Software Demonstration: **DicoPro: An Online Dictionary Consultation Tool for Language Professionals**

Susan ARMSTRONG, Colin BRACE, Dominique PETITPIERRE,
Gilbert ROBERT, Derek WALKER,
Genève, Switzerland and Amsterdam, The Netherlands

Abstract

The DicoPro project addressed the need for a uniform, platform-independent interface for accessing multiple dictionaries and other lexical resources via the Internet/intranets. Based on HTML and Java, the client software is designed to shield the end-user – typically language professionals (writers and translators) as well as those in the educational sector – from storage and file format concerns. The DicoPro server is a robust cross-platform Java program which gives the user access to multiple dictionaries via the Internet. In order to handle data coming from multiple sources, a powerful transformation tool (XML-Trans) was developed for conversion, validation and indexing of the dictionary data. The multilingual dictionaries currently available are professional, commercial-grade reference works, licensed for testing purposes within the DicoPro project from several prominent European publishers.

1 Background and Overview

1.1 Context

The DicoPro project, which ran from April 1998 to September 1999, was a project funded within the EU Multilingual Information Society Programme (MLIS).[1] The aim of the project was to develop a uniform, platform-independent interface for accessing multiple dictionaries and other lexical resources via the Internet/intranets. The project brought together technical experts for program development, major dictionary publishers providing data and insight into usage of the data and language professionals for testing and validation of the tool.

The background to this project was a dictionary server (DICO) developed at ISSCO in 1990. DICO included similar functionalities to those described below, but was developed to run on a local area network [Robert/Petitpierre 1997]. The DICO system was based on a client-server architecture and offerred two interfaces *xdico* and *tdico*, to accommodate Unix workstatons running X-Windows and PCs via a simple terminal mode. After more than a decade, the program is still operational on the University of Geneva network. It provides access to ten monolingual and bilingual dictionaries and is regularly consulted by hundreds of users. The MLIS DicoPro project can thus be seen as a natural next generation of dictionary servers, taking full advantage of the Internet and the growing potential of e-commerce.

1.2 Project aims

The development of a dictionary server entails more than just the design and implementation of a set of programs. The DicoPro project had three specific goals:

- Data: the conversion of several dictionaries to electronic form, stored in standard SGML format.
- Tools: the development of a uniform, cross-platform tool to enable translators and other language professionals connected to an intranet (or internet) to consult dictionaries and related lexical data from multiple sources.
- Licensing: the exploration and assessment of various distribution and licensing schemes for electronic dictionaries whereby usability and convenience of the end-user is balanced with the need to protect the property of publishers.

1.3 Target audience: language professionals

In their day-to-day work, language professionals, in particular translators and technical writers, consult a variety of dictionaries and other reference works. For example, a single task may require the translator to consult several mono- and bilingual dictionaries to ascertain the precise meaning of a word in a given context. Traditionally, these were all paper-based works, but today more and more such resources are available in electronic format, distributed on CD-ROM.

Unfortunately, the power of today's advanced search techniques is hindered by the variety of formats and proprietary interfaces of these offerings. Even among the offerings of a single publisher, not all CD-ROMs have the same interface. Users are required to deal with multiple search programs, cluttering their screens with redundant windows and toolbars. Concurrent access to the same dictionary by multiple users may not be possible, due to the lack of site licensing schemes enabling shared access, forcing users to pass CD-ROMs back and forth by hand.

This lack of user-friendliness often leads language professionals to prefer traditional paper volumes, but they thereby lose the potential of the powerful search capabilities offered by desktop computers to improve and facilitate their work. As such, companies and organizations involved in translation and technical writing invest less in electronic resources than one might expect, leaving some language professionals with less access to online dictionaries than the average home user.

In the future, as the networked PC becomes ubiquitous, local access to a CD-ROM may not always be possible or desirable, hence the need for a site-oriented client/server approach, encompassing multiple end-user platforms (Windows, Mac, Unix, etc) and based on non-proprietary standards (i.e. SGML, HTML). Though resources are becoming available on the Web, access to the data suffers from many of the same problems described above. As with CD-ROMs, for each publisher and each edition, a new window is required. The plethora of resources found on individual sites is often not properly documented and of limited, if not dubious quality.

2 The DicoPro "Open" Solution

The DicoPro consortium developed what is anticipated will be a commercially viable tool based on existing open standards. The data formats used in the system rely on SGML, HTML and XML technologies. The client and server tools have been developed to run on a wide range of platforms. In particular, all development was done using the portable programming language Java. In this section, we describe the core components of the system in somewhat more detail.

2.1 The dictionary data

A number of bilingual and monolingual dictionaries were supplied by the DicoPro consortium partners for use in the project. Typically, source data obtained from project partners was marked up in SGML-like fashion. This data was converted to valid SGML where necessary. This involved correcting coding errors, extracting the underlying DTD, and then verifying the results. These SGML files were then converted to XML using the standard SGML tool "sx" developed by James Clark.

2.2 Converting the data : XMLTrans

To transform dictionary entries for display in HTML, a transformation tool, XMLTrans, was developed for DicoPro [Walker et al 2000]. The XMLTrans transducer takes as input a well formed XML file and a set of transformation rules and gives as output the application of the rules on the input XML file. It was designed for the processing of large XML files, keeping only the minimum necessary part of the document in memory at all times. The program was written in Java and uses an XML DOM parser. For each dictionary, a set of XMLTrans transformation rules was written and then iteratively improved until the resultant HTML was satisfactory.

XMLTrans was also used to extract relevant fields from entries for indexing. For instance, the translation component of a bilingual entry can be extracted and indexed to allow the user to search the dictionary using only the translation fields of entries.

2.3 The DicoPro server

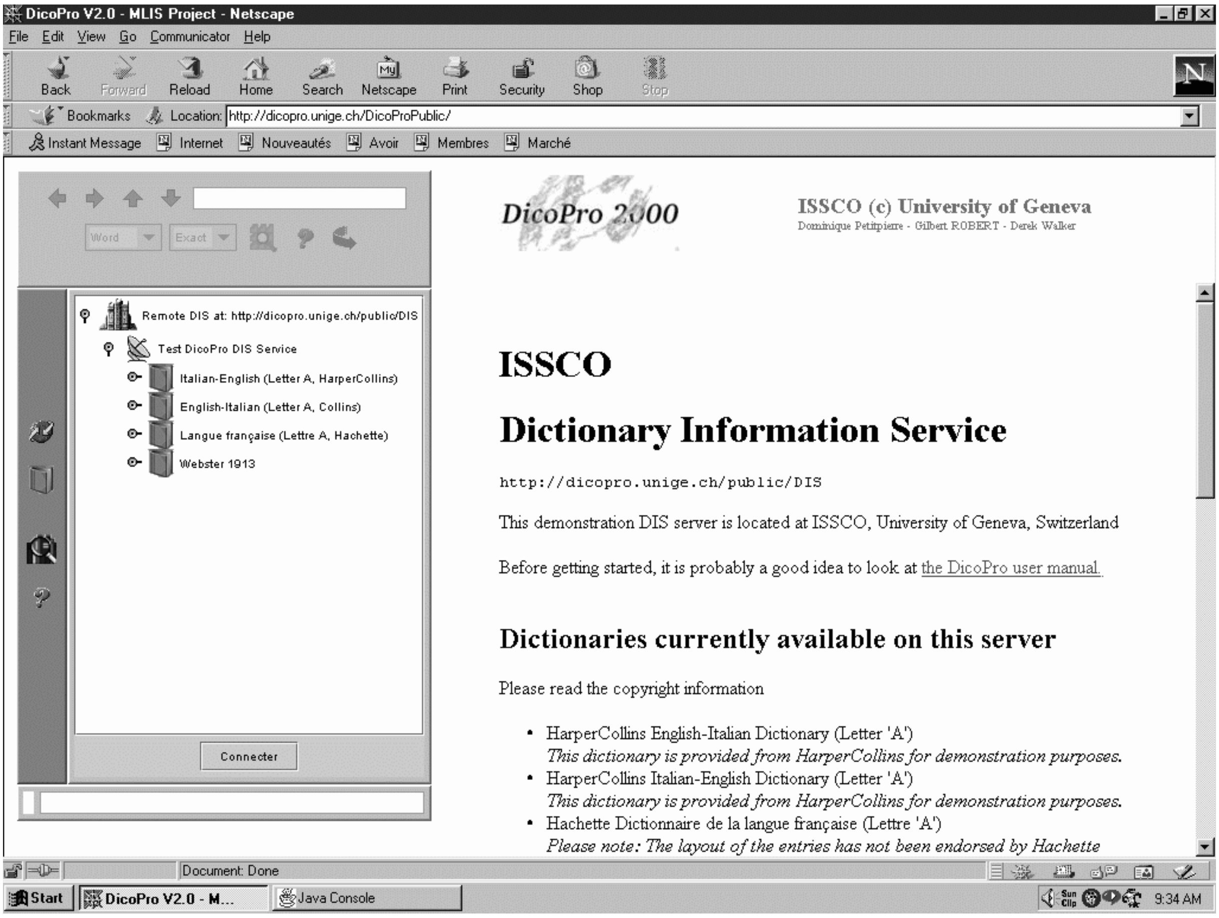
Once prepared, data is stored on the DicoPro server, which is a robust cross platform Java program. It was developed using a threaded design, allowing it to handle many concurrent users accessing diverse data. Connections to the server are made using standard HTTP requests, and are thus capable of passing firewalls. The server can be run as either a standalone application, or as a Servlet from within a web server such as Apache. This second model permits filtering of clients by IP address and the use of SSL encryption.

Dictionaries are distributed in encrypted bundles which can only be decrypted when accompanied with the correct license file. A number of security features were added to the server to prevent tampering with the data or license parameters. Any tampering with the data or license file renders the data useless. Facilities were also added to enforce the license requirements described in Section 3 below. For instance, the time on the client machine must correspond within a reasonable range to the time reported by the server for a connection to be established. Thus the time on the server cannot simply be "rewound" to prevent the license from expiring.

2.4 The DicoPro client

The client is also a cross platform Java application which can be run on Windows, Unix or Macintosh systems. An applet version of the client runs from within a web browser.

The client connects to the Dictionary Information Server (DIS) which provides it with a list of available dictionaries (fig.1). Once opened, each dictionary has its own space with its own menus and options for searching and displaying results. Multiple dictionaries can be opened and consulted at the same time (fig. 2). A number of indexes such as prefix, suffix, regular expression, and inflected form are available.


DicoPro V2.0 - MLIS Project - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://dicopro.unige.ch/DicoProPublic/
Instant Message Internet Nouveautés Avoir Membres Marché
Word Exact
Remote DIS at: http://dicopro.unige.ch/public/DIS
Test DicoPro DIS Service
Italian-English (Letter A, HarperCollins)
English-Italian (Letter A, Collins)
Langue française (Lettre A, Hachette)
Webster 1913
Connecter
Document: Done
Start
DicoPro V2.0 - M...
Java Console
9:34 AM

DicoPro 2000

ISSCO (c) University of Geneva
Dominique Petitpierre - Gilbert ROBERT - Derek Walker

ISSCO

Dictionary Information Service

http://dicopro.unige.ch/public/DIS

This demonstration DIS server is located at ISSCO, University of Geneva, Switzerland

Before getting started, it is probably a good idea to look at the DicoPro user manual.

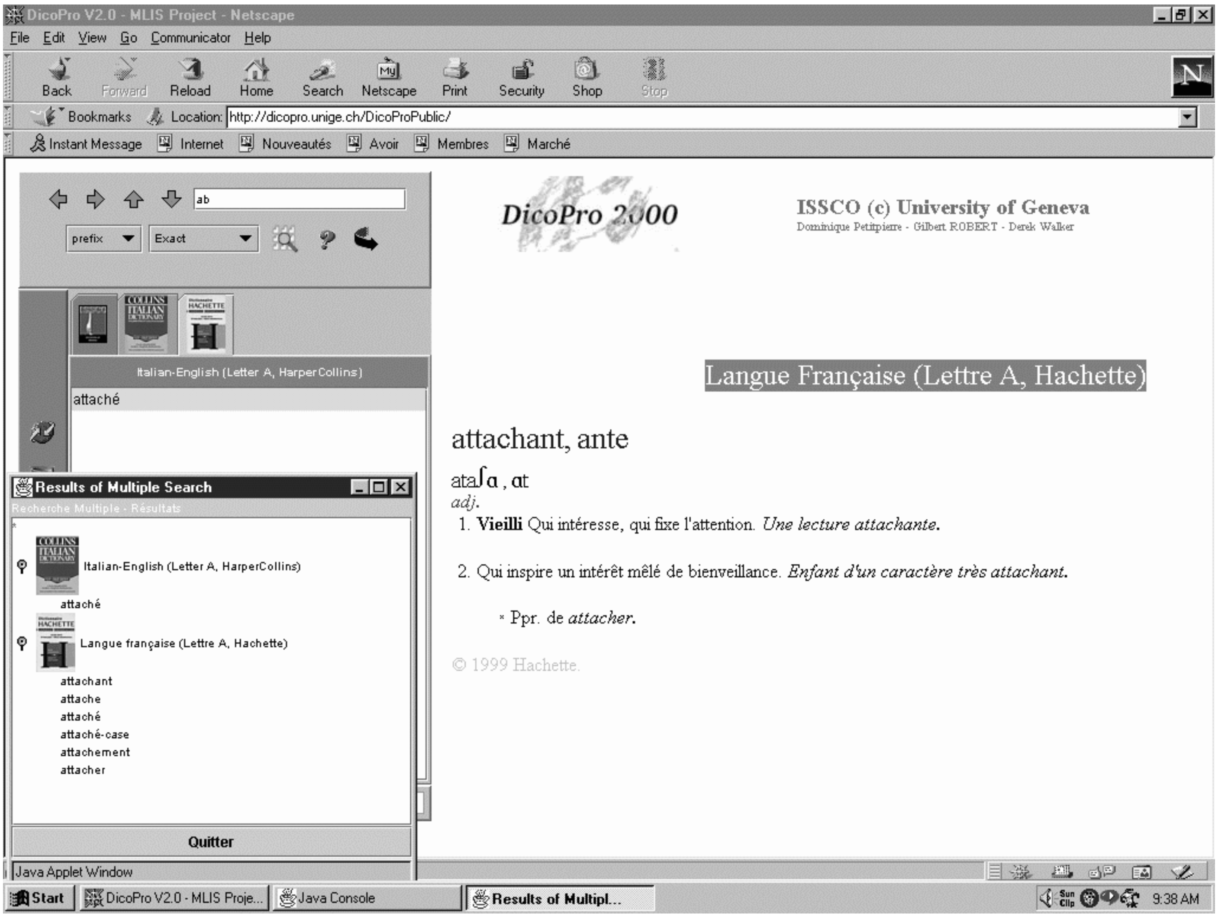
Dictionaries currently available on this server

Please read the copyright information

- HarperCollins English-Italian Dictionary (Letter 'A')
  This dictionary is provided from HarperCollins for demonstration purposes.
- HarperCollins Italian-English Dictionary (Letter 'A')
  This dictionary is provided from HarperCollins for demonstration purposes.
- Hachette Dictionnaire de la langue française (Lettre 'A')
  Please note: The layout of the entries has not been endorsed by Hachette

Figure 1: DicoPro Interface

The client software enables simultaneous access to multiple lexical resources from a diversity of well-respected publishers. It provides a uniform interface allowing parallel queries in multiple dictionaries, regardless of the actual physical location of the resource. Each user (or user group) can select the set of dictionaries to be consulted.


DicoPro V2.0 - MLIS Project - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location:
Instant Message Internet Nouveautés Avoir Membres Marché
ab
prefix
Exact
DicoPro 2000
ISSCO (c) University of Geneva
Dominique Petitpierre - Gilbert ROBERT - Derek Walker
Java Applet Window
Start DicoPro V2.0 - MLIS Proje... Java Console Results of Multipl...
9:38 AM

http://dicopro.unige.ch/DicoProPublic/

Italian-English (Letter A, HarperCollins)

attaché

Results of Multiple Search

Recherche Multiple - Résultats

Italian-English (Letter A, HarperCollins)

attaché

Langue française (Lettre A, Hachette)

attachant
attache
attaché
attaché-case
attachement
attacher

Quitter

Langue Française (Lettre A, Hachette)

attachant, ante

ataʃɑ , ɑt

adj.

1. **Vieilli** Qui intéresse, qui fixe l'attention. *Une lecture attachante.*

2. Qui inspire un intérêt mêlé de bienveillance. *Enfant d'un caractère très attachant.*

* Ppr. de *attacher.*

© 1999 Hachette.

Figure 2: Accessing multiple lexical resources

3 Licensing schemes

Offering access to multiple dictionaries online for a fee over the internet is a relatively new development for which no standard models exist. The DicoPro project devoted considerable time to investigating the issues related to this topic [Petitpierre/Murphy 1999]. Two primary licensing options have been implemented. Each dictionary can have its own licensing scheme.

One licensing option is an unlimited license for a certain number of users. This scheme is intended for users who have a constant need for either the technical or general dictionaries, and are prepared to pay a high price for the initial license fee in order to have unlimited use. The only restriction is on the number of concurrent users. This is a traditional way of licensing software, and is already more user-friendly than many software packages which are tied to specific workstations.

A second implemented licensing option is a time limited license for a certain (or unlimited) number of users. This option allows frequent users to budget a more manageable price per month, per several months or per year.

As mentioned in Section 2.3, the license parameters are protected by a number of security features to prevent subversion or tampering of the license restrictions.

4 Conclusion

The DicoPro consortium developed what is anticipated to be a commercially viable tool based on existing open standards. With the tools written in Java and with HTML, SGML and XML as standards for data mark-up and display purposes, the system is assured to be flexible and platform independent. The client software enables simultaneous access to multiple lexical resources from diverse publishers. It provides a uniform interface allowing queries across multiple dictionaries, regardless of the actual physical location of the resource. The licensing schemes available allow the data providers to stipulate the cost and access conditions while providing secure interaction mechanisms for site and user validation.

Notes

[1] The project was funded by the European Union and the Swiss Federal Office of Science and Education. For a full list of partners, detailed project reports and an on-line demo cf. `http://www.issco.unige.ch`.

References

[1] Petitpierre, D. and Murphy, D. (1997). Proposals and Specifications for Licensing Schemes. DicoPro Project Report D4.2., ISSCO, Geneva.

[2] Robert, G. and Petitpierre, D. (1997). Dico: un outil de consultation de dictionnaire en réseau., in META, XLII, 2, pp. 283-290.

[3] Walker, D., Petitpierre, D. and Armstrong, S. (2000) XMLTrans: A Java-based XML Transformation Language for Structured Data. To be presented at COLING 2000, Saarbrücken.