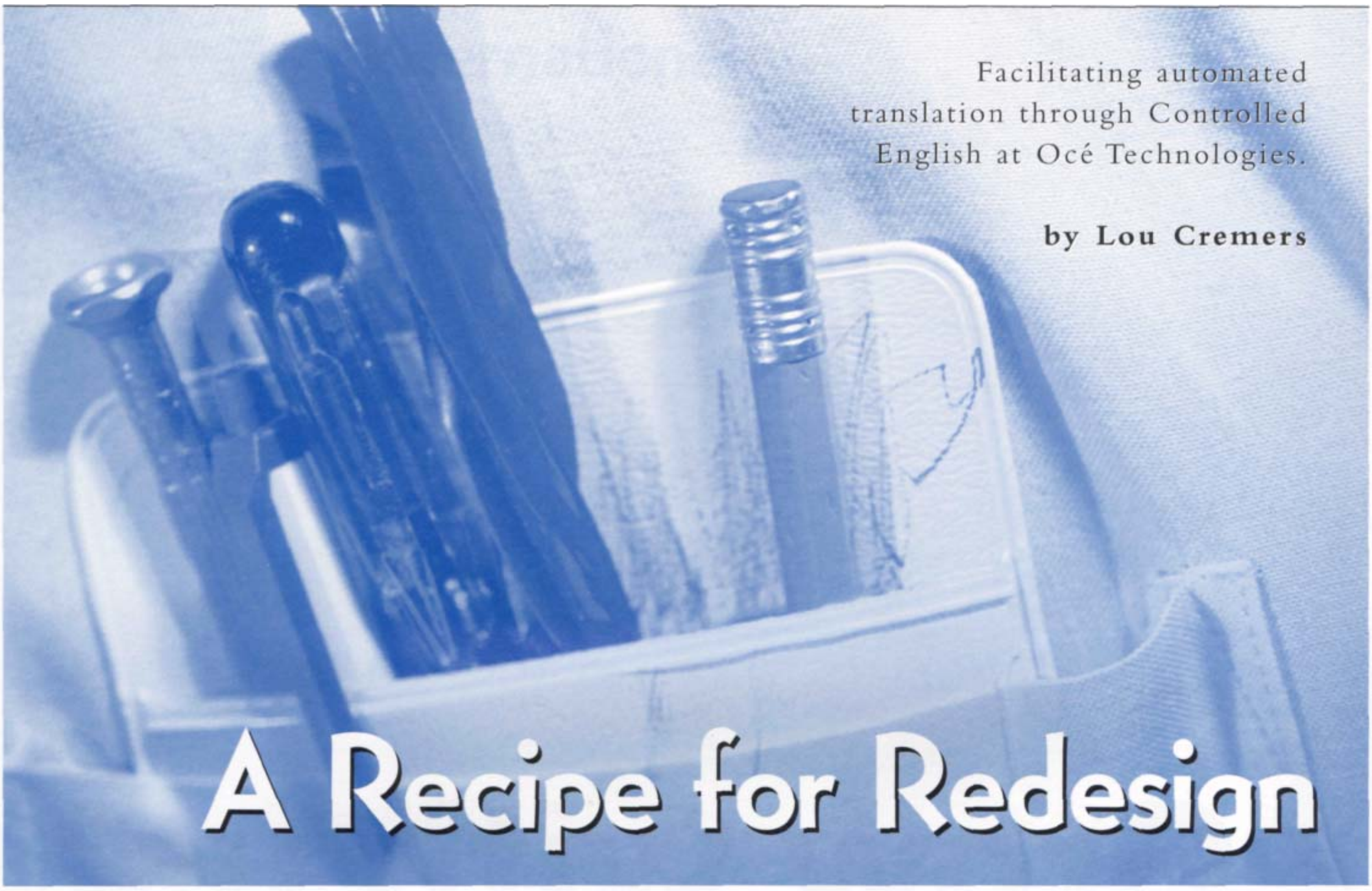

Facilitating automated translation through Controlled English at Océ Technologies.

by Lou Cremers

A Recipe for Redesign

Océ Technologies has a small internal department, Translation Services, which is responsible for the translation of user reference documents and service documentation and some of the software. There are no in-house translators, all translations are outsourced. The department is mainly responsible for the automation of the translation process.

A few years ago, we described in this magazine the implementation of machine translation (MT) at Océ, (*LI* 9.6, 1997). Océ used MT (the Logos system) on a commercial basis for the first time in 1996. Soon afterwards, a TM (translation memory) system from Trados was added. Currently, Océ is implementing additional modules to further improve both documentation and translation workflow by the introduction of XML, Controlled English, and Text & Structure. The sequence of implementation (MT before CE) is rather unusual, being due to historical and organizational reasons.

Reasons for Change

There were a number of reasons that led to a redesign of the documentation process, regarding both the creation of documents and the translation of documents.

Although automated translation through MT and TM resulted in a considerable reduction of cost and time required to translate user reference documents and service documentation, there was still space for improvement. Océ faced an increasing number of types of documents to be translated, ranging from software to online help and Lotus Notes databases. All documentation was written for a specific publication, so a lot of time was spent on DTP. Each document format required specific conversions to enable the translation of each source document through a common workflow.

The fact that all documentation was being written in "uncontrolled" language did not help the translation process. Different styles, inconsistent use of terminology, and typical second-language phenomena resulted in documentation that was difficult to translate—not only for machines, but often also for human translators. In addition, documentation was not generally available online. Consequently there was little reuse of documentation, except for updates.

Translation Services reached a situation where it was difficult to continue to improve automation of the final phase, namely the translation process. Further, the production cycles were shortened. To further increase quality, speed, and cost efficiency, Océ needed to revert to the documentation creation process, and introduce changes going far beyond the translation phase alone.

XML Database

Central in the new approach is a multilingual XML document-management database, which typically contains modular information elements instead of chapters and books. These information elements will subsequently be combined to form specific publications (whether as HTML, help, or a FrameMaker book). Publication will be a largely automated process.

To avoid the problems outlined above, it was essential to develop a completely new way of working. To achieve this, two modules were introduced before the actual XML implementation to ensure optimal standardization and control of the written texts: Text &

Structure and Controlled English. In addition, a careful analysis of the types and contents of current and future information shaped the database for document layout.

Text & Structure

The first step in the documentation redesign is an Océ-internal course called Text & Structure, introducing the general principles of how to present information and how to group, label, and "chunk" information.

Controlled English (CE) is complementary to Text & Structure, describing how to formulate the information elements. CE follows the AECMA Simplified English rules, but it is less strict and is adapted to the Océ environment. The program works with a rule set and a vocabulary. The vocabulary consists of three parts: a core dictionary, a technical dictionary (Océ specific), and a synonym dictionary.

Océ has adopted the MAXit CE checker from Smart Communications. It is fairly easy to customize, use, and maintain. It is available as a plug-in for the various editing environments currently in use, as well as for the XML editing environment that is to be introduced in the near future.

The CE checker can be set to check at various levels for paragraph, page, or document, employing a color code to mark the violations encountered against rules and CE terminology. Violations are described by one of a set of 40 messages. Additionally the program offers explanations and/or suggestions for correction.

A few months were spent on building the Océ-specific CE dictionary, and some rules were modified. One of the customizations was not to have MAXit prompt users to write out numbers below 10 in letter form. This would have created many fuzzy matches from 100-percent matches when analyzing service documentation against a translation memory.

The introduction of Controlled English has also made it possible for the first time to have some control over the use and introduction of terminology—a critical requirement for the success of the entire design.

Overcoming Obstacles

Since it was crucial to adopt a different writing method before starting to fill the database, many different skills had to be taught and acquired practically simultaneously. First, authors had to learn to think differently about the structure and presentation of information. Instead of thinking in terms of the final publication, they now need to think in terms of information modules, independent of the final appearance.

Next they had to learn to phrase the information according to CE principles. This is very time consuming, especially in the early phases. However, as a result of the continuous feedback, new material should soon be produced at the same speed as before, perhaps even more quickly. Although resistance to a tool that heavily constrains creativity would have been quite expected under the circumstances, most authors were quite happy with it. They recognized the usefulness and, indeed, the necessity for such tooling. Finally, the editing environment has changed as well. In phases, all documentation will be written in XML, which will require knowledge of XML and the XML editing environment.

The Océ translation department had to change their workflow to include and adapt to XML as an additional format. The old document formats will continue to persist for a number of years. Océ decided not to convert and rewrite existing documentation. A new translation-memory component will be installed to suit the new workflow, and this will be able to handle current formats plus native XML. A special workflow had to be established for terminology maintenance: a central terminology database to maintain the new CE terminology as well as existing dictionaries. From this central database it will be possible to export, for example, a dictionary for MT or TM. New terminology, proposed by the writers, goes through validation phases before it is added to both MAXit and the database. Authors have additional terminology support available either through a local browser or an intranet Web browser, giving them access to terms, definitions, translations, synonyms, and even translation memories. These browsers need to be fully synchronized with the MAXit CE dictionaries. The browsers are likewise fed from the same central database.

> **The first step in documentation redesign is an Océ-internal course called Text & Structure, introducing the general principles of how to present information. [...] Authors learn to think differently about the structure and presentation of information. Instead of focusing on the final publication, they now need to think in terms of information modules, independent of the final appearance.**

Benefits

In the new documentation process, the quality of documentation will improve quite dramatically. The documentation will be more clear, concise, and consistent in the use of terminology. A side-effect of using the CE tool is a reduction of superfluous information (the "need-to-know" as opposed to the "nice-to-know"). In the end, only one source text format should remain—XML. This will eliminate the need to convert each document type for the translation process to one common format, which to date has been RTF. Reuse of existing text should improve, since the online database will contain publication-independent generic information modules. There will also be a further reduction in translation efforts. Fully reused text elements (100-percent matches) will no longer need to be processed, which is not the case in the current workflow with translation memories. Because the source material will meet a higher quality standard, the MT system produces much better draft translation for FIGS languages (French, Italian, German, and Spanish). Translators need less time to post-edit the MT-drafts, and additionally save time when translating from scratch. Finally, the authors are freed from DTP work, and can fully concentrate on content.

Lou Cremer's special interests include computational linguistics, speech technology, and information science. Since 1995, he has been responsible for automating translation at Océ's International Training Centre, including MT and translation-memory systems. Contact him at lcr@oce.nl