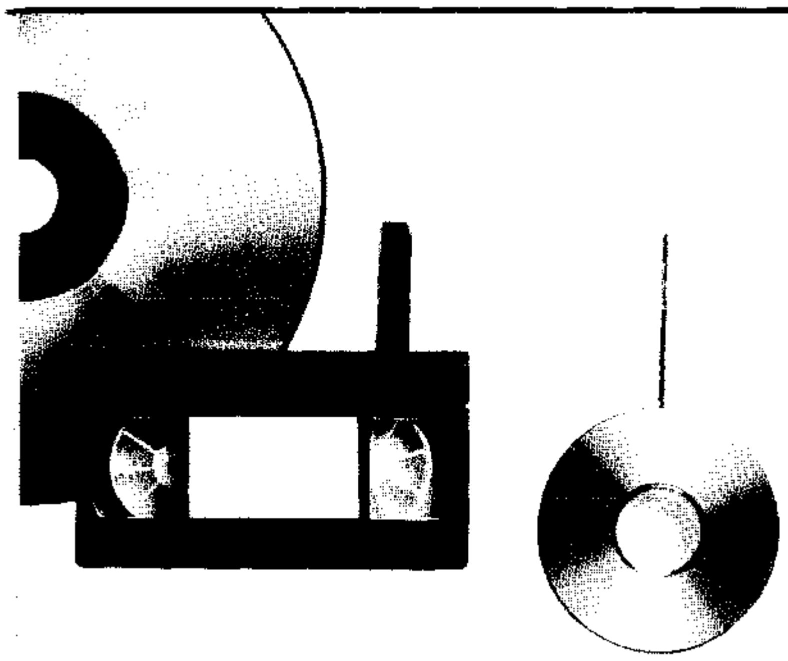


Behind the Scenes

Data-exchange standards are unsung heroes revolutionizing the language industries

By Alan Melby

Data-exchange standards are not usually in the headlines, and you don't think about them until you need them.

Suppose that two previously unrelated companies decide to merge. Each company has a sophisticated employee database stored in a relational database-management system. Assuming the designers of the two databases did not use the same database layout, it is unlikely that the creation of a single employee database for the new merged company will be a trivial task. Each database will have its own types of data elements and relationships among them. Information from one database cannot simply be exported to a comma-delimited file and imported into the other database. One way to get employee information from both companies into the same database is by defining a data-exchange standard. One approach to defining data-exchange standards is to use XML, the subset of SGML that has recently received a lot of press coverage in the computer world. Putting the information into XML will solve the problem, but only if both sides agree to use the same Document Type Definition (DTD) and associated constraints.

To cite an analog from the hardware side of the computer world, suppose you need to get a crucial database onto your computer and you receive it on disk by express mail since it is too large to send as an email attachment. Wouldn't it be irritating to receive a disk that looks like a Zip disk but is just a little too big to stick into your Zip drive?

In the case of the employee-database example, there was no data-exchange standard. You can't just connect two separately developed applications and expect data to flow smoothly between them. And it is not because of lack of adherence to standards, as in the disk story. Until recently, there have been no data-exchange standards to adhere to in the world of translation technology. People have been investing heavily in the creation of large, concept-oriented terminological databases (termbases, for short), translation-memory databases, and machine-translation dictionaries. Unfortunately, most of these language resources are in proprietary formats tied to particular software applications. When users need to re-use these language resources in other software applications, the need for data-exchange standards suddenly becomes painfully obvious.

The good news is that most pieces of translation technology have some provision for exporting and importing data to and from the outside world. The bad news is that the export format for one product usually does not match the import format of another product. Clearly, a solution is needed, but just as clearly, no one vendor of translation technology can solve the problem unilaterally. This is where two standards organizations are coming to the rescue: Oscar and ISO/TC37.

Oscar to the Rescue

Oscar is the data-exchange standards group within the Localization Industry Standards Association (LISA). Technical Committee 37 of the International Organization for Standardization (ISO) is charged with establishing principles and coordination of terminology.

Oscar recently announced that the LISA General Assembly had voted overwhelmingly to approve a standard for translation-memory database exchange, called TMX (Translation Memory eXchange), and ISO/TC37 recently voted to invite the submission of a New Work Item Proposal for terminology interchange, to be called Blind Martif. A European Union project called Otelo has produced OLIF, a format that facilitates the exchange of machine-translation dictionaries. The announcements concerning the three data-exchange formats—TMX, Blind Martif, and OLIF—are the headlines. What is behind the headlines?

TMX is an excellent example of cooperation among competitors. In June 1997, the first Oscar meeting was held. Major software developers as well as major suppliers of translation services called together representatives of the primary suppliers of translation technology, including Trados, Star, Alpnet, Systran, and Logos, to see if the time was ripe for the development of a data-exchange standard for translation-memory databases. Franz Rau

Translation technology will enter a new era, with the end-users winning big.

Once Trados, Star, and Alpnet have implemented TMX, no translation-memory product will be able to survive without implementing it.

of Microsoft chaired the meeting, and the group agreed to function as a LISA special interest group. The suppliers of translation technology acknowledged that users were demanding an easier way to re-use their translation-memory resources and that the creation of a translation-memory exchange standard would "level the playing field," that is, allow software products that use translation-memory databases to stand or fall on their technical merits rather than holding their users captive because they cannot easily get their translation-memory databases out of one product and into another.

Common Cause

Not only did the participants in that June 1997 meeting agree to cooperate, they agreed to do so quickly. The task of developing a translation-memory exchange format was divided into two stages: the first would define the "container," the high-level structure of a translation-memory exchange file, down to the segment level but no lower. The second would define a way to preserve valuable formatting information inside segments. This two-stage approach gave rise to the name of the group, Oscar, which stands for Open Standards for Container/content Allowing Re-use. The most surprising thing about Oscar is that it actually produced TMX and submitted it to the LISA General Assembly in about a year. The momentum of that first meeting in Arlington, Virginia, was maintained as Oscar met in several locations at conferences, summits, and meetings around the world throughout the year. Now the various Oscar representatives are busy implementing TMX into their respective translation-memory products, and translation technology will enter a new era, with the end-users winning big. Once Trados, Star, and Alpnet have implemented TMX, no translation-memory product will be able to survive without implementing it.

Why did Oscar work so well? Having been in attendance as technical secretary at every Oscar meeting, including the founding meeting, I can say that it is because the core members displayed a rare combination of qualifications: solid technical knowledge of their respective translation-memory products, willingness to compromise when theoretical struggles threaten deadlock, and ability to work hard until the details are finalized.

What about Martif, the terminology interchange format, and why is there a blind variety? Actually, blindness is highly desirable in data-exchange standards, where it contrasts with repeated negotiation. The blind/negotiated contrast can be explained in terms of library catalogs. It is one thing for two libraries to negotiate a way to include each other's catalog in their online system by automatically converting between their two respective catalog formats. It is another thing for an association of libraries to agree on a single-exchange format that will allow anyone from the group to send a catalog entry into a common pool without "seeing" who is currently in the association.

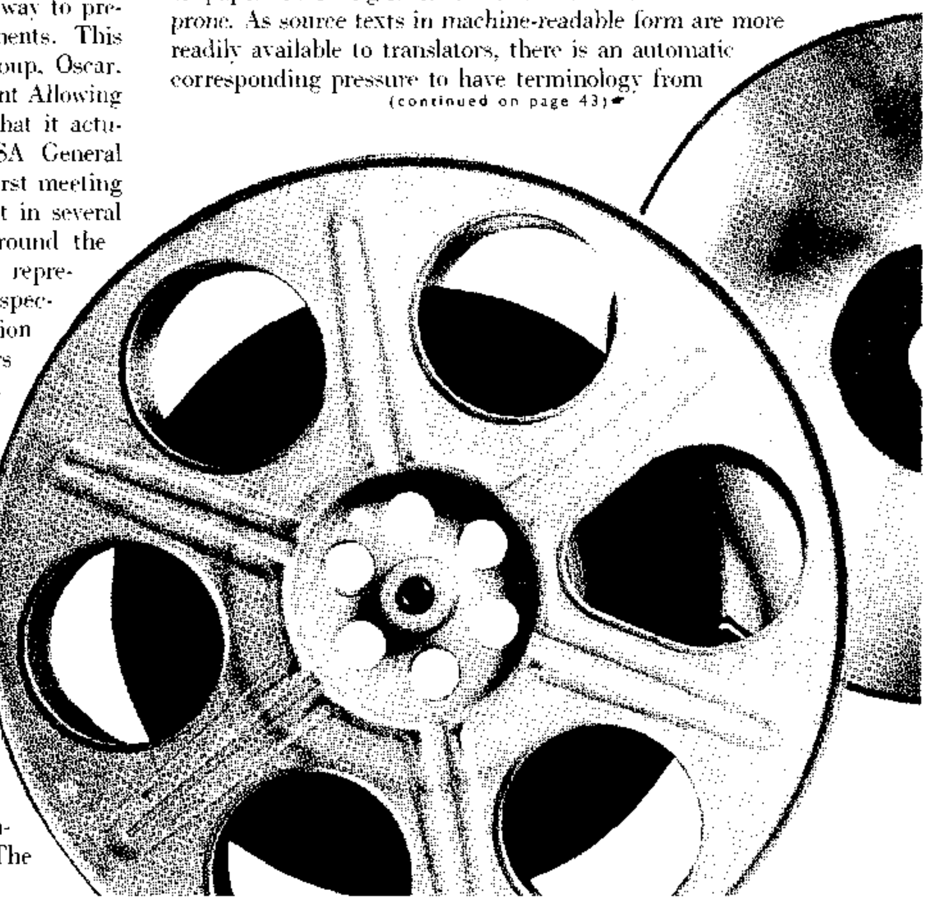
In blind exchange the details are prenegotiated, so that a single import/export routine can be written that will work even for exchanging information with partners not yet in the association, provided each new member agrees to use the established format. This kind of blindness is desirable because it avoids the need to see with whom you are exchanging information.

Martif (Machine-readable Terminology Interchange Format) is a negotiated interchange format (note: interchange and exchange are used "interchangeably" here). However, what real-world requesters and suppliers of translation need is a blind format that will allow organization-specific terminology to be transmitted along with a source text without needing to know what technology will be used to assist in the production of the translation. Simple glossaries consisting only of source-target term pairs are insufficient. Various pieces of information concerning the terms, concepts, and administrative status are needed in today's sophisticated terminology databases.

Going to the Source

Even though nearly every document that is now translated was created in some kind of word processor or desktop-publishing system, many translators still do not have access to the source text in machine-readable form. This will change because of simple economic pressures: many translation tools require the source text to be available in machine-readable form rather than on paper. Scanning is not the answer, since it is slow and error-prone. As source texts in machine-readable form are more readily available to translators, there is an automatic corresponding pressure to have terminology from

(continued on page 43)



the requesting organization also available in machine-readable form. This creates the need for a blind interchange format for terminology because the same tools that take advantage of machine-readable source text also need machine-readable terminology entries to support automatic lookup.

Not If, But When

Therefore, the real question is not whether a blind interchange format for terminology is needed. The question is whether it will be ready soon enough. Although the ISO process has resulted in many important and frequently used standards, it must be accompanied by industry standards creating processes like those used in Oscar. Interestingly enough, the recent ISO vote to invite submission of a new work-item proposal called Blind Martif quickly resulted in an Oscar decision to consider Blind Martif in its work-item form as the basis part of a forthcoming Oscar standard to be called TBX (tremble exchange) that will accompany and complement TMX. This means that there will be parallel work in Oscar and in ISO Technical Committee 37 toward the same objective: a blind interchange format for termbases. It will be interesting to see how they interact, especially since I am both the project leader for TBX within Oscar and the editor of Blind Martif within ISO TC 37.

Exchange standards are increasing relevant for "gisting"—producing a translation that gives a rough indication of the content of a document.

The third type of language resource mentioned at the beginning of this article was machine-translation dictionaries. As machine translation is increasingly used for "gisting"—producing a translation that gives a rough indication of the content of a document—and for domain-specific translations of controlled-language texts, there is an increased need to synchronize the terminology in a machine-translation dictionary with the terminology in a termbase for use in tools for human translators. Another Oscar project is to look at using OLIF (from the Otelo project) and TBX together, along with some specialized software, to facilitate the synchronization process.

The development of the standards discussed above is being hastened by: (1) more mature attitudes of cooperation by competitors in the language industries; (2) underlying standards that they can all build on, such as XML, Unicode, and the Text Encoding Initiative; and (3) insistence on the part of users that drawn-out discussions be translated into rapid progress.

Alan K. Melby is professor in the Department of Linguistics, Brigham Young University at Provo, Utah. He is a member of the Board of Directors of the American Translators Association and active in several organizations dealing with data-exchange standards. On the theoretical side, he is part of a research group studying possible connections between language and agency.