DITA for Enterprise Business Documents
Sub-committee Proposal Background

Why an Enterprise Business Documents Sub-committee

Documents initiate and record business change. It is easy to map some business documents, for example a purchase order, into a data structure that can be integrated with computer software, such as an accounting system. There are other business documents that do not share the organized structure of a purchase order, and while these documents are more difficult to integrate with computer systems, the potential benefits of this integration are still very high. These less-structured documents of widely varying lengths are found across the enterprise, and are referred to with the term *narrative business documents*.

Narrative business documents gained attention in the 1990's as organizations came to understand the need to manage this type of "unstructured content" in a controlled manner. Today, the increase in tightly integrated systems and automated business processes has once again brought attention to these documents. A large number of organizations now see the unstructured content that exists in narrative business documents as standing in the way of processes that could be automated end-to-end. The lack of structure leads to inconsistency, poor readability, and the inability to reuse content. These same documents contain a significant amount of the organizations' intellectual property, which because of the inherent lack of structure, remains hidden from business intelligence and other software tools.

Because of this, organizations are no longer satisfied to just *manage* unstructured content; they now want to *structure* it. In fact, it may be that the term unstructured content is a misnomer that served its purpose for a decade or so. We would suggest that a more modern view of content would be that there is some content that we have successfully structured, and other content that we do not know how to structure yet.

Narrative business documents are an example of one type of content that has proved more difficult to structure. Early attempts to apply XML to these documents used a variety of document definitions and approaches to XML editing. From a business perspective, lessons were learned that pointed out the need to provide more natural word processing and collaborative experiences. On the technology side, information strategists needed a DTD or schema that represented a standard, was extensible, and adhered to principles of object orientation that govern most enterprises.

In the past year, a growing number of organizations have come to believe that DITA not only provides the best basis from which to start addressing their technical requirements for narrative business documents, but that characteristics of DITA simplify the usability issues as well. DITA appears to work so well, that the absence of a sub-committee focus on narrative business documents has not stopped several organizations from embarking on the use of DITA for narrative business documents. This appears to be an ideal time for the DITA technical committee to address the issues and provide guidance concerning the use of DITA for this class of documents.

Enterprise Business Document Sub-committee Goals

Goal One: To develop and recommend an enterprise business document meta-model with sufficient detail to:

   a. define the scope of documents addressed by this subcommittee, and

   b. allow these documents to be discussed in a non-ambiguous way.

Rationale

When an organization decides to implement DITA for narrative business documents, and begins to compare these documents with the DITA schema, it becomes clear that there are some structural issues to resolve:

- Narrative business documents are generally authored and presented as contiguous sections of content that are larger than the normal DITA Topic, leading to questions as to how topics should be aggregated into a document that can be validated against a DITA DTD.

- Business documents are made up of sections and sub-sections that are roughly analogous to DITA Topics, but the topic segmentation is not as clear as it is for technical documents

- Business documents contain hierarchical relationships between sections and in-line content that can be difficult to harmonize with the DITA model for sections and recursive information types.

In addition to these structural issues, there are also requirements concerning specific element types. For example, there may be business requirements for authors to view and manipulate the numbering of sections or lists in-line. Even as this list grows, the types of issues found are all well within the scope of what can be resolved through specialization or the minor additions to the DITA standard that might be expected to be suggested by any technical sub-committee.

What makes the task of arriving at recommendations for narrative business documents somewhat different from that faced by other DITA sub-committees, is the lack of a vocabulary that adequately describes these documents. As pervasive as business documents are, words to describe some of the structures found in these documents simply do not exist. As a result, business analysts and end-users who try to quantify the requirements associated with these documents resort to illustrating specific instances or examples of document structures that they feel are problematic—without being able to provide terminology for the generalized case.[1] A few examples will highlight not only some of the document structures that need to be addressed by a sub-committee, but the lack of adequate terminology to support effective discussion:

- When attempting to discuss the differing types of section structure that must be harmonized with the DITA standard, terminology immediately becomes an issue. When a section of a document contains both sub-sections and in-line content interspersed with each other, what is that called? If a more formal

---

[1] A good example of this is the Elkera Comparison of XML Schema for Narrative Documents, which describes certain structures that are found in narrative business documents and compares how various schemas would handle these structures.

outline is followed and sections contain only sub-sections without in-line content between them, what is that called?

- Referring to what *might* be a simple topic specialization, narrative business documents include blocks of text that must be kept together for contextual reasons. If an author wants to indicate that three paragraphs within a section form a single idea, and that any reuse of one paragraph must also reuse the other two paragraphs, what is that grouping called? Are there other similar groupings that should be included in the analysis, and what would they be called?

- Sections are often recursive in business documents, with no semantic differences based upon the position of the section in the hierarchy. However other documents apply specific semantics to sections at each level of a hierarchy. What is this type of structural semantics called?

The lack of an adequate vocabulary to describe narrative business documents will make it difficult to progress in efforts to approach these documents with the DITA standard. Such a vocabulary is normally part of a model that identifies, in the most general sense, the characteristics of a business document. This type of model is called a meta-model, and is described for readers who might be less familiar with the role such a model plays.

The meaning of a document meta-model

Most people understand that an XML schema is a definition that states how a document will be structured, the types of content a document may (or must) contain, and the metadata that may (or must) be used to describe the content in the document. The physical representation of this contract is one or more .xsd files, which contain technical notation that defines the rules of the contract.

Perhaps a little less obvious is that an XML schema also represents an abstract model of a particular type of document, and that this abstract model had to exist in someone's mind before the schema was created. It is a model because it defines how to create new instances of a document type. It is an abstraction because it is not based on any particular instance of a document type—instead it is abstracted from the sum of all known instances of a document type and their particular variations.

This particular type of abstract model is called a meta-model. A meta-model attempts to describe the component parts of something, and the meaning each component part has for a particular purpose.[2] So when applied to narrative business documents, a meta-model would first describe the types of components that occur (simple examples are a *title* or *caption*), and the meaning each component conveys to the reader of the document.

All this information about meta-models is important, because in order to create rules that express the needs of narrative business document authors, it is essential to describe how these documents differ from other document types (especially in the case of this sub-committee, technical documents).

---

[2] While not technically accurate, this definition of meta-model is designed to supply the essence of the term for business readers without detracting from the primary purpose of this document. This definition is derived from a number of sources, including www.metamodel.com.

Doesn't this type of meta-model already exist?

It would be very helpful if some field of science or industry had already defined a meta-model for narrative business documents that business analysts could use to examine the meaning of the different structures that exist in these documents. Linguistics immediately comes to mind since it is the science of language. However, linguistics is primarily concerned with the spoken word. When linguists do study the written word, it is normally at the sentence, or sub-sentence level.[3] For this reason the methods that linguists use to look at language might be of general interest to business analysts, but few if any direct writings of linguists address the structural semantics in documents that are so important to understanding how to apply DITA to these documents.

Moving from science to industry, typography and related publishing activities come the closest to addressing documents in a way that is useful for business analysts. There is a publishing document meta-model, with a vocabulary that all of us are familiar with to varying degrees. Common terms such as *titles, captions*, and *sections* are derived from this model; as well as less common terms such as *widows* and *orphans.* Since a meta-model describes something with a particular purpose in mind, the typographical meta-model, which is designed to identify how components are arranged on a page, does not fully meet the needs of business document analysts. However, it provides a starting point that is relevant to the narrative business document discussion.

The goal of the sub-committee would be to start with the useful components of the typographical document model and expand these to create a light-weight narrative business document model that is sufficient to support the remaining sub-committee activities.

Goal Two: To develop and recommend an approach to harmonizing the narrative business document meta-model with the DITA standard.

Rationale

The term *harmonizing* is used carefully here to suggest that there may be some give-and-take on both sides to bring the benefits of DITA to business documents. The goal is to implement DITA for narrative business documents with as little business disruption as possible, while suggesting as few changes as possible to the standard itself. Ideally, the narrative business document meta-model can be implemented without requesting changes to the schema, but to launch the effort with this requirement would seem to be self-defeating.

We anticipate that the sub-committee will focus primarily on structural specializations, with some domain specializations that relate to the meta-model itself. We do not expect to address the domain specializations that may be required for narrative business documents in a specific industry. For example, we would view *task* as a domain specialization for the technical document meta-model, while *assembly* and *disassembly* might be further specializations of task for a particular industry. It is our intention to focus at the task level of granularity.

---

[3] In 2002 Richard Power, Donna Scott, and Nadjet Bouaya-Agha wrote the report, <u>Document Structure</u>, for *Computation Linguistics*. They stated, "While there is a long tradition and rich linguistic framework for describing and representing speech prosody, the same is not true for text layout." It is one of the few writings from the field of linguistics that addresses narrative document structural models in some detail, and does so from the viewpoint of Natural Language Generation (NLG).

Goal Three: To develop and recommend a standard approach for fully expanding DITA Map references into an editable process instance that may be validated against an approved DITA DTD or schema.

Rationale

This may not be needed as a separate goal, since it could be viewed as part of the previous goal. It is listed separately for three reasons.

- There are a number of projects underway that involve the aggregation of multiple topics into a single document for authoring, as this approach has appeal for subject matter experts of certain document types. Various approaches have been taken to create this document instance, often by creating a hybrid schema from DITA Map and Topic. These hybrid schemas, while similar to DITA, do not conform to the standard.

- It is possible to create a valid DITA document that serves the purpose of representing a runtime version of a DITA Map, both with and without specializing any existing DITA elements. There may be shortcomings to this approach that suggest some extensions to the DITA standard are needed, but it is important to examine the issue in detail before assuming this.

- While it will take a fair amount of time to address all of the issues associated with narrative business documents, starting at the document, or highest level of the hierarchical structure, might make sense. It remains to be seen whether the sub-committee could quickly come to consensus on how to approach editable topic aggregation without having fully described the underlying document meta-model.

Goal Four: Long-term to develop and recommend guidance for organizations that intend to adopt DITA for enterprise business documents.

Rationale

Regardless of the type of documents being authored, the ability to think about content as components requires that authors modify the way they think about writing and that organizations modify the business processes they use to publish content. When compared to technical writers, authors of narrative business documents will find it more difficult to adjust to these changes without disruption.

Change is often difficult, and unless there is significant business process improvement to be gained, organizations tend to oppose change. The approach to content creation and management that DITA promotes has the kind of clear benefits that motivate people to change. The sub-committee can play a role in enabling change by not just suggesting that change be done in phases, but by actually providing best-practice based guidance on the phases and approaches that might be considered.

Narrative Business Document Sub-committee Deliverables

Proposed initial work products for the subcommittee include the following:

1: Recommended approach for fully expanding DITA Map references into a valid editable process instance.

2: Recommended baseline enterprise business document meta-model.

3. Recommended harmonization of enterprise business document meta-model with the DITA Standard.

4. Long-term recommend guidance for implementing DITA for enterprise business documents.

For more information please contact:

Ann Rockley
rockley@rockley.com

Michael Boses
mboses@invisionresearch.com