database, as the documents arrive. The data from the top document would be added to information already placed in the database, from the bottom two documents, months and years before. When the November data is added, the automatic checking of overall information uncovers the noteworthy link across all three documents--and triggers a highlight message to the requesting employee. In this example, the analyst had provided instructions to report "any new transaction involving New World Enterprises," and the automatic process found the correlation shown in Figure 5.

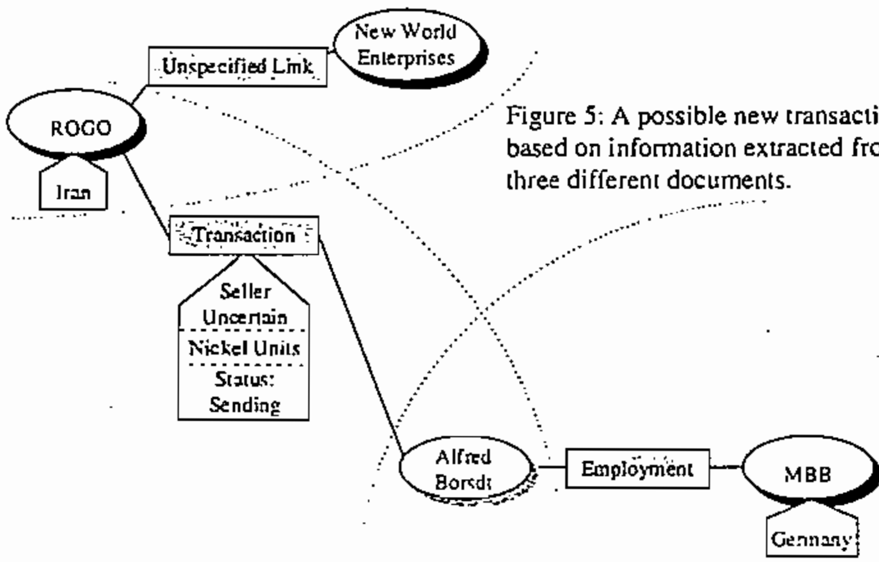


Figure 5: A possible new transaction, based on information extracted from three different documents.

With a bit of effort, any person can trace through the extractions shown in Figure 4 to find the correlation shown in Figure 5. The value of the automated process is that it can trace through extractions from millions of documents to find such links, and repeat the process whenever new information is added to the database.

From this example, we can see how the proposed approach offers a major breakthrough:

- The analyst in this example did not have to read hundreds of documents each day to discover the interesting link that these three documents provide. The three documents were chosen automatically because they were found to represent a link that seemed to match the analyst's prescribed interests.

- The intelligence process did not have to rely on one analyst, reading one set of documents--and a second or third analyst reading other documents--to discover that they each happened to have partial data on an important link. The corporate database searched out correlations across all the data.

- The discovery did not require the analyst to re-read all previously collected information that may have been surveyed by his or her predecessor. The correlation process compared new data to all the previous reporting.

- New information that did not get delivered to the analyst at this time does not go unused. Instead, it remains with all other information in the database, to help connect important bits of information that enter the database in the future.

The employee's responsibility for keeping track of new information would be much less time consuming in the proposed approach, because the pressure to read all documents would be removed. The employee can define information patterns that matter, and then depend largely on the automatic process to catch incoming data. A key responsibility would be to understand the basic kinds of information that machine processing can extract and correlate, and then describe the patterns that he or she would like to be alerted to.

The proposal here is for corporate services to extract and organize the most basic information elements from incoming documents, but a natural side effect could be local contributions of more detailed data elements. Local intelligence components probably would develop specialized machine processing capabilities to extract additional information from the centrally processed documents. They would wish to add more kinds of information to the corporate database, specific to their interests--so that more patterns they care about could be detected automatically. (This would be feasible because of local databases that contain speeialized information to aid in automated extraction--pertaining to national economies, weapons proliferation technologies, or key political figures, for example--or because the local component chose to do some expert data entry manually to augment the basic corporate processing.) Local components might opt to make such domain specific operations compatible with the corporate process because of simple self interest--i.e., because the central process delivers the bulk of the basic extraction and correlation capability required. The institution would gain from integrating such local operations into the corporate process.

The extraction process may be largely automatic, but humans would still be in the loop--operating downstream and just in time.

The approach outlined here entails automatic generation of database entries without much human intervention, as discussed in the appendix. The resultant entries often would be less reliable than data elements reviewed by a human being, but the overall process would manage error gracefully: Since a key purpose of the database is to do a much better job of selecting information for human review, the whole operation assumes there will, in fact, be human review for any information extractions that <u>really matter</u> to someone. The degree of perfection required from the initial machine proeessing can therefore be considerably lower than we are accustomed to thinking about for intelligence databases.

Human review could be a fairly simple process, in which the employee's train of thought remains focused primarily on substantive issues described in the text, rather than a separate administrative chore. If so, the burden imposed on employees would not be onerous. As an analyst checks over data that purportedly represents a noteworthy new development, he or she would validate or correct the data entries as a natural by-product of the work he or she wants to be doing.

Figure 6: Data entries "nominated by a machine will become "validated" when they matter to someone--as an employee reviews highlighted developments and either endorses or corrects the nominated entries.

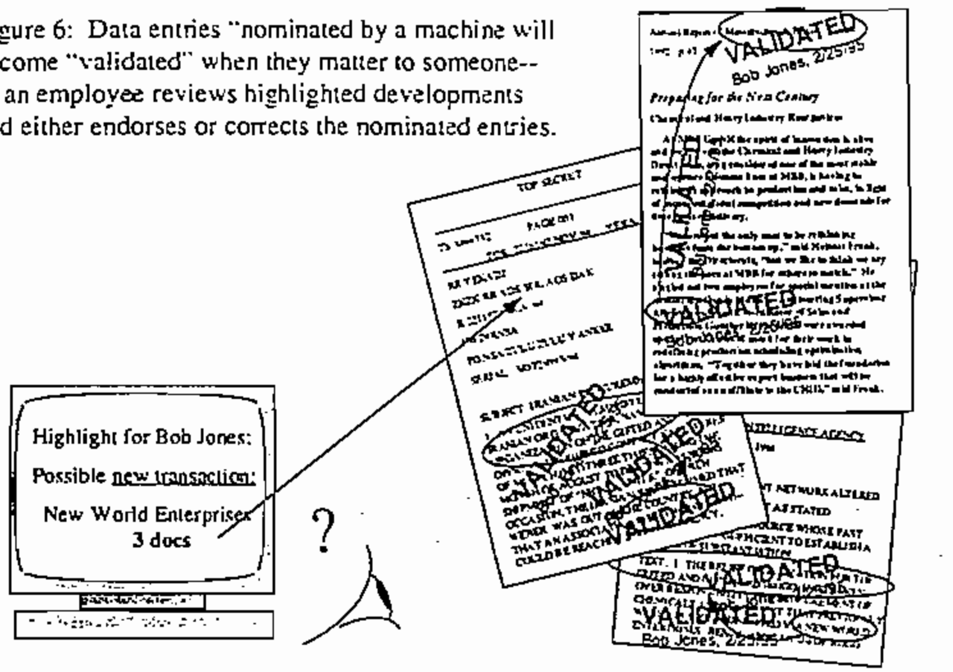


Figure 6 underscores the idea that the proposed approach does not ask highly skilled employees to spend time on extensive and "mindless" data entry and cleanup operations. Instead, the process captures the work that employees would naturally be doing as they explore the information that they have reason to believe is important. Designing the process to facilitate data validation--and designing the employee reward system to reinforce it--will be critical, for two reasons. Much of the automated extraction process will depend on direct access to information in the database, to help make sense of the text found in newly arriving documents (see appendix). Validating more of the corporate data will therefore improve the processing of incoming documents. The second reason for validating data entries is to create a trusted subset of the corporate database that can be relied upon in answering a question or preparing a briefing on short notice. In this kind of situation, productivity will be greatly enhanced if employees can directly query the trusted elements in the corporate database.

By adopting a corporate approach to managing incoming information, any conversion of data from "nominated" to "validated" entries will benefit all workers, whenever a single employee finds the information to be worth checking. The approach therefore represents a means for the institution to learn from its members. It offers a means to build corporate memory and share institutional knowledge. And, the benefits could extend to many employees who are not cleared for the most sensitive information, because the basic data elements would be inherently less revealing than the full-text source documents. (Employees could be alerted to potentially important developments because of linkages found in the corporate relational database, even if their passwords do not permit immediate access to some of the full-text source documents.)

Cultural Changes

Redesigning our process for tracking new information, away from a document-based approach and toward a corporate database approach, would not require extensive training to learn new skills. Most of the changes in work activities required of collectors and analysts could be designed to be intuitively obvious, and certainly the underlying thinking about information interpretation would not change greatly. Nevertheless, the shift would imply some significant cultural changes for intelligence:

- The notion of a corporate database implies connectivity and compatibility among organizational components that have traditionally maintained independent information systems. Issues involve not only hardware and software but also information formatting and naming conventions for the information elements (people, organizations, dates, places, etc.) that are to be captured in the corporate database.

- The idea of having employees do their routine work in and through a corporate database--so that the whole enterprise benefits and works more efficiently--forces the question of how to handle information that should be maintained locally. This implies partitioning of some local files while still doing formal work in the corporate realm--which requires a new way to think about one's corporate responsibilities.

- Employing people in an integrated corporate operation--as distinct from commissioning individual work that is captured in documents that are passed around--implies a significant shift of performance evaluation and rewards. To achieve the benefits of an intelligence "assembly line," as compared with today's artisans, we will need to be able to measure performance in new ways, recognizing and rewarding work that contributes to each corporate success.

- The corporate database would become important to all information workers, a key to maximizing each component's impact on the intelligence mission. There would need to be agreed methods for determining when and how to make changes in the corporate data structure and content--because components will be so dependent on a system that is not within their exclusive control.

- The idea of using a corporate database to build corporate memory and exploit corporate knowledge, as a natural byproduct of each person's routine work, means that the process cannot be individually supervised. It can be audited, and the work of selected employees (e.g., new hires) can certainly be reviewed, but the total number of database contributions involved would be huge. The fundamental approach would entail much less hierarchical control and more reliance on individual quality control, compared with document-based work.

Many years may be required to make such changes in an organization's culture. This suggests that a longterm plan with clearly stated objectives would be needed to help bring about a graceful shift from document-based operations to the approach proposed above.

Appendix

Making Data Entry Affordable--Automated Extraction of Information from Documents

Using Information Already in the Corporate Database to Enhance Document Processing

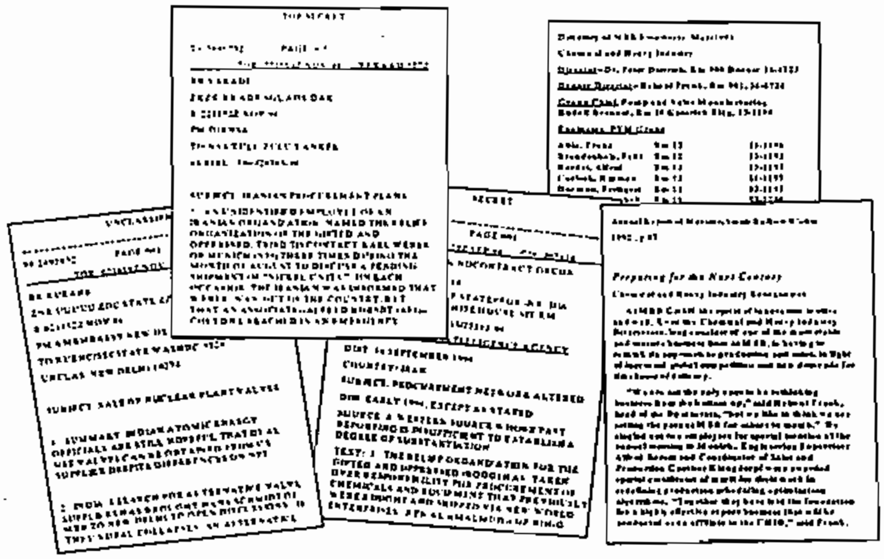
Text processing technology on the market today is capable of identifying many of the basic entities of interest in a document--including the names of people and organizations, for example. But commercial software cannot understand text well enough to recognize many of the important relationships among these entities. If we wish to extract critical information with high reliability, therefore, a human being is generally thought to be required to review the extractions and fill in information that relates one extracted entity to one another. This requirement is changing, for some applications.

Extracting information from incoming documents to fill a relational database can be done automatically, if the data and relationships to be extracted are not too complex, because validated information already in the database can be used to check and assist in the processing of new documents. If it can be done with sufficient accuracy, this processing will allow automatic correlation between what is new and what is known--including what is known about the information needs of each employee. The involvement of human experts can thus be shifted from the front end--where documents have a low concentration of critical information--to a point downstream, where employees want to be exploiting key information in the conduct of their daily work. And, because human reading speed would no longer be a factor in the up-front processing of incoming documents. information can be extracted and organized from many more documents than the work force would ever find time to read.*

The kinds of data and relationships that can be extracted include people, organizations, dates, locations--items that are of critical importance to work throughout the intelligence process. The following pages describe the database-aided extraction process, by examining a few mock intelligence documents and extracted information. The content is fictional, but the language is representative of text from which data must be extracted.

* Reading new documents consumes approximately two hours per day for typical all-source intelligence analysts--representing approximately one fourth of their activity. Even spending two hours per day, however, most analysts are not able to look at all the documents that are likely to have information bearing on their assigned areas of responsibility. And, of the documents that are seen, most include information that is relevant but not of immediate value. Therefore, the current approach to tracking incoming information leaves considerable room for productivity gains--if the analyst can be relieved of the task of looking at all the relevant new documents and instead can focus attention only on the small amount of data that represents important new information.

Figure A1: Simulated intelligence documents...



...and the kinds of data likely to be worth extracting:

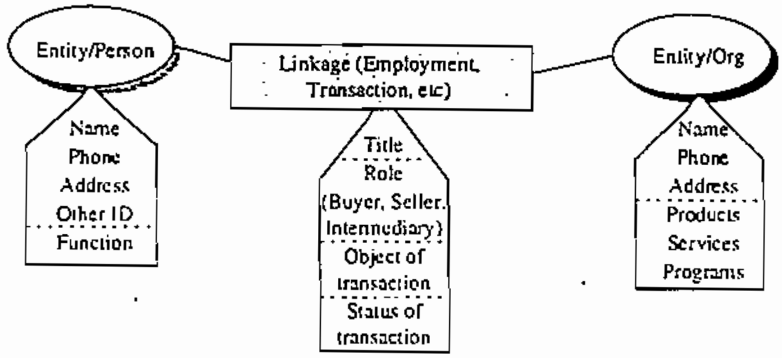


Figure A1 above presents several mock intelligence documents and shows the basic kinds of entities and relationships that intelligence officers might look for. As suggested by the figure, there are some important documents, such as company directories, that are very rich sources of basic factual information but that are unlikely ever to be read or retained in human memory. Such documents can greatly enrich a corporate relational database.

Figure A2 below shows some of the information that a person would be likely to extract from one document. It is easy to see how the entities and relationships can be inferred from the text in this document. A human reader can see that there is a transaction being considered, that it involves an outfit that goes by the initials MBB and which sells nuclear plant valves, and that the transaction is only in the discussion stage. It is also evident that the potential buyer is the Atomic Energy Commission and that the buyer's real objective is to acquire the valves it needs from the US rather than from MBB.

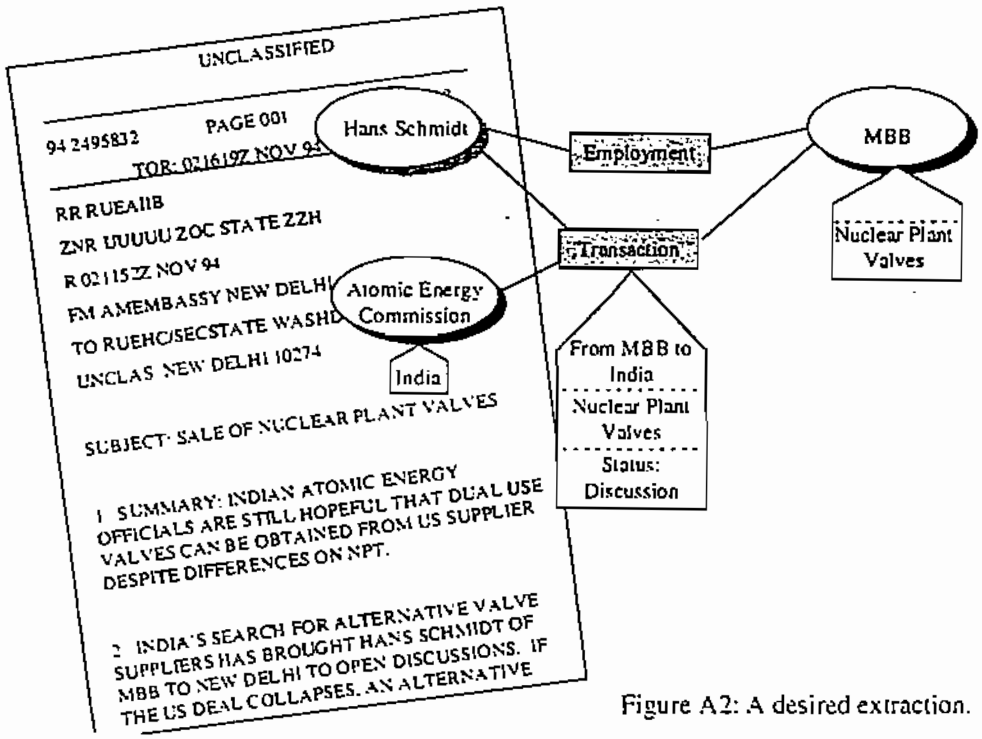


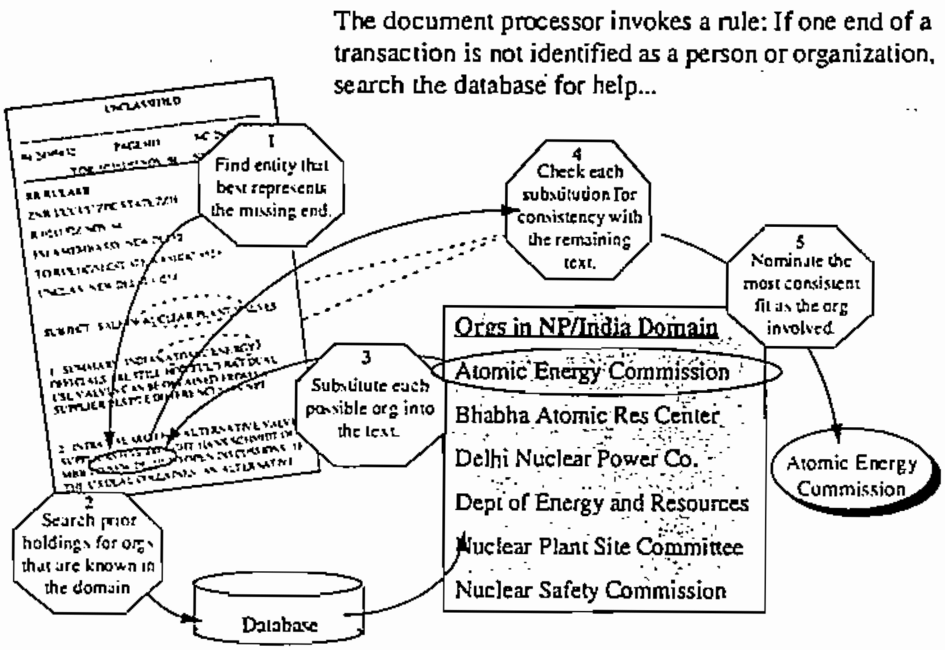
Figure A2: A desired extraction.

This example shows why automating the extraction of information is so important. A human being can get the gist of the transaction from the document in Figure A2 in a few seconds, and can then move on to the next document for review. But if the human must stop to capture this information in a relational database, the time required can be much greater--at least several minutes. The use of human beings to extract and enter such information into a relational database represents a pacing factor that would greatly restrict the number of documents that could be processed in a day. Hence the need for ways to extract at least some of the critical information from documents automatically.

Unfortunately, information which is so easy for a human to infer from the document in Figure A2 cannot be recognized reliably by today's commercial text processing software. The host organization (Atomic Energy Commission) is never named in the document; we simply infer it from the reference to "INDIAN ATOMIC ENERGY OFFICIALS." The fact that MBB is a vendor rather than a buyer cannot be automatically derived from the text; it requires a sophisticated understanding of the situation that is being described. Similarly, while a person can easily see that the sale of nuclear plant valves is merely under discussion, it is unlikely that today's commercial software can be made to recognize reliably the many different ways in which the English language could convey this fact.

We can do better than simple processing of the text in the document, however. Our options for automated extraction and entry of information can take into consideration not only the text in the document but also the wealth of knowledge that already resides in our files. Figure A3, for example, illustrates how an automated search of institutional files could help identify the Indian Atomic Energy Commission as the host organization alluded to in the document.

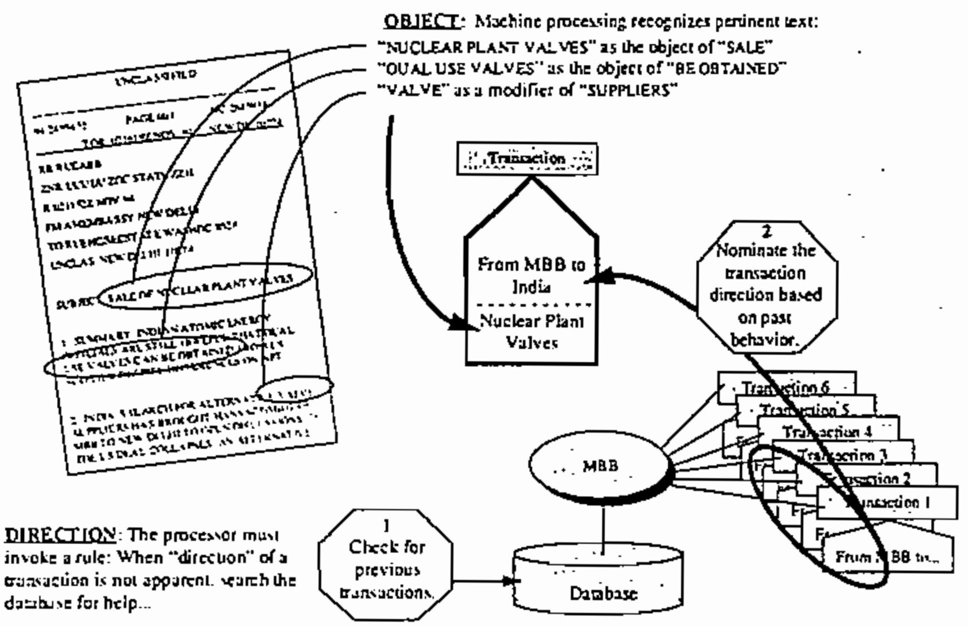
Figure A3: Recognizing an organization implied by the text.



The steps taken in Figure A3 will not always produce accurate information. For example, the host organization may actually be Delhi Nuclear Power. This illustrates why machine nominated entries should be tagged as such--so that each automated entry is distinguished from human validated information until verified. Also, the correct organization may not yet be listed in the database. This illustrates why the corporate database would improve as more data is entered over time.

An institutional database can also be used to clarify the nature of the transaction described in the example document. Commercial software is apt to recognize the object of the transaction as "NUCLEAR PLANT VALVES" or "DUAL USE VALVES" or just plain "VALVE." But which end is the buyer and which is the seller? A human reader can see immediately that India is searching for suppliers, and that MBB must therefore be a seller--but this reasoning is a bit much for today's text processing software to achieve quickly and reliably. In this case, if the institutional database holds previously reported information about transactions involving MBB, this information can be used to increase the accuracy of the document processor, as shown in Figure A4.

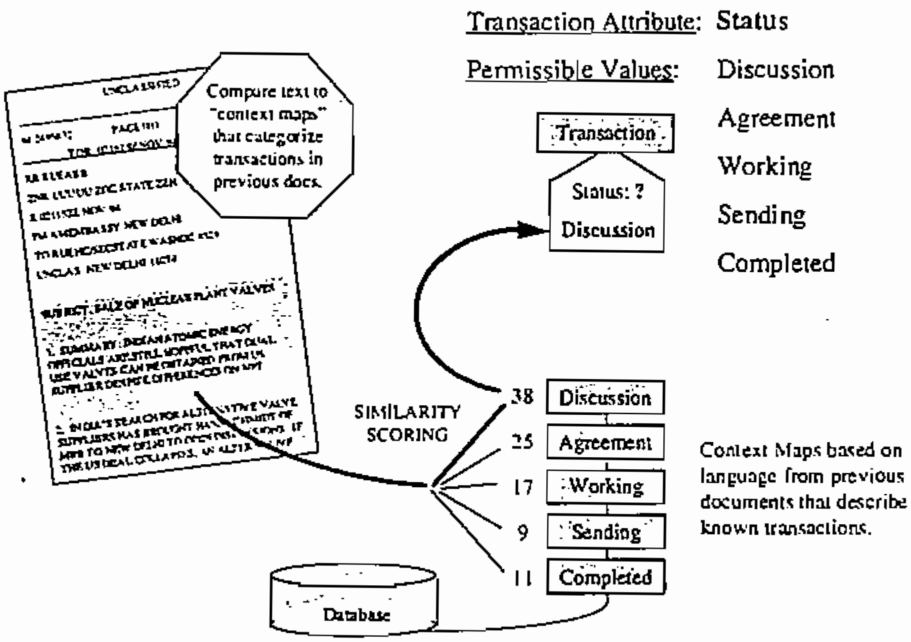
Figure A4: Determining the object and direction of a transaction.



Determining the direction of a transaction by examining past behavior will not always work, even if the files contain information on past behavior. We are apt to learn what works best only as we gain experience. The premise in this paper is that we will find many independent automated operations, which individually are simple and inconclusive, but which together yield highly successful database stuffing.

In some cases, we will need to characterize events or attributes in which even the central concept is beyond the capabilities of text processors to understand. In the example, the phase or "status" of a transaction is such a concept. In the real world, and in the English language that describes real transactions, the variety of eonditions that make up "status" is enormous and complex. Nevertheless, human readers could easily categorize most transactions as being in one of a few general phases (e.g., under discussion, in agreement, working, in delivery, or completed). When the goal is to put new information into one of several groups in this way, records of past documents can be a valuable guide. The language used in previous documents that describe transactions in known categories can be used to maintain a special file of representative language, or "context maps." Each new transaction can be compared to the language on file for each category--and therefore can be scored to see which category is most likely, as illustrated in Figure A5. Each kind of attribute will require some experimentation to determine the applicability and reliability of this technique, but we should find that this approach is highly successful for making some kinds of characterizations without any sophisticated automatic language understanding.

Figure A5: Nominating an attribute based on bulk language comparisons.



Summary

The examples above show three different ways in which an existing database can be used to enhance the performance of commercial extraction software. One application is to find a known entity, such as an organization, that the text refers to implicitly but does not identify by name. Another is to survey all that is known about an entity that is found in the text--such as its past role in transactions--so as to understand better what is described in the newly arrived document. The third application is to use the database to refer back to language in previous source documents. The original language can be used to maintain context files, as in Figure A5, or can be processed "on the fly" to obtain contextual information that is used once and then thrown away. As we gain experience, more ways to exploit the existing database for enhancing the processing of new documents are certain to be discovered.