USING UNSTRUCTURED DATA AS AN ANNOTATION OF STRUCTURED TO IMPROVE RETRIEVAL, INTERPRETATION AND REUSE

The ability to communicate efficiently and effectively across the extended enterprise has led to great advances in the way we conduct business. Information can be shared almost instantaneously with colleagues in distant locations and can be compiled and collated from a wide variety of sources to provide a detailed analysis of operations, markets and opportunities. Approaches such as Business Intelligence and Data Mining allow us to monitor the performance of organisations and infer fresh insight from stores of operational data and information. Despite these advances, however, there is a fundamental flaw in the way we store and retrieve information; one which is preventing organisations from gaining as much value as they could from the data they hold.

INTRODUCTION

The problem has arisen with the emergence of two separate disciplines to handle the different types of data and information that organisations need to collect. The first focuses on structured information which, as the name implies, provides defined content for specific data fields. Such data can be recognised by systems, interpreted and acted upon. Structured information tends to support high volume transactions (credit card payments for example) and normally refers to the state of things, such as prices, locations and identities.

Unstructured information, on the other hand, tends to support communication between people. This information explains and provides a context for the patterns between and across structured information.

These two are currently treated completely separately within our IS portfolios, despite the dependencies between them, where the unstructured provides an insight into the structured information, which in turn contextualises the unstructured.

The alignment of the mechanisms used to document and exchange structured and unstructured information would provide both richer messages and a more enduring record. Importantly, it would also enable 'human' knowledge and experience to be encapsulated alongside the factual content.

It is an area which still poses a significant challenge to IS, however.

The contention of this paper is that by considering the unstructured or more explanatory information as an annotation of the structured, we can begin to build up a set of exchanged messages and historical records that define not only how a business activity was conducted, but that also incorporate key stakeholders' views and insights relating to that activity. So identification and retrieval would be easier and different viewpoints could be incorporated in layers of annotation, superimposed on to the operational information, and bringing together insight from different areas of the business.

THE DISLOCATION BETWEEN STRUCTURED AND UNSTRUCTURED INFORMATION

To facilitate analysis, information needs to be treated in a variety of forms. In many cases this involves taking information out of its original context, interpreting it and in many cases transforming it into a form suitable for subsequent interrogation. To achieve this, the structure of the underpinning data (the elements which, when combined, form the information entity) must be computationally explicit so it can be interpreted according to pre-defined rules.

Structured information is best suited to the definition and communication of statements of fact, which are required by other people to complete a task or make a decision. These would include the identity and account number of a customer who placed an order, the date of an event, or the value of a set of shares.

Much of the information we communicate these days, however, is accompanied by human 'inference', via reports, presentations, emails, diagrams, charts and, increasingly, in forms such as video clips and webcasts. These methods have much less structure. Computers lack the ability to interpret information in such a form, not only in terms of identifying the salient detail but also through an inability to understand metaphor or analogy, or to apply context given situational clues. We rely upon these to interpret and act upon such information; to provide an implicit structure which is not defined or directly articulated

Highly structured approaches are less suited to articulating reasons for, or ramifications of, a given state or situation. Let us consider an example frequently encountered in supply chain operations. The warehouse receives repeated orders for the same replacement part, from the same area. There is a structured element to this process (take order, pick, pack and transport) and this structured data can be handled within the supply chain transactional system. What is omitted, however, is the communication of the reason behind the repeated failure of the part and the need to constantly replace it. This information is often provided by the maintenance engineer, but in an unstructured format - perhaps by email or in a telephone call. The individuals who provide replacement parts may ultimately notice an unusual level of demand, but are powerless to rectify the underlying issue unless the reason for the high demand is correlated to the items and activities in question. In other words, a richer narrative is required to supplement the structured data. Such unstructured information relies upon human interpretation to convey its message; the computer lacks the ability to assess or act upon the *meaning* of the data.

THE CONTEXT OF INFORMATION

Much unstructured information is transitory. It supports a perspective at a given point in time and space, where the broader context is absent or weakly defined. An email sent between colleagues will include only cursory contextual detail, unlikely to provide clear insight to anyone reading the content at a later date. A document written to summarise a piece of work may include greater contextual detail, but only within the confines of the immediate programme of work. In other words, the reader will need to be familiar with the subject to be able to fully interpret implicit assumptions and dependencies. He or she will probably also need to know the context of the programme of work within which it was produced.

This leads into the issue of ‘assumed knowledge’. Where a document is produced in the course of an event, it will be interpreted in the context of this event, with many assumptions made as to the reader’s understanding of context and prior knowledge. Subsequent readers do not share this knowledge, so will have difficulty in interpreting it, and assessing its relevance to a new situation.

The dilemma is that while unstructured information represents a potentially richer resource, it lacks computational interpretability and an associated context. Structured information, conversely, suffers from the very rigidity which gives it clear interpretability and context, but which constrains the freedom of expression necessary to articulate abstract concepts.

ALIGNING STRUCTURED AND UNSTRUCTURED INFORMATION

An apparently simple solution would be to align the two: the context of the structured information would clearly locate the associated unstructured information; reciprocally, the unstructured information would provide a narrative through which the ramifications of the structured could be conveyed.

Many instances of this can be seen. Google, for example, provides a highly structured view of the physical environment, with photos at satellite and ground level associated with a structured mapping of the world. The view is enhanced with user commentary, such as reviews or descriptions of restaurants within the map location or with further photographs taken at various times and during various events. Note the interplay between structured and unstructured information: the structured shows the relative position of specific entities - in this example, how close a given restaurant is to car parks, main roads and public transport facilities; the unstructured annotates this to provide more explanatory, judgment-based narration along with a pictorial view of the immediate environment. In other words, this aligned view will help you to decide if there is a restaurant that is both accessible and worth the trip!

This utopic state of alignment is difficult to achieve in many organisations, however, for a number of reasons.

The main block comes from the use of different systems, polices and approaches to the management of data in all its different forms: Enterprise Resource Planning (ERP) systems and Product Data Management (PDM) applications are used to compile and publish views on structured data; the supporting, unstructured data (the reports and presentations that arrange such views into a narrative structure) are published in separate document or content management systems. The latter will arrange these documents or content by whichever categorisation has been selected –by business function, project or customer, for example. Since some of this content will be meaningful only to those involved in its creation and immediate consumption, it becomes increasingly difficult for any one user to collate and derive insight from all unstructured ‘narrative’ corresponding to a given set of structured information.

So what can be done?

USING UNSTRUCTURED DATA AS AN ANNOTATION OF STRUCTURED

Unstructured information should be treated as an annotation of the structured, just as when we draw a graph, we collate and annotate the relevant underlying structured data entities to illustrate a concept; or when we compile a report, we arrange a collection of observations and facts into a comprehensible structure and punctuate it with commentary.

Unfortunately, at present, the separation between systems that define ‘fact’ and those that provide ‘annotation’ prevents the annotation from serving in so literal a role. As such, many of the ‘facts’ described in a report are themselves abstractions, with no direct computational linkage to the underpinning structured data. This introduces problems of provenance and of reuse: we cannot establish the veracity of the structured information used to inform a position (ie. the 'content' of the document) and we cannot use the document to retrieve the underlying data for re-examination and further analysis.

This total 'decoupling' of the two sets of systems has implications for immediate operational performance and for the subsequent reuse of the operational information and related narrative - for example, to redesign a product or to establish successful marketing campaign strategies.

In terms of operational use, formal message exchanges allow structured communications to take place in a highly automated and controlled manner. Analysis of this information may highlight issues or areas for concern – provided that there is enough content to allow this insight. Generally, however, further investigation is required to identify the underlying reasons and to propose reparative measures.

In many cases, the individuals directly involved in the operations will have an opinion on what is causing the problems and what is required to address them. Such information is vital and tools such as 'Lessons Learned' databases are intended to collate and share this kind of experience and insight. The problem is that these are retrospective measures that do not capture great detail. They are also normally stored in document management systems, with few if any linkages to the operational data which could 'locate' these observations against specific areas of operations.

Rather than looking back, it is preferable to build up annotations on an ongoing basis. These experiences should be transmitted as part of daily working practice, so that the information may be acted upon. This has the dual benefit of ensuring the quality of the information and motivating its exchange.

So, all users should have the capability to give their perspective on structured information at the point of need, namely, when they want to articulate a concept to others within the enterprise. The recipients will then be in a much better-informed position to address the issues. What is key is to be able to compile supporting evidence from a variety of sources and structures. Where statements are decoupled from the actual evidence, the receiver will not know if the inference expressed by his colleague in another part of the organisation is one that he can trust.

The real value in an organisation lies with its people. It is very much a question of finding the best way to communicate the valuable information and experiences held by these people, by attaching it to the structured evidence.

And since information tends to come in tandem with an individual's perspective on the situation, different stakeholders must be allowed to provide their own annotations. So, for example, a warehouse operator may consider an inadequate storage facility to be the cause of breakages, whereas the 'breakages' issue may actually be due to an inadequate packaging of the products. Taking this forward, the warehouse operator may choose to compile and annotate views of different warehouse locations against breakages, whereas an area manager might compile information describing wastage at a given site, compared to a regional average. Thus, an annotation may refer to any given set of structured information, at any given level of granularity.

It is this that sets the idea of annotation apart from Metadata Management (MDM, although annotation may be contained in such metadata. MDM considers that metadata 'is about' a given entity, whereas annotation refers to the addition of some form of narrative above a freely-compiled set of elements, with the set existing purely for the purposes of that single annotation. A given entity can be included in any given annotation.

ADOPTING ANNOTATION AS AN APPROACH

To adopt annotation rigorously means linking information from disparate sources. The ability to collate a set of observations, assign a narrative to explain some feature or pattern of interest amongst these observations, and share that with a broader audience allows for the expertise and insight of each member of the enterprise to be built up to form an 'overlay' within which operational data may be interpreted. Technically, this is achievable: many of the mechanisms for accomplishing this already exist. Hyper-linking, metadata and mark-up are commonplace and all of these can be used for the annotation of structured data.

Effort is still required, however, to provide an IS environment within which to accomplish this, as well as an operating framework to ensure that such an approach supports working practice.

There is a governance challenge too. The annotation approach requires a 'single version of the truth' and it is not easy to achieve this in most organisations. Imagine, for example, an entity such as MoD where annotation could involve access to everything from a supply chit to top-secret operational information and where much information is held in siloed repositories, as both security and operational concerns may dictate. Access to trusted sources of 'evidence' is a key stage in allowing users to usefully build annotation around such structured information

CONCLUSION

Annotation links together a series of disconnected facts and weaves a 'story' around them. It is often this ability to pull together different facts that enable an organisation to make sense of discrete information or situations.

And the benefits go beyond providing a richer set of information to support current operations. As these annotations embed elements of the knowledge and understanding of stakeholders, they enable this knowledge to be stored against the structured information entities which define the organisation's activities. In other words, the knowledge is located against structured information and arranged according to a given task or activity, thus facilitating its retrieval, interpretation within context, and ultimately its reuse across the organisation. This can occur over extended lifecycles, which is vital in a world where service contracts may run for many decades, and in many cases beyond the careers of those initiating a project.

In simple terms, the annotation of structured data with unstructured information allows the organisation to encapsulate a moment in time, providing access to the background position or view of the world which existed when a specific statement was made or an item of information collected.

Glue Reply is UK’s leading consulting services organisation focused exclusively on optimising IT/Business alignment and minimising the cost of business and IT technology change. Our core proposition is to help organisations maximise the value from their change and technology investments by helping them define, design, implement and resource best practice:

– Enterprise architecture and business/technology change management
processes, roadmaps and competencies;
– Business design and process management initiatives;
– SOA, integration and data management platforms.

Glue Reply UK
www.replyltd.co.uk