Implementing Z39.50 in a multi-national and multi-lingual environment

Makx Dekkers
Pica - *Centrum voor bibliotheekautomatisering*
Leiden, The Netherlands
(dekkers@pica.nl)

Abstract

Z39.50 provides a very useful tool for intersystem communication, but it also demonstrates that differences in language and culture have an impact on the scope and usefulness of international services.

A number of problems arise when implementing Z39.50 in a multi-national and multi-lingual environment. In the U.S. implementers group, these problems are not always obvious.

Problems that are being identified by current implementers in Europe include differences in character sets between countries, different sorting order of characters in different languages, different rules for conversion from 8-bit to 7-bit ASCII for indexing dependent on country and language, difficulties in translating system messages, variety of MARC formats and differences in cataloguing rules between countries.

Internationalisation of the standard can solve some of the problems. Hopefully, through the implementation and use of EXPLAIN some of the others can be explained to users. The aim should be to make it possible, through Z39.50, to provide services to a wide international audience, respecting the multitude of cultures and languages in the world.

Introduction

As a result of the possibilities offered by the introduction of network technology in the past decades, exchange of information through communication between computer systems has become an everyday phenomenon in today's world.

With the success of Z39.50 in North America, organizations in non-English speaking areas are becoming interested and now are identifying problems with the essential Anglo-American scope of some aspects of Z39.50. Obvious examples are: the use of basic 7-bit ASCII, and the relation with the Anglo-American Cataloguing Rules AACR2. It should be stressed that the issue is not a American-European dichotomy but rather one of language, i.e. English versus non-English communities.

While it is clear that Z39.50 provides a very useful tool for intersystem communication, experience from international projects indicates that solutions will have to be found for national differences such as language, character sets, cataloguing rules and data formats, to make the standard globally acceptable.

Problem areas

It should be stated that many of the problems that are identified by non-English implementers are not inherent to Z39.50, but are merely made visible because of the new possibilities of intersystem communication.

A major category of problems is that of external incompatibilities. These problems are related to cultural differences: national bodies are responsible for definition and maintenance of national rules for bibliographic descriptions; more fundamentally, different countries use different languages with different character sets and sorting rules.

Almost all countries have bodies that define the national rules for cataloguing of publications. Large bibliographic utilities sometimes define their own (additional) rules. Although some coordination is taking place, such as with the AACR2 rules, the independence of these national bodies introduces some fundamental incompatibilities when organisations in

different countries want to interchange bibliographic descriptions through Z39.50.

An illustrative example of this is the treatment of multi-volume publications; under some rules these are catalogued as one single record with repeated elements, under other rules they are described as separate entities with relations between them. Another example is the use of standard phrases in national language within the cataloguing rules, such as for title changes for journals (in Dutch cataloguing the description would contain the phrase "Voortgezet als:"). International exchange of such bibliographic descriptions would ideally involve automatic translation; however, this is not being done in practice.

For searching, standardised keyword lists are usually defined in national language and subject code systems are agreed in a national context. These different language and country related rules and practices cause incompatibilities that are difficult or sometimes impossible to overcome.

A number of interworking problems are associated with differences in character sets. There are many scripts being used in the world and most of the existing library systems are unable to handle them all. Transcription rules or character set conversions will sometimes lose information since they are not always one-to-one reversible. Transcription rules are generally dependent on the combination of source and target language, e.g., Russian Cyrillic will be transcribed differently in the Netherlands, Germany and France. Furthermore, transcription rules sometimes change over time. Especially in searching, it is difficult for a user to determine what transcription should be used: will it be Nabokov, Nabokow, Nabokoff? How does one search for the person referred to in the Netherlands as Aleksandr Isajevitsj Solzjenitsyn? In practice, in current multi-national communication, the textual information is transferred in 7-bit ASCII (ISO 646). This limits the scope of communication to languages with a Latin character set and even then the results are fundamentally insufficient.

In sorting, the situation is even more complicated. The same character set may be used in two different languages, yet the sorting order might be different. In some languages "o-umlaut" is sorted as "oe", in others it might appear at the end of the alphabet. Even when the same language is used in two countries, there might be differences in sorting order of names: in Belgium a personal name of "Van Dam" will appear under "V", in the Netherlands under "D".

All these practices make it difficult to predict what the results of search commands will be; sometimes items will not be found that are in the database, sometimes they will just appear in unexpected places in a sorted list. Users do not always realise these difficulties and the occasional user might turn away in disappointment.

Users would like to look at title descriptions that were found as a result of the search action. This is not a trivial task. The target database records might be stored in a format that the user is unfamiliar with. The systems will then have to provide some form of format conversion. Even between MARC formats, which are at least structurally compatible, format conversions almost invariably lose some of the information. If for some reason format conversions are not possible, the user will either see an unfamiliar format or a description can only be displayed as unstructured text, which might be fine for an end-user doing reference work, but would be close to useless to a librarian.

Finally, there are also some elements of the standard itself that pose problems to users in multi-lingual environments. For several services, the target system may convey diagnostic information to the origin. The standard does not prescribe the origin behaviour when such messages are received, so the origin might not know anything better to do than show the diagnostic information to the user. This information could be in a language the user is not familiar with. Obviously, the standard is not supposed to prescribe any external behaviour but to users this is not really helpful.

Solution scenarios

In practical implementations of the standard, solutions to the above problems take a very pragmatic approach. In one way or another, they all take the lowest common denominator and make the best of it. This either leads to solutions that cut away all complexity and settle for very limited functionality, or to highly parameterised implementations of the form: "If talking to A do this, if talking to B do the other". The language and character set problems are commonly ’solved’ by standardizing on English and basic ASCII.

For real solutions, a first essential step is for developers and implementers of the standard to become aware of the problems that are introduced by its use in multi-national and multi-lingual environments. This awareness can not be of a simple theoretical nature; it can only lead to practical steps if there is a business case to justify extra investments.

One general step towards solving the incompatibilities outlined above is the introduction of negotiations, dynamic conversions, and powerful explanation techniques. Negotiation aims at establishing a mutually agreed environment. If this cannot be achieved, the data that is exchanged can be converted from one format to another. If that is impossible, the user should receive some information to explain why the result is not as expected and to suggest alternative actions to maximise the user's efforts.

In the final text of the 1995 version of the standard, a mechanism has been incorporated for negotiations between origin and target. This mechanism may be used to negotiate the character set used for textual information and the language of messages. This is an essential improvement over the 1992 standard, allowing multi-lingual systems to take advantage of their capabilities across a Z39.50 communication. The character sets that can be used are ISO 10646 and ISO 2022, or mutually agreed private character sets. If negotiation cannot be completed successfully, the situation is basically that of the 1992 standard: the target determines the language and the only safe assumption for the character set is that it will be 7-bit ASCII.

For the data formats, it is clear that national or local rules will remain important to determine the way information is stored in databases. Possibly some of the MARC formats will become widely accepted as an exchange format; at the moment, USMARC and UNIMARC seem to be the dominant formats in several projects. It is also clear that this will not be a solution for systems not governed by purely bibliographic rules that do not usually store or export their data in MARC. A positive development is that some European projects are building table-driven, public domain toolkits for format conversions. Although 100% accuracy in conversion cannot be achieved, this might help in broadening the scope of Z39.50 interoperability.

In areas where negotiation or conversions cannot solve the problems, the use of the Explain facilities defined in the standard would have to provide the solution. This facility is probably the most powerful feature of Z39.50. It can be seen from the complexity of the Explain facility that the problems are manifold. Through Explain, the client is given information that can be used to help the user to better understand what goes on behind the scenes and to allow him or her to make sense of the results of certain actions. The problem, of course, is that what makes perfect sense for one user might be completely illogical to another. At best, Explain will provide a general information level to an average user. Still, this is better than nothing and fortunately, all messages in Explain have been designed for multi-lingual environments. As a drawback, the maintenance of an Explain database can be a considerable task.

Even with the above solutions, there will remain elements that cannot be negotiated, converted, or explained during a session. There will always be out-of-band bilateral agreements between Z39.50 partners and there will always be situations where two Z39.50 systems cannot communicate to provide a useful service to users. As an example of the latter, imagine that the Z39.50 target provides access to a Chinese database and the origin has no way of displaying Chinese characters to the user. It should be clear that a standard like Z39.50 is just a vehicle for communication and will never be capable of solving all the problems that exist; neither should it try to do this.

Conclusion

Z39.50 provides a very useful tool for intersystem communication but it is clear that differences in language and culture have an impact on the scope and usefulness of international services. Internationalisation of the standard can solve some of the problems. Hopefully through the implementation and use of Explain some of the others can be explained to users. The aim should be to make it possible, through Z39.50, to provide services to a wide international audience, respecting the multitude of cultures and languages in the world.