

The Deployment of Language Technology in an Industrial Setting

Jörg Schütz
IAI
Martin-Luther-Str. 14
D-66111 Saarbrücken
joerg@iai.uni-sb.de

Introduction

This paper gives a short overview of the first results obtained in a case study within the MULTIDOC project that aims at designing and building an integrated architecture for multilingual document processing and production in the automotive industry. The MULTIDOC intelligent technical information management system shall bring together leading SGML editing tools with sophisticated repository software, to support the process of authoring and managing shared corporate information in a supportive, structured environment, and to enable that information to be published in a variety of formats. The MULTIDOC system shall also provide full support for translated text in all formats, radically reducing the cost of managing multilingual translations.

Working with document content means working with language. Therefore we are also investigating the potential of state-of-the-art language technology for its integration into the MULTIDOC architecture.

Business Requirements

The MULTIDOC partners belong to different European car and truck manufacturers, in particular to the departments which are responsible for the production and publishing of technical information, aimed at helping technicians to maintain and repair units and their constituent components. These departments are dedicated to the production of all the various types of technical information used in dealers workshops, as well as the customer-oriented information provided with a car when it is delivered.

The information provided at dealerships includes not only traditional hard copy, but information delivered electronically by the companies computer systems. The content of the electronic and paper version of a particular information set may be similar, but the output format is very different, and the means by which information is accessed and navigated differs radically.

In most companies information has historically been authored and managed on a variety of different, incompatible platforms. There were/are different publishing methods, and different authoring approaches, for substantially the same information; a lot of the effort expended in such a situation is redundant. Since all of the MULTIDOC partners information is published in different languages (up to 23 languages), there are major translation overheads incurred by a revision to any existing publications, as well as upon release of a new publication.

Most cars are designed with re-use of existing components in mind, it makes sense that the information relating to those components, and the major units they make up, should be prime for re-use also. This would help to relieve the authoring workload, at the same time as reducing or even eliminating necessary duplication; it would also help to reduce translation costs if, in the event of changes to a piece of information, the computer system could identify exactly what changes had taken place, and constrain the translator to re-translate only those elements.

In the automotive industry the question was: *how do we ensure that the information we author is re-usable, rather than adding to the burden of managing a pool of increasingly technical information?* The answer was prompted by the imminent SAE J2008 legislation, which affects all automotive manufacturer who sell into the US market. J2008 prescribes the use of SGML (Standard Generalized Markup Language) as the data tagging protocol of electronic service-related information, where that information is pertinent to the maintenance of emission-related components of a car. Therefore most/all European car manufacturers have developed an SGML authoring capability within their companies.

In addition, efforts are under way to devising an object-oriented approach to information management, which penetrates deeper than the document, to address individual information units. These information units, SGML elements which are tracked throughout their lives, can be stored as discrete objects and shared many times between output publications and other information-related activities.

Whereas the evolution of this approach is open for addressing a corporate information warehousing strategy, which will enable many users within the enterprise to share a common corporate information pool, there is obviously a gap between the envisaged information technology being deployed (e.g. thin client/server models) and the effective integration of language technology to support the information management system by linguistic means, which, on the one hand, ensures better intelligent information access and update, and on the hand, significantly improves the authoring process in terms of consistency, accuracy and translatability.

Language Technology Deployment

Currently only standalone solutions have been elaborated such as multilingual terminology databases, monolingual syntax and style checkers and computer-aided translation utilities (translation memories and fully automatic translation), but the aim is to include linguistic technology (or linguistic intelligence) in the overall workflow of information processing and production to eliminate the existing overheads in multilingual technical information management.

In this context ALEP was considered as an open architecture approach which could satisfy some of the language-related requirements of MULTIDOC, which are SGML-compatibility, platform-independency and openness for distributed linguistic applications and language resources, extensibility of language resources (included the use of different resources), and low linguistic maintenance.

At the moment, the results of our investigation are not such promising in that the current ALEP release is not yet mature enough for a real integration into the envisaged technical infrastructure of the MULTIDOC architecture. However, we did also not find any appropriate alternative neither in the language technology market nor in R&D laboratories.

According to our evaluation procedure the major drawbacks of ALEP within the context of the MULTIDOC architecture are:

1. ALEP is a monolithic system.
2. ALEP networking capabilities seem to be quite limited or not existing, e.g. for client/server applications.
3. There is no easy integration potential on the application layer (although the ALEP documentation argues the opposite way).
4. There are only slowly emerging language resources available for the ALEP core system (e.g. from projects such as LS-GRAM, GramCheck, etc.).
5. ALEP is Unix- and Emacs-based, only the former can be seen as an industrial standard.
6. There is no real development environment for language engineers (developers of linguistic technology), e.g. visualisation utilities for grammars and lexicons, and tracing and debugging, which in addition could be re-usable with other resources and processing modules.

Prospects

At the moment it is too early for final conclusions. The evolution of the ALEP toolbox is however of great interest because today no other system with similar capabilities is on the market. The developments foreseen in the MELISSA project certainly will contribute to the further evolution of this linguistic platform within an industrial setting. This in particular because the MELISSA architecture makes use of very advanced information technology features (e.g. middleware based on CORBA), which have not yet found an operational basis in the industrial field of MULTIDOC.