

Research Fact Sheet

Multilingual Document Authoring

“MDA (Multilingual Document Authoring) is an interactive tool which assists and guides monolingual writers in the production of multilingual documents.”

The MDA (Multilingual Document Authoring) project provides interactive tools, such as context-aware menus, for assisting monolingual writers in the production of multilingual documents. These tools extend conventional syntax-driven SGML or XML editors so that choices down to the word-level are possible when authoring the document content. In addition, dependencies between two distant parts of the document can be specified in such a way that a change in one part of the document is immediately reflected in a change in some other part of the document.

The author's choices have language-independent meanings (example: choosing between a *solution* and an *emulsion*), which are automatically rendered in any of the languages known to the system, along with their grammatical consequences on the surrounding text. Although the author is not explicitly following standards, the text produced by the system is implicitly controlled both:

- Syntactically: the choice of the standard term for expressing a given notion is under system control, as is the choice between grammatical variants (such as active/passive sentences) for expressing a given information;
- Semantically: the consequences of a choice somewhere are reflected across the whole document, the author cannot forget to provide some information that the system requires, dependencies

between semantic parameters such as *gender* and *pregnancy* can be described.

Background

We are witnessing a general trend towards environments in which the writer is guided in the authoring process by context-aware systems such as XML DTD's or MS-Word wizards. Such programs are still linguistically poor, in at least two dimensions:

- Limited capabilities to handle sub-sentential (or even sub-paragraphical) chunks of text as units. For, at this fine-grained level, grammatical aspects such as number agreement, coordination, or choice of pronouns start to play a prominent role, for which non-linguistically aware tools such as DTD's are inadequate. Such tools then have to have recourse to "PCDATA" nodes, under which any free text is acceptable.
- Limited capabilities to specify dependencies between pieces of text, which appear at a certain distance from each other in the document. For example, in a drug package-insert, if the "Dosage and Administration" part specifies dosages for *adults and children above 15, children between 12 and 15, and children between 6 and 12*, but nothing else, then in the part "Contra-indications", *children under the age of 6* should normally be mentioned. Such

relationships are mostly ignored in conventional DTD's, both because the granularity of the elements entering into such dependencies is typically smaller than the units they can handle, and also because they have no formal mechanism permitting the expression of such dependencies.

These observations are strong indications that there is in this area a large space in which language technology can play an important role and complement conventional structured document authoring.

MDA Functionalities

At XRCE, we have developed a prototype system, which has the following characteristics: **The system supports the whole authoring process**, from document macro-structure to choosing words. Both linguistic knowledge and knowledge about the document are represented in a uniform formalism. A grammar in this formalism can be seen as an enriched DTD, in the following sense: if one ignores certain aspects of the grammar rules, then one gets a standard DTD (but with finer-grained units than traditional); but if one takes into account these aspects, then (i) long-distance dependencies start to arise between different units in the DTD, (ii) these units are provided with linguistic realizations in the different languages and with conditions for combining these realizations into a grammatical text.

The author is guided in his/her choices. During the authoring process, the system maintains a typed "abstract-tree" structure, which fully represents, in a language-independent way, the communicative content of the document. The authoring process consists of adding new nodes to this abstract-tree in a stepwise fashion, until the tree cannot be extended further. At any given moment only certain extension choices are possible (compatible with the tree-typing). The system presents these choices to the author in the form of menus. In this way, the author can quickly identify the valid possibilities without performing external checks and his/her choices are guaranteed to eventually lead to a valid document. Such a process eliminates frustrating *a posteriori* control procedures telling the author that the paragraph/sentence/term/word s/he has decided on is forbidden or not recommended.

The authoring process is monolingual, but the results are multilingual. At each point of the authoring process, the author can view *in his/her own language* the text s/he has authored so far, and areas where the text still needs refinement are highlighted. The menus presented to the author for the choices associated with these areas are also displayed to the author in his/her own language. Thus, the author is always overtly working in the language s/he knows, but is implicitly building a language-independent abstract-tree representing the content of the document. From this tree, the system automatically builds a grammatical text in any of the several languages it handles.

The granularity of linguistic description can vary depending on needs. To each leaf ("concept") in the abstract-tree corresponds a text fragment (continuous or discontinuous piece of text) in every language handled by the system, and these fragments are automatically combined to produce grammatically correct documents. The size of the text fragment associated with one abstract-tree concept can vary from a single word to a multiword expression, a full sentence, or even a whole paragraph, depending on such factors as productivity and reusability of concepts in the domain under consideration. A single concept "responsibility-waiving" can correspond to a stereotypical paragraph denying company responsibility for misuse of a product, and there would be no practical value in a fine-grained analysis of the frozen expressions appearing in such a paragraph, because they are never re-used in different contexts in the same domain of discourse. On the other hand, in the pharmaceutical domain, the type of packaging used for a drug (tablets, powder, solution, etc.) will typically be expressed by a single word or a short expression and will be re-used several times in different situations.

A prototype application: package inserts for drugs

Pharmaceutical package inserts are short, highly structured documents that have to conform to norms and to be made available to their intended public in many different languages.

On the basis of documentation regulating the appearance of French package inserts and of a number of examples of such inserts, we have

developed a prototype which generates grammatical French and English inserts, and which handles semantic dependencies between distant elements in the document.

Figure 1 shows the interaction window at an early stage in the authoring of an insert, using English as the interaction language. The author is making a choice, which results in a new state of the document, shown in Figure 2. The author can choose to display the document in the other languages handled by the system, as shown in Figures 3 and 4.

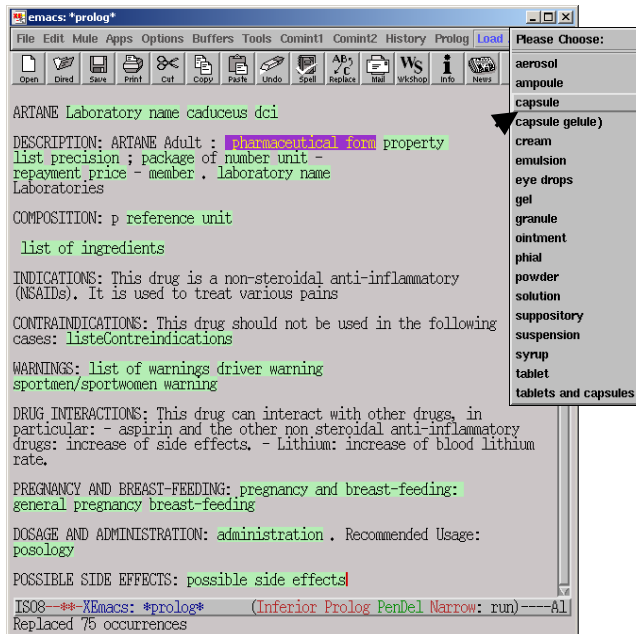


Figure 1: A step in the authoring process using English as the interaction language

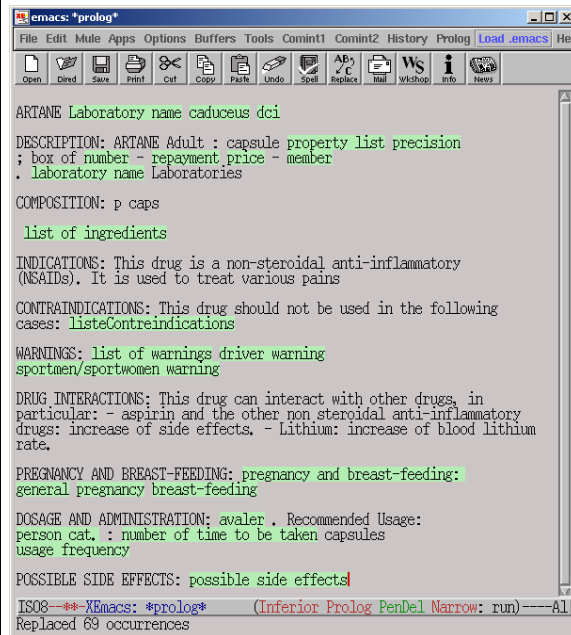


Figure 2: The resulting document in English ...

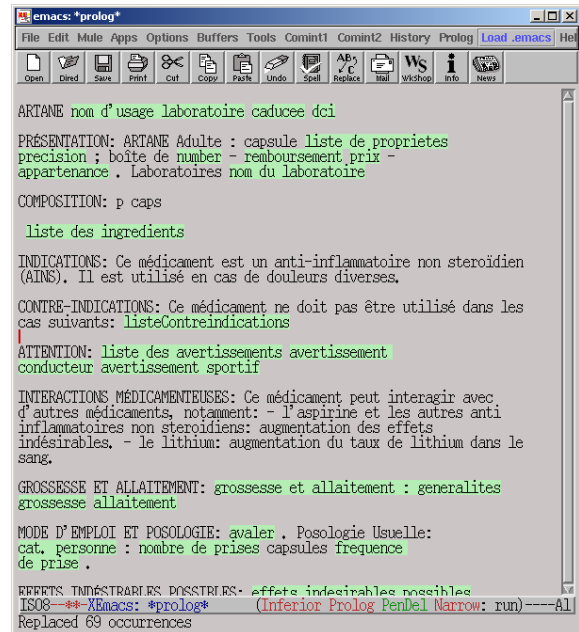


Figure 3: ... in French

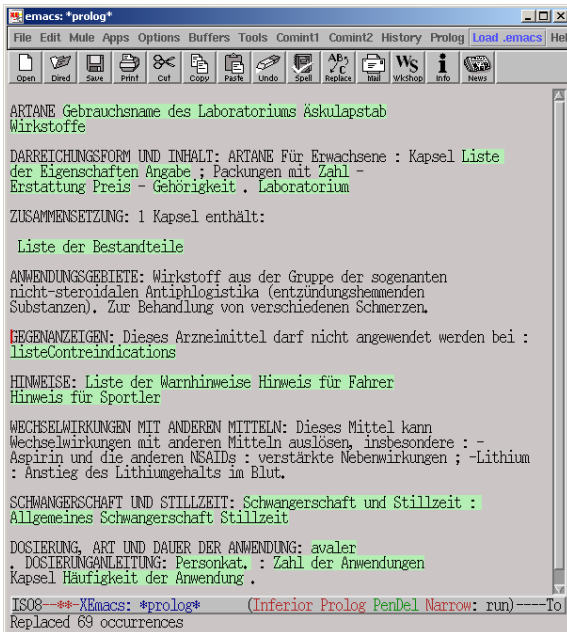


Figure 4: ... in German

Future Developments

The MDA approach has potential to be applied to situations where one needs to produce validated – both semantically and grammatically – multilingual versions of documents under time-pressure conditions. It is especially valuable in cases where the whole document (pharmaceutical notices, letters in response to customer requests) or critical parts of it (for instance portions such as Warnings-Cautions in aeronautical technical documentation) can be composed out of a well-delimited set of communicative elements.

Next steps involve better identifying such domains and establishing close connection with potential users of the technology. We expect that their needs will require adaptations of the current prototype and we can already foresee several areas for potential improvement:

- Better user-interfaces, going further than the top-down refinement mode now in place, and permitting users to do such things as starting from an existing MDA-authored and modifying it by pointing and clicking to obtain a related valid

document (the situation where several minor variants of a given document need to be produced seems to be frequent).

- Allowing more support for “memory-based” authoring, by integrating into the system data-bases of re-usable text fragments.
- Using MDA as a basis for further automation of the authoring process in cases where information normally directly elicited from the author can be obtained directly from existing formal representations or data-bases.

For more information on the **Multilingual Document Authoring** research project:

Information: <http://www.xrce.xerox.com/competencies/content-analysis/dcm/>

Video: <http://www.xrce.xerox.com/competencies/content-analysis/dcm/mda-demo.html>

please contact:

Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France

Marc Dymetman

E-mail: Marc.Dymetman@xrce.xerox.com

WWW: <http://www.xrce.xerox.com/>

Tel: +33 (0)4 76 61 50 50

Fax: +33 (0)4 76 61 50 99