Finding & Fixing Document Differences in Drug Product Labeling

Part II—Comparing documents that are the same, but different

August 2011

Editor's note: This is a companion document to *Finding and Fixing Document Differences in Drug Product Labeling (June 2011)*. That document should be read first.

Like its companion, this document assumes basic familiarity with i4i's A4L technology and products. Details can be found at i4i.com.

Background to the problem

Comparing two documents to identify differences is a core activity in any document production environment.

Compare tools highlight content differences, inserts and ~~deletes~~, between two or more documents. This is used to support collaboration and review: see what has changed in the content (is different in the context of the original), intentional or otherwise, and easily process these differences.

Compare is a fundamental tool in an organization's *information agreement*[1] arsenal. In the world of drug product labeling, it is needed to reconcile the many versions and variants of a document to ensure information agreement.

Compare assumes that the documents being compared are closely related. Its objective is to uncover differences between content objects that have the strong possibility of being the same (i.e., between document versions or variants). The more the documents are the same, the more useful compare is in exposing differences. The presence of many differences creates "noise", which the user finds hard to process.

There is a case, however, where compare fails completely despite the fact that the documents being compared are closely related. In fact, they are supposed to be the same, but are not: this is the case of translated documents.

Compare also fails as a tool when the content is restated. That is, the content is materially the same but said in a different way: for example, warnings for the practitioner as opposed to warnings for the consumer.

This paper discusses i4i's approach to helping users deal with this problem. Like all of i4i's solutions, it is predicated on the use of XML in documents.

The core idea

At the core of the solution is the fact that every piece of content in a document expresses an idea or *concept*. Some concepts are high level (e.g., warnings or indications). Other concepts are specific (e.g., arrhythmia, neonatal, or epiphyseal injury).

It is assumed that two documents that may be the same share the same concepts. In fact, it could be argued that the starting point for uncovering differences in documents should be uncovering differences in the concepts discussed in the documents.

If the concepts are the same, that tells the user that at the conceptual level the documents are the same—and what will be uncovered by a traditional compare will be differences in the articulation of the concepts.

---

1 Information agreement: where what is being said in one document is the same, or materially the same, as in another document—and, if not, the reason why not is captured. See the white paper.

If the concepts are *not* the same, that tells the user that there are fundamental differences between the documents that need to be addressed before dealing with the specifics.

When comparing documents that are the same but different (i.e., translations), i4i's solution is to compare the concepts in the documents. Differences in the concepts are an indication that, at that level, the documents are not the same and further review is required. Concepts are identified using XML.

Why this works

Consider a drug that is indicated for acne and rash, and may cause nausea and/or restlessness in pregnant women.

There are three concepts here: indications, side effects, and populations. Each concept has a concrete form or *term*: acne and rash for indications, nausea and restlessness for side effects, and pregnant women for populations.

A label for this product must identify the concepts and their concrete form. It might say:

> **This product is indicated for acne and rash. It may cause nausea and/or restlessness in pregnant women.**

A variant of this label, for a jurisdiction that does not recognize rash as an indication, would say:

> **This product is indicated for acne. It may cause nausea and/or restlessness in pregnant women.**

A compare of the two identifies the difference as:

> **This product is indicated for acne ~~and rash~~. It may cause nausea and/or restlessness in pregnant women.**

But, if the jurisdiction's language was Spanish, the label would say:

> **Este producto está indicado para el acné. Puede causar náusea y / o inquietud en las mujeres embarazadas.**

A compare of the English and Spanish content results in:

> ~~This product is indicated for acne and rash. It may cause nausea and/or restlessness in pregnant women.~~
>
> Este producto está indicado para el acné. Puede causar náusea y / o inquietud en las mujeres embarazadas.

This is not, from the point of view of understanding if there are meaningful changes, a helpful result. Applying XML tags to the content, to unambiguously identify the concepts and their concrete forms, results in:

> **<para>This product is indicated for <indication>acne</indication> and <indication> rash</indication>. It may cause <side effect>nausea</side effect>and/or <side effect>restlessness</side effect> in <population>pregnant women</population>.</para>**

Sending this for translation[2] to Spanish results in:

> **<para>Este producto está indicado para el <indication>acné</indication> y el <indication>sarpullido </indication>. Puede causar <side effect>náusea</side effect> y / o <side effect>inquietud</side effect> en las <population>mujeres embarazadas</population>.</para>**

The local user changes this to satisfy the local regulatory regime which does not recognize rash as an indication. The result is:

> **<para>Este producto está indicado para el <indication>acné</indication>. Puede causar <side effect>náusea</side effect> y / o <side effect>inquietud</side effect> en las <population>mujeres embarazadas</population>.</para>**

A compare of the XML in the English and Spanish documents results in:

> **<para><indication></indication>~~<indication></indication>~~<side effect></side effect><side effect></side effect><population> </population></para>**

This is a meaningful result. It unambiguously informs that an indication, specifically the second indication, has been removed.

i4i's A4L has a solution

A4L's authoring tool lets the user apply rich XML as shown above. This rich XML is used to identify concepts, manage the structure of the document, ensure that it can be repurposed, and make it fully computer-processable.

The A4L web services provide specialized *XML-compare* services that allow a comparison of *just content* or *just XML*. In some instances, the business case for rich XML cannot be made. The documents are not intended for repurposing, their structure is so simple that XML structure management is unnecessary overhead, and XML output is not required.

Despite this, the need remains for comparing documents that are the same, but different.

2 Translation systems do not translate the XML tags.

A4L authoring tool

Traditional XML is the rich XML described above. A4L can be configured to support this—for example, the "A4L for SPL configuration" or the "A4L for European labeling" configuration.

The A4L authoring tool can be configured to provide support for only *sparse markup*. *Sparse markup* is XML markup that identifies concepts. Sparse markup is not used to enforce structure or traditional repurposing. Sparse markup can be placed at any location in the document.

In the context of A4L's *information agreement* solution, sparse markup is the application of *concept* tags, added as needed to the document. There can be many different types of concept tags.

In the above examples, there are indication concepts, side effect concepts, and population concepts, each term of which would be captured by a concept tag. A sparse markup configuration of A4 authoring allows a user to create a Word document as they normally would.

A4L's sparse markup functionality is brought into play only when the user wants to identify content that instantiates core concepts and must be migrated to versions and variants of the document.

The above example, with sparse markup for the concepts, would be:

> **This product is indicated for <indication>acne</indication> and <indication> rash</indication>. It may cause <side effect>nausea</side effect>and/or <side effect>restlessness</side effect> in <population>pregnant women</population>.**

A4L's XML-compare

The A4L web service for *XML-compare* is run to determine whether a variant that should be the same as its base version is actually the same. It compares the sparse markup of the base with the sparse markup in the variant, and returns a report as shown:

XML Compare: yyyy-mm-dd hh:mm		
Base document: docID, title, version	Change document: docID, title, version	Result
<indication></indication>	<indication></indication>	
<indication></indication>		~~<indication></indication>~~
<side effect></side effect>	<side effect></side effect>	
<side effect></side effect>	<side effect></side effect>	
<population></population>	<population></population>	

Close

This informs the user that an indication term, specifically the second indication term, has been removed from the variant document. A review is required.

Summary

Document compare has traditionally been limited to comparing content streams. This model assumes that the compared documents could be identical. In production environments, there are cases when in fact the documents are identical but are not, the best example being translated documents. Traditional compares are not appropriate for this common scenario.

Appropriate use of XML allows a user to mark up a document in a way that captures its core concepts in XML. Identifying differences in documents that are the same, but not, is done using specialized XML tools that compare XML tags.

This provides a richer analysis of document differences that is not reliant on the details of language.

Infrastructures for Information Inc. (i4i)

116 Spadina Ave. 5th Floor

Toronto, Ontario, Canada M5V 2K6

www.i4i.com