i4i

Beyond Compare

A look at concept analytics in the context of drug product information

i4i

January 2015

Compare: is that all there is?

A look at concept analytics in the context of drug product information.

How does one compare two documents that should have a degree of material similarity? It depends on what the user wants the compare to uncover.

If the user is trying to uncover topographical differences such as changes in word order, the addition or deletion of words or characters, changes in formatting (i.e. normal to italic), or changes in punctuation, then the "track changes" model is a good one. It identifies, against the base document, differences in characters, words, formatting, white space, etc. found in the target document and highlights them as inserts or ~~deletes,~~ or sometimes as moves.

Thus, comparing two versions of a document makes good sense; there will be differences between V1.2 and V1.3 but one does not expect V1.3 to be a total rewrite of V1.2. A few words or even paragraphs may have been added or deleted, and some formatting changed. The "track changes" will highlight, for the most part, incidences of change within a continuity of no change. This paradigm works because of what we are looking for: minor physical differences in what should otherwise be identical documents.

It is true that uncovering word changes can have greater than topographical implications; a word can be an instantiation of an important concept. Its insertion or ~~deletion~~ can change the meaning of the document. This discovery is a beneficial but undesigned product of the "track changes" function, which does not and cannot weigh changes—that is, what is important and what is not. A coma is undifferentiated from nausea.

What if the user is not trying to uncover topographical differences? Uncovering topographical differences between an SmPC, a PL, and SPL is futile as these are expected to be topographically different. So if the user is not trying to uncover these, which by definition exist, what is the user trying to uncover? The user is trying to uncover conceptual similarities and differences—are the concepts presented in the documents the same or different?

Since the track changes model of compare cannot uncover conceptual similarities or differences, the user has to do it manually. The first step is for the user to create a working data set upon which to base the analysis. To do this, the user reads the base document and identifies, in an essentially arbitrary manner[1], what are believed to be key words or terms. These are, again in an essentially arbitrary manner, classified. When this manual analysis of the base document is complete, the target document is read and key terms are identified and placed in one of the base document's classification buckets—or, if needed, new buckets are created. This preparatory strategy provides structure to what would otherwise be a random activity. This is a strategy than can only be described as: time consuming, exhausting, arbitrary, and demanding a high level of subject matter expertise. Value can be realized only after this labour intensive work is done.

There is, for life sciences documents, an alternative. Software can do the laborious preparatory work so that the user can focus on the high level analysis. For the purpose of the discussion as it continues, "software" and "tool" are used in reference to i4i's ALiCE[2] solution.

[1] The identification is not 100% arbitrary; it is informed by the expertise of the reader. However expert, the reader is only human and as such variable—thus assignments may vary for no discernible reason.

[2] ALiCE – **A**uthoring **Li**fecycle & **C**ollaboration **E**nvironment

Concept analytics

Concept analytics does two things: it creates a database of metadata for terms found in the documents of interest and then uses that metadata to compare documents.

Metadata database

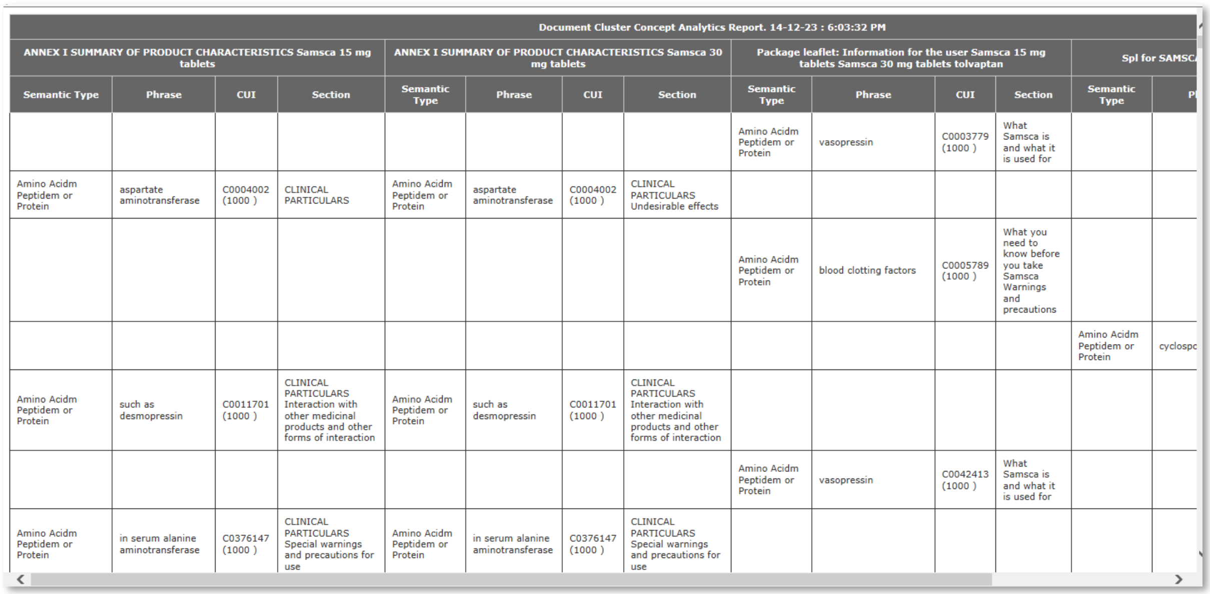
The tool creates the database by parsing the document and identifying for each term the possible semantic or conceptual role(s) of the identified terms. It then stores in a database as metadata to the document: the terms, their conceptual role(s) and information on the location of the terms in the documents, the SNOMED or MedDRA or NDA preferred expression for the term, and a unique identifier for the term. This metadata is developed by using software that in part includes the UMLS[3] semantic analysis tools. The UMLS tools have been selected so that the user of the i4i concept analytic software will be working with datasets that are recognized as authoritative by the regulator. The metadata and document management aspects of the metadata database as well as all the metadata analytics software are specific to i4i.

Using the metadata: analysing document clusters

Document clusters are collections of documents that are related in some manner. An example of a document cluster is a drug product’s set of regulatory information documents: SmPC, PL, SPL, CCDS, etc. Each of these documents describes, in its own distinct manner, the same product or some aspects of the same product. They are of course topographically different; thus comparing them with a track changes model returns an effectively meaningless result. They should however have a degree of conceptual similarity: what the product is indicated for, possible adverse events, contraindications, substances used, and so forth.

[3] The Unified Medical Language System® includes SNOMED, MedDRA, and NDF, as well as other datasets.

Concept analytics is used to create a Document Cluster Report[4], as shown below.



Document Cluster Concept Analytics Report. 14-12-23 : 6:03:32 PM													
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 15 mg tablets				ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 30 mg tablets				Package leaflet: Information for the user Samsca 15 mg tablets Samsca 30 mg tablets tolvaptan				Spl for SAMSC	
Semantic Type	Phrase	CUI	Section	Semantic Type	Phrase	CUI	Section	Semantic Type	Phrase	CUI	Section	Semantic Type	Pl
								Amino Acidm Peptidem or Protein	vasopressin	C0003779 (1000)	What Samsca is and what it is used for		
Amino Acidm Peptidem or Protein	aspartate aminotransferase	C0004002 (1000)	CLINICAL PARTICULARS	Amino Acidm Peptidem or Protein	aspartate aminotransferase	C0004002 (1000)	CLINICAL PARTICULARS Undesirable effects						
								Amino Acidm Peptidem or Protein	blood clotting factors	C0005789 (1000)	What you need to know before you take Samsca Warnings and precautions		
												Amino Acidm Peptidem or Protein	cyclospc
Amino Acidm Peptidem or Protein	such as desmopressin	C0011701 (1000)	CLINICAL PARTICULARS Interaction with other medicinal products and other forms of interaction	Amino Acidm Peptidem or Protein	such as desmopressin	C0011701 (1000)	CLINICAL PARTICULARS Interaction with other medicinal products and other forms of interaction						
								Amino Acidm Peptidem or Protein	vasopressin	C0042413 (1000)	What Samsca is and what it is used for		
Amino Acidm Peptidem or Protein	in serum alanine aminotransferase	C0376147 (1000)	CLINICAL PARTICULARS Special warnings and precautions for use	Amino Acidm Peptidem or Protein	in serum alanine aminotransferase	C0376147 (1000)	CLINICAL PARTICULARS Special warnings and precautions for use						

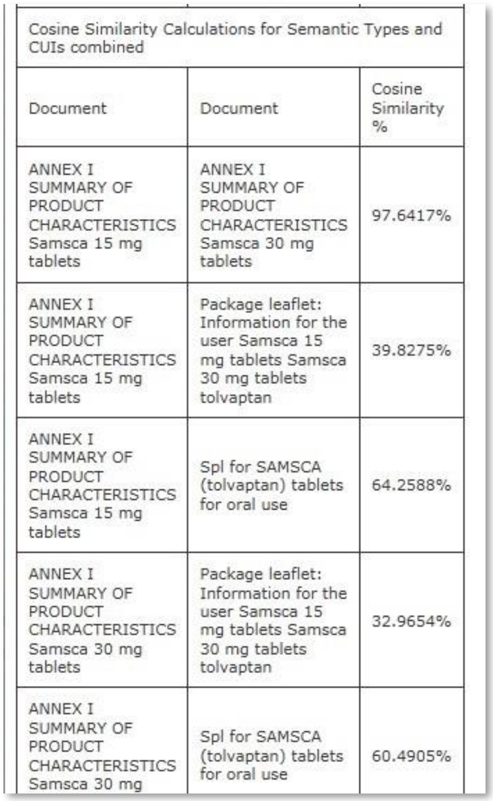
This is a report of terms used in the document, grouped by semantic classification.

The documents selected for the report shown above are related by virtue of being label documents for the product, Samsca tablets. What is shown is that all three documents, the SmPC, the PL and, the SPL, discuss the concept *Body Substance* and instantiate that concept in each document with the term *fluid*[5]. As can be seen, the SPL has two further instantiations, *water* and *urine*, to the SmPC or the PL. Two documents, the SmPC and the PL, discuss the concept of *Age Group* instantiated in the term *children*, while the SPL is silent on that matter. In all, over 600 terms from the three documents are classified and grouped for the user to review and determine the level and specifics of the documents' similarities and differences. The user has at their immediate disposal the question: why is *age group* discussed in the SmPC and PL but not the SPL?

[4] The report is available in two formats, HTML for viewing a browser and csv for viewing and further manipulation in a spreadsheet.

[5] It must be noted that term *fluid* is ambiguous as it can also be an instantiation of the concept *Qualitative Concept*. This is identified further in the report. The user elected to organize the display by concept; they could as easily have elected to display by term or section.

Supporting the data in the Document Cluster Report is a statistical analysis of the similarities of the documents in the cluster. Using the Cosine similarity algorithm[6], the report determines a statistical degree of similarity.



Cosine Similarity Calculations for Semantic Types and CUIs combined

Document	Document	Cosine Similarity %
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 15 mg tablets	ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 30 mg tablets	97.6417%
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 15 mg tablets	Package leaflet: Information for the user Samsca 15 mg tablets Samsca 30 mg tablets tolvaptan	39.8275%
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 15 mg tablets	Spl for SAMSCA (tolvaptan) tablets for oral use	64.2588%
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 30 mg tablets	Package leaflet: Information for the user Samsca 15 mg tablets Samsca 30 mg tablets tolvaptan	32.9654%
ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS Samsca 30 mg	Spl for SAMSCA (tolvaptan) tablets for oral use	60.4905%

Using the metadata: finding similar documents

Concept analytics can also be used to find documents. Consider the situation where the user is developing text for a regulatory document, an SPL or SmPC. Knowing which text has already been approved by the regulator can help in the defense process: "*we are saying it this way because the following approved labels present the same concepts this way*". For instance, if the document in hand discusses *toothaches* and *vertigo* and a*llergic reactions*, it would be helpful to find other documents that discuss the same matter.

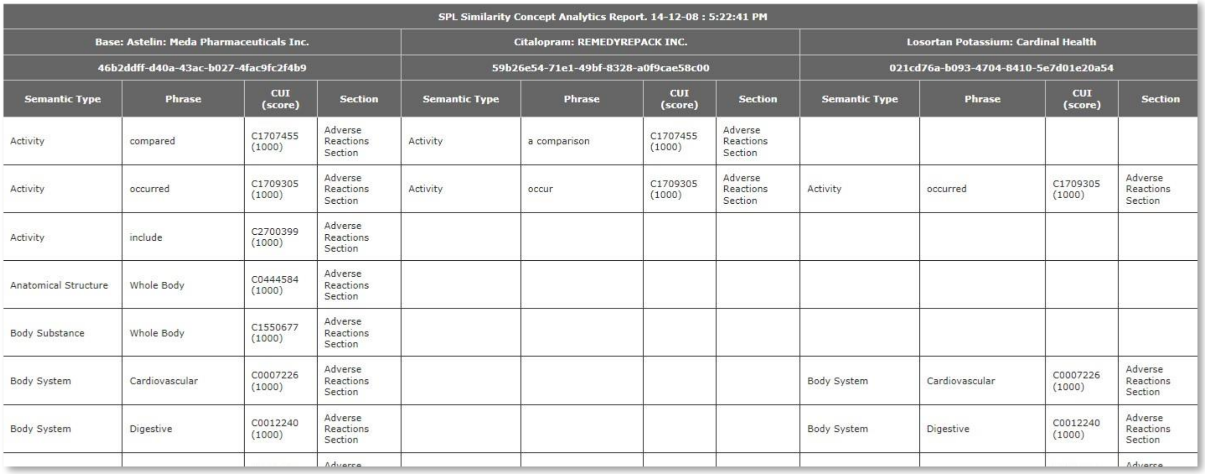
A simple text search for these terms will miss documents that discuss *dental pain* or *dizziness* or *hypersensitivity*. *Toothache* and *dental pain* are instantiated variants of the concept *sign or symptom dental pain*, while *allergic reaction* and *hypersensitivity* are variants of the concept *pathologic function.*

The concept analytics *SPL Similarity* report finds similar documents by building a concept profile of the base document and finding SPLs with the same or similar profiles. A typical profile is complex and can have hundreds if not thousands of data

[6] http://en.wikipedia.org/wiki/Cosine_similarity

points. Finding similar profiles in a database of thousands of documents, each with their own complex profile can be time consuming and return many positives of marginal value. Because SPLs are structured content, the search can be narrowed to specific sections. In addition, the search can be limited to specific concepts found in the base document.

This report lists the concepts and their instantiations in the base document, and the documents that have the same concepts and synonymous instantiations.



SPL Similarity Concept Analytics Report. 14-12-08 : 5:22:41 PM											
Base: Astelin: Meda Pharmaceuticals Inc.				Citalopram: REMEDYREPACK INC.				Losortan Potassium: Cardinal Health			
46b2ddff-d40a-43ac-b027-4fac9fc2f4b9				59b26e54-71e1-49bf-8328-a0f9cae58c00				021cd76a-b093-4704-8410-5e7d01e20a54			
Semantic Type	Phrase	CUI (score)	Section	Semantic Type	Phrase	CUI (score)	Section	Semantic Type	Phrase	CUI (score)	Section
Activity	compared	C1707455 (1000)	Adverse Reactions Section	Activity	a comparison	C1707455 (1000)	Adverse Reactions Section				
Activity	occurred	C1709305 (1000)	Adverse Reactions Section	Activity	occur	C1709305 (1000)	Adverse Reactions Section	Activity	occurred	C1709305 (1000)	Adverse Reactions Section
Activity	include	C2700399 (1000)	Adverse Reactions Section								
Anatomical Structure	Whole Body	C0444584 (1000)	Adverse Reactions Section								
Body Substance	Whole Body	C1550677 (1000)	Adverse Reactions Section								
Body System	Cardiovascular	C0007226 (1000)	Adverse Reactions Section					Body System	Cardiovascular	C0007226 (1000)	Adverse Reactions Section
Body System	Digestive	C0012240 (1000)	Adverse Reactions Section					Body System	Digestive	C0012240 (1000)	Adverse Reactions Section
			Adverse								Adverse

For completeness, concept instantiations not found in the base document are presented at the bottom of the report.

Using the metadata: other applications

If concept analytics is used in a structured authoring environment, the metadata developed in the analysis process can be inserted into the source documents. This is done using XML markup as shown: *<concept CUI='C0040460' semantic_type='sign or symptom'>toothache</concept>*. This markup can be used to help navigate the document when, for instance, the user is looking to identify indications for an IDMP submission. The structured content is highlighted and a button navigates the user through the *sign or symptom* concepts in the document.

Concept analytics can even enhance the translation process by providing specificity to concepts when the term in the base may be ambiguous. Example: *cold,* as in *the cold is* *annoying* can be a *natural phenomenon*, a *disease or symptom*, or a *physiologic function.* Potentially ambiguous terms are identified and the user selects the desired meaning. This is held in the concept markup so as to inform the translator of the conceptual intent of the term. And of course if the concept is incorrectly interpreted and translated, a concept analytics of the back translation will uncover the conceptual difference.

Summary

Concept analytics is a tool. It eliminates or rationalizes the heretofore manual task of preparing topographically dissimilar documents for comparative analysis. It will not decide whether the identified differences or similarities are important or of no consequence; that is the role of the user of concept analytics.

Concept analytics is not the tool of choice for a user who is asking the question: “*What are the topological differences between these two documents; how do the documents visually line up?*” Answering that question is the task of “track changes”. Concept analytics is the tool of choice for a user who is asking the question: “*What is the same and what is different in the information these two documents are trying to deliver; how are these documents in information agreement?*”

Contact i4i

116 Spadina Ave., 5th Floor
Toronto, Ontario, Canada M5V 2K6

info@i4i.com

www.i4i.com