

*The Data Center*

## **Multi-Lingual Display of Business Documents**

David L. Brock, Edmund W. Schuster, and Chutima Thumratranapruk

**The Data Center**, Massachusetts Institute of Technology, Building 35, Room 212, Cambridge, MA 02139-4307, USA

### **ABSTRACT**

This article introduces a simple approach to multi-lingual display of documents. Using keyed words from the M Language Dictionary, a detailed mapping of words and phrases from different human languages becomes possible. Employing the “wiki” process, keyed words can be added, changed, or deleted, while still maintaining a complete history of all dictionary transactions. Once issued as a Data Center release, the M Language Dictionary and associated language mappings will be available through a standard web interface and various web services. This proposed approach to multi-lingual display has a great deal of practical value for global businesses, specifically in the area of supply chain management.

## ABOUT THE AUTHORS

**David L. Brock** is Founder and Director of the MIT Data center and Principal Research Scientist at the Massachusetts Institute of Technology, as well as co-founder and a former Director at the Auto-ID Center (now EPCGlobal, Inc. and Auto-ID Laboratories). David is also Assistant Research Professor of Surgery at Tufts University Medical School and Founder and Chief Technology Officer of endoVia Medical, Inc., a manufacturer of computer controlled medical devices. Dr. Brock holds bachelors' degrees in theoretical mathematics and mechanical engineering, as well as master and Ph.D. Degrees, from MIT. Dave can be reached at [dlb@mit.edu](mailto:dlb@mit.edu)

**Edmund W. Schuster** is Co-Director and research engineer at the MIT Center Data Center. His interests are the application of models to logistical and planning problems experienced in industry. He has a bachelor of science from The Ohio State University and a master's in public administration from Gannon University with an emphasis in management science. Ed also attended the executive development program for physical distribution managers at the University of Tennessee and holds several professional certifications. Ed can be reached at [edmund\\_w@mit.edu](mailto:edmund_w@mit.edu).

**Chutima (Pom) Thumratranapruk** is a second year MBA student at the MIT Sloan School of Management. She has a Bachelor of Engineering in Industrial Engineering with Honors from Chulalongkorn University, Bangkok, Thailand. Pom has work experience with Michelin in Shanghai and is founder of a manufacturing company in Thailand. She can be reached at [chutima@mit.edu](mailto:chutima@mit.edu)



## 1.0 INTRODUCTION

Businesses are increasingly becoming involved in international operations where speed and accuracy is essential in communication. Trends in foreign trade, global outsourcing, multi-national supply chains, and international labor forces located in many different countries are creating a dynamic and tightly inter-dependent worldwide business environment. Given these developments, many documents associated with supply chain processes, such as purchase orders, invoices, and sales receipts, now appear in digitized form. This allows for almost instant transmission anywhere in the world.

In addition to the globalization of business, the volume of data and information companies must manage is growing exponentially. By one estimate, the amount of digital information organizations maintain as part of everyday data processing operations grows at 40 to 60 percent per year (Park 2004). The Internet revolution, along with the movement toward the eXtensible Markup Language (XML), web services, and service-oriented architectures (SOA), means that the growth in digital information will continue at an even greater rate (W3C 2006).

While English is the established language of international trade, the new world of internet connectivity means many more local workers will be required to participate in the global communication network. This means that the issue of language translation is far more important than it ever has been. Of particular interest is the ability to translate digital documents *automatically* from one language to another as a way of reducing errors in interpretation.

While rapid and accurate machine translation of words, phrases, and sentences would have a wide-ranging impact on business and society, this article presents a simpler method for display of business documents in multiple human languages.

The core of this approach involves a simple keyed word mapping using the M Language Dictionary, followed by decoding of business documents into native languages (Brock 2005). This is in contrast to translation from one language to another via various computer programs that depend on Artificial Intelligence (AI) to establish the context of words, phrases, and sentences.

The next section of this article discusses how the proposed method for using the M Language contrasts with traditional machine language translation. After this discussion, the following sections outline the approach to multi-lingual display of business documents, including the generation and use of language “mappings.” The final section presents an actual implementation and software solution that demonstrates the near-term business value.

## 2.0 SIMPLE LANGUAGE DISPLAY VERSUS TRANSLATION

Historically, various types of AI programs have attempted to translate business documents automatically from raw text (Dorr 1999). The fundamental problem with this approach is that translation programs must infer the meaning (semantics) of words, phrases, and sentences with little idea of the context in which these appear. Further, AI translation



programs require a deep knowledge of the basic mental heuristics that humans use to comprehend the meaning of written language. This “common sense” reasoning ability, often the result of many years of learning, is far more complex than any AI researchers had ever imaged (Lenat and Guha 1990). For these reasons, the automatic translation of business documents is currently not possible to accomplish in practice with any degree of accuracy or reliability.

The approach presented in this article takes a different and far more practical direction, and involves an innovative application of the M Language. This approach relies upon the fact that keyed words contained in the M Language Dictionary have only one meaning (Brock, Schuster, Kutz 2006). These keyed words are “mapped” into other languages by manually entering their equivalents into the M Language Dictionary. Using new methods for internet collaboration and social networking, the global online community plays an important role in achieving the mapping of words from different languages.

The remaining sections of this article describe the details of this approach along with observations concerning practical implementation.

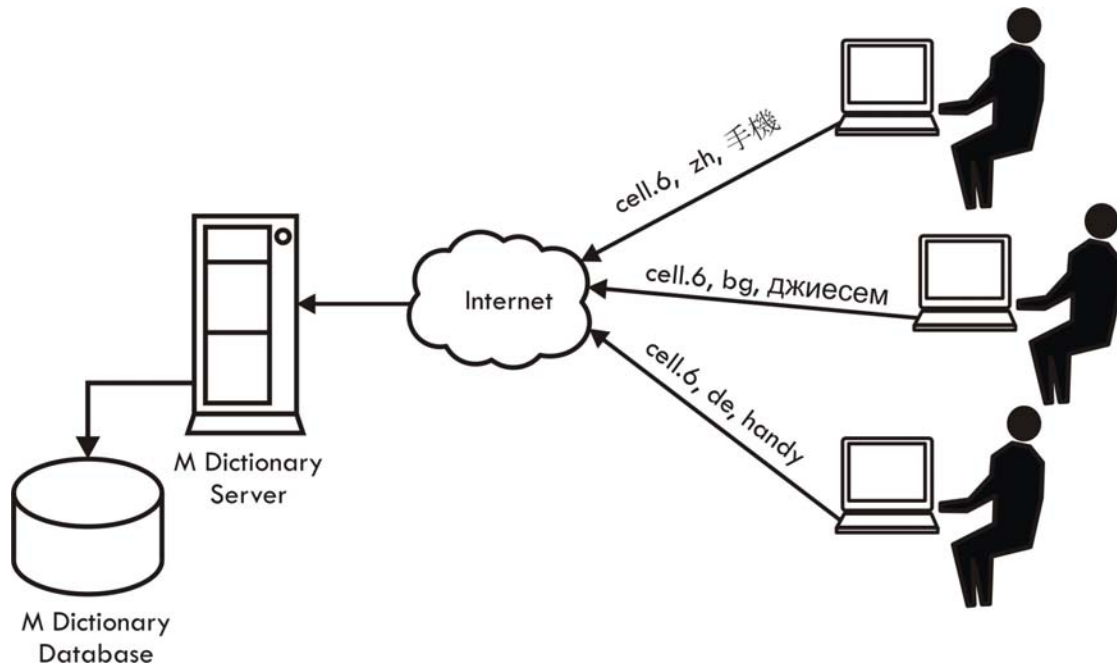
### **3.0 BUILDING THE LANGUAGE MAPPINGS**

In the process of building the M Language, users have the ability to contribute language mappings for each of the keyed terms contained in the dictionary.

Since every term in the M Language has only one definition, it is easy to provide a representation of that term in a human language. This human language representation need not be a single word, but may be a phrase or short description.

During Dictionary construction, users from anywhere on the internet can add a language mapping to a term by first entering a language code, describe by the two or three letter ISO 639 code (part of ISO2002), and then a character string representing the actual “translation,” as illustrated in Figure 1.

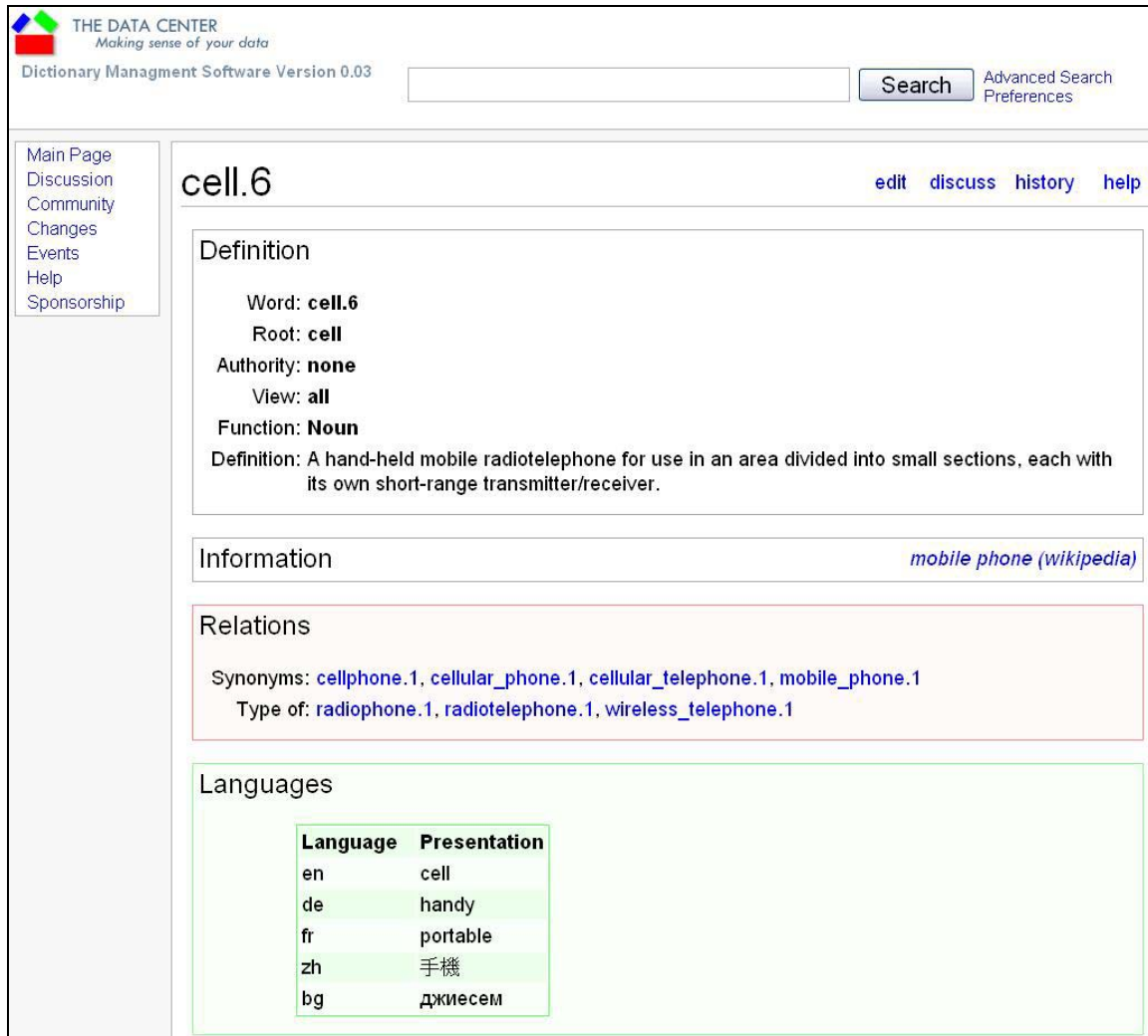




**Figure 1** - Users on the internet can contribute language mappings for any of the terms in the M Language dictionary.

As an example, the resulting M dictionary entry represented in Figure 1 for cell.6 (cellular telephone) might appear as shown in Figure 2. The “language” section of the entry contains representations for a variety of languages. These mappings, upon final release of the dictionary, will provide the means by which M documents are displayed in multiple languages.





THE DATA CENTER  
Making sense of your data  
Dictionary Management Software Version 0.03

Search [Advanced Search](#) [Preferences](#)

Main Page  
Discussion  
Community  
Changes  
Events  
Help  
Sponsorship

## cell.6

[edit](#) [discuss](#) [history](#) [help](#)

### Definition

Word: **cell.6**  
 Root: **cell**  
 Authority: **none**  
 View: **all**  
 Function: **Noun**  
 Definition: A hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver.

### Information

[mobile phone \(wikipedia\)](#)

### Relations

Synonyms: [cellphone.1](#), [cellular\\_phone.1](#), [cellular\\_telephone.1](#), [mobile\\_phone.1](#)  
 Type of: [radiophone.1](#), [radiotelephone.1](#), [wireless\\_telephone.1](#)

### Languages

Language	Presentation
en	cell
de	handy
fr	portable
zh	手機
bg	джисем

**Figure 2** - The M Language dictionary contains presentations of keyed terms in various human languages as contributed by users.

#### 4.0 USING THE LANGUAGE MAPPINGS

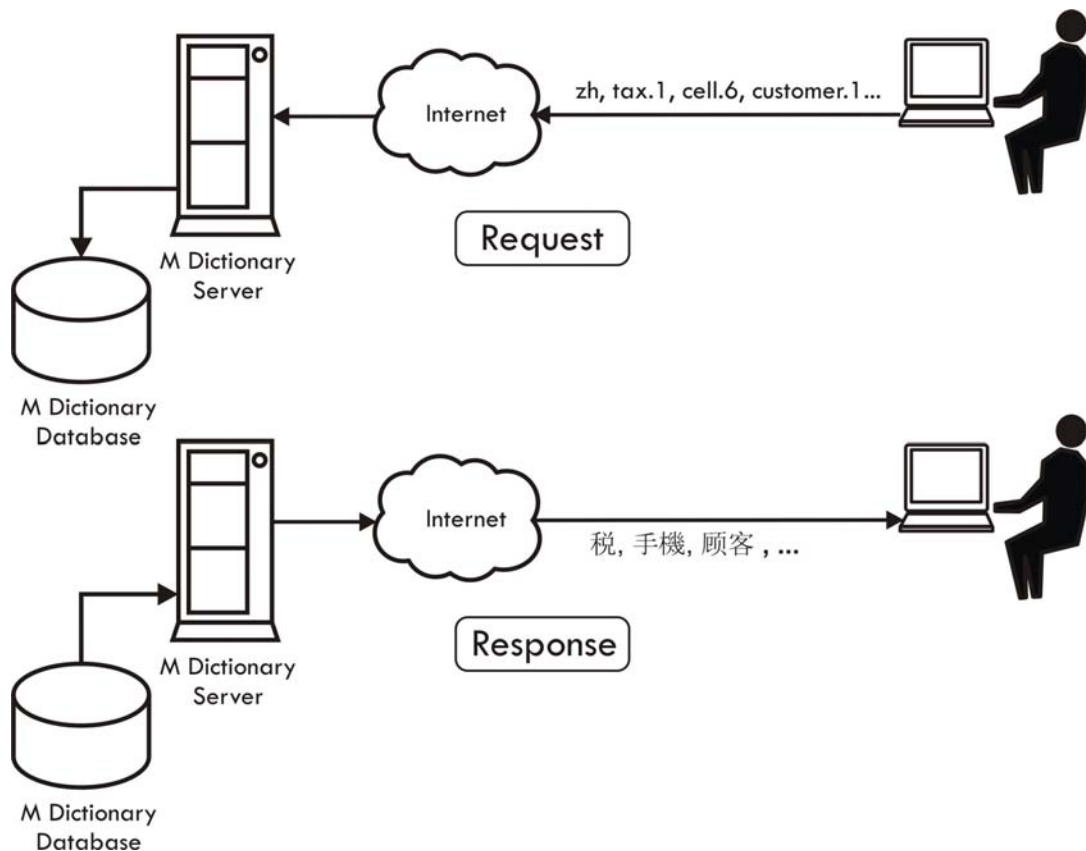
To keep things organized, the dictionary is developed using an open “wiki” process where users freely add and edit entries using a web browser (wikipedia 2006). All changes, including time, date, and the person responsible for entering the word or phrase are recorded and logged. Periodically, the process of entering words will be suspended. The resulting dictionary will then be subjected to a thorough review where the data will be checked and validated. Upon completion of this editing and review process, the dictionary will be released as a recommended vocabulary. After release, the dictionary is then open for use by various web clients and end-user applications through a standard Web interface as well as various web services.



Some of the web services that have been implemented include access to entry data based on a specific keyed term, connections (if any) between two terms based on a pre-specified relation, and the representation of a keyed term in a specified language.

Figure 3 illustrates the data flow for this later web service, in which one or more terms are sent to the server along with a specified language (represented by a two or three character ISO 639 code).

The specific web service protocols available for the M Dictionary include Simple Object Access Protocol (SOAP), as specified by the Web Services Description Language (WSDL) (W3C SOAP 2003, W3C WSDL 2006). Dictionary services are also provided using the much simpler Representation State Transfer (REST) methodology (Fielding 2000). Either of these methods may be employed by browser based Asynchronous JavaScript and XML (AJAX) and standalone client-side applications (Garrett 2005).



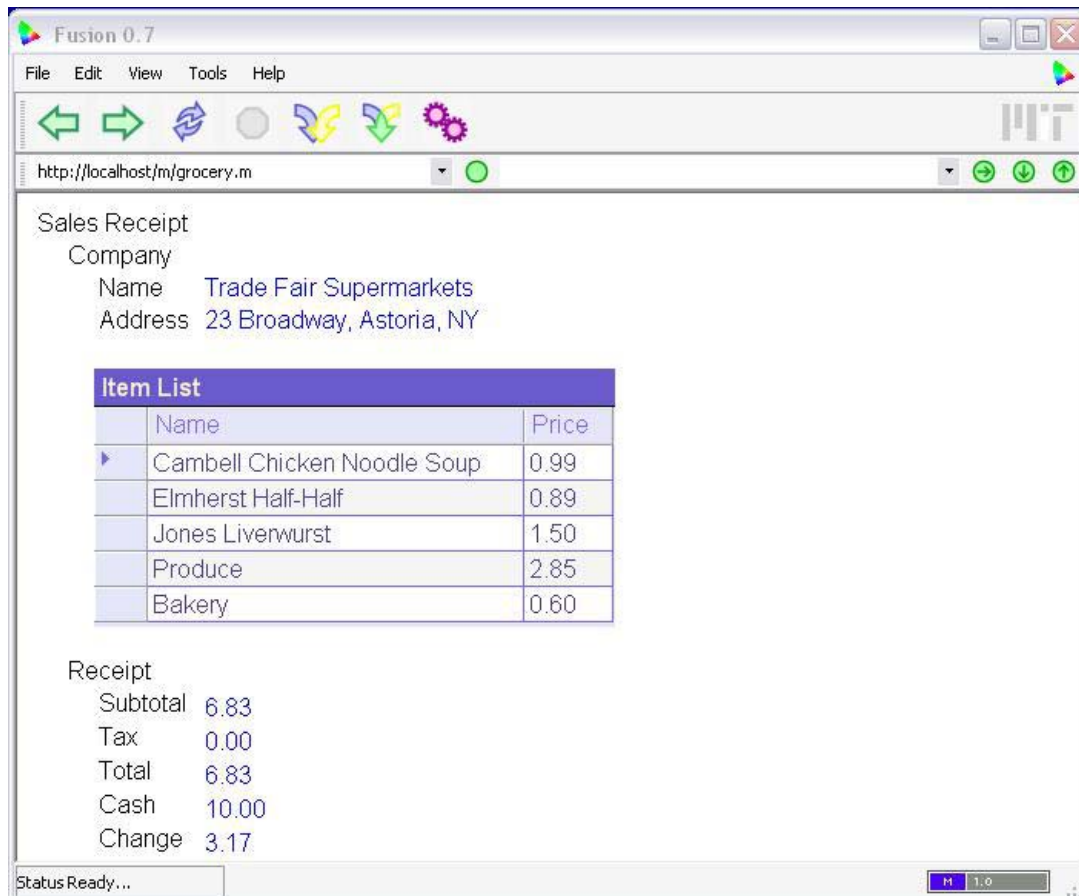
**Figure 3** - The M Language dictionary contains presentations of keyed terms in various human languages as contributed by users.

## 5.0 "FUSION" – AN M BROWSER

As part of the M Language development, the MIT Data Center is building a reference client-side application that allows M Language documents to be viewed and manipulated. Code named "Fusion," this M "browser" uses the rules of the language to format the data



and the words of the language to perform “intelligent” search, validate data, merge documents, and apply algorithms, as shown in Figure 4.



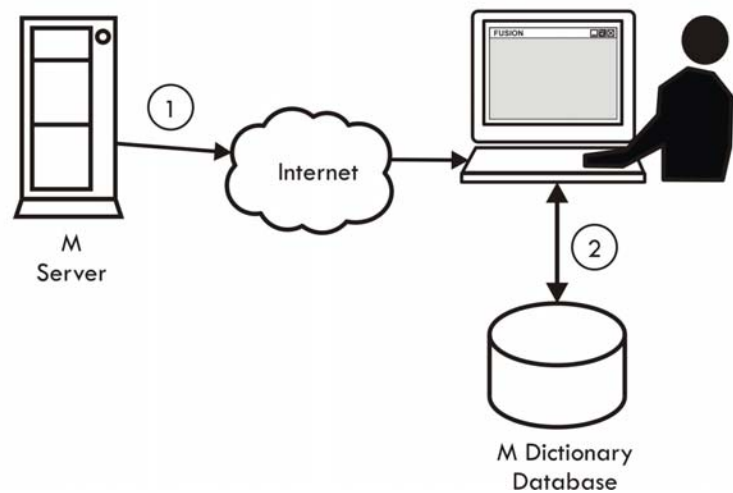
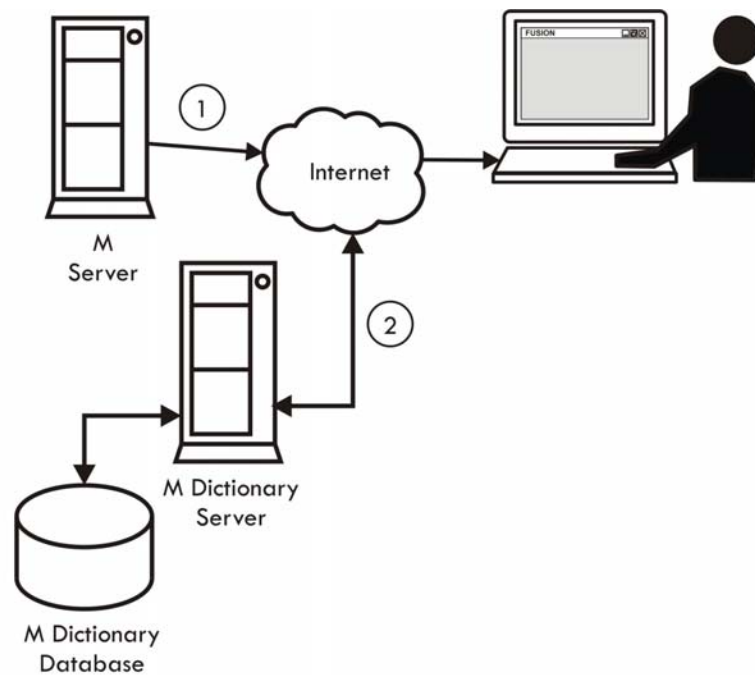
**Figure 4** - The Data Centers reference M browser uses the language rules to format the data and the words to search, validate, merge, and manipulate documents.

The Fusion Browser accesses one or more M documents over the network using standard internet protocols (Brock 2005). All terms in the documents are extracted by the browser and sent to a remote M Dictionary server or a local copy of the dictionary to recover at least some of the information associated with those terms, as shown in Figure 5. These data may be recovered automatically or in response to a user query.

In the case of language display, the user may select from any of a number of languages. The browser then passes this selection to the remote or local copy of the dictionary to recover the specified language representations of those terms, as illustrated previously in Figure 3. An actual display of the prototype Fusion browser is shown in Figure 6 for both English and Mandarin (Simplified Chinese) representations. The Mandarin words were entered into the Dictionary by Zhang Qinli of Hong Kong University of Science and Technology.

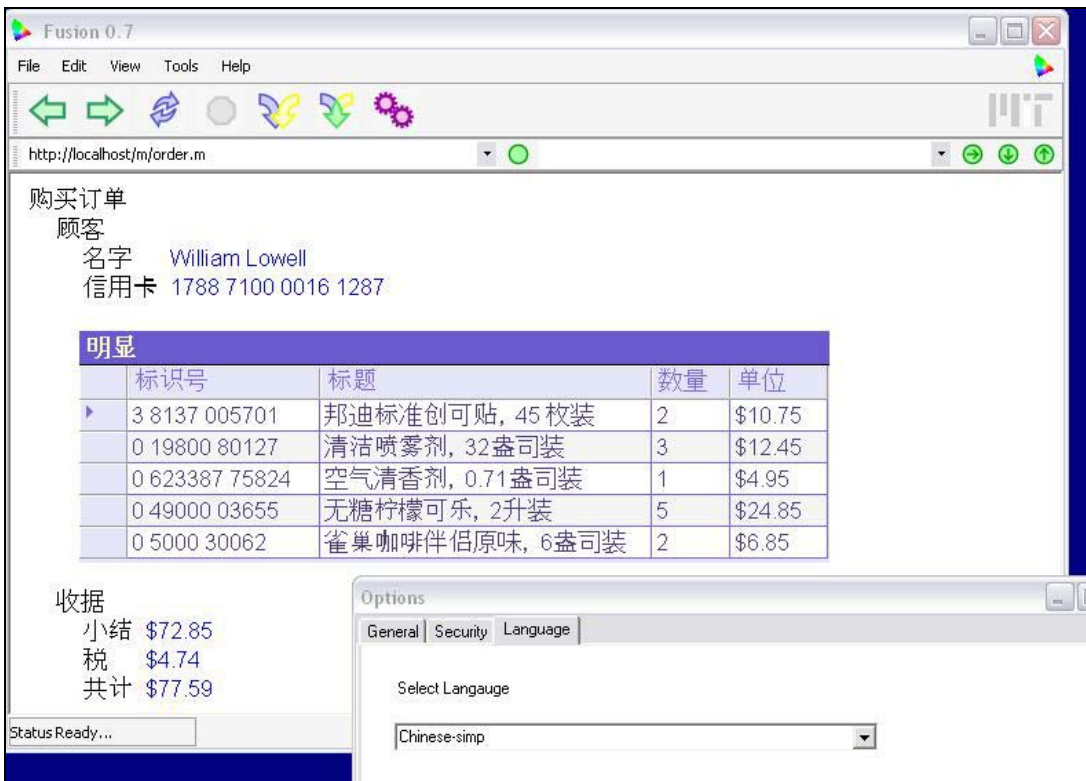






**Figure 5** - The M browser extracts all the terms and sends them to a remote M dictionary server (above) or a local copy of the dictionary (below) to recover all associate information about those terms.





**Figure 6** - The M browser can display terms in any of a number of language depending on the mappings supplied by the dictionary.



## 6.0 CONCLUSION

The method presented here is a simple, yet practical approach to business document display in multiple languages. The approach depends on large-scale voluntary contribution along with international collaboration. To get the process started, the MIT Data Center is working with the MIT Sloan School of Management to organize a group of students who will in turn lead teams responsible for making dictionary entries and keyed word mappings. As this effort gains momentum across the Internet, rapid growth is expected from a host of parties interested in using the M Language for multi-lingual display of documents.

## 7.0 ABOUT THE MIT DATA CENTER

The MIT Data Center is currently looking for industrial sponsors to participate in this exciting prototyping process involving human language representation. As a sponsor of the Center, individual companies will have the first opportunity to apply the M Language to supply chain processes. Besides language translation, there are a number of other benefits to participating in the consortium, which already includes MorganFranklin Corporation, LG, Raytheon, and Siemens.

Those interested should contact David Brock [dlb@mit.edu](mailto:dlb@mit.edu) or Ed Schuster at [Edmund\\_w@mit.edu](mailto:Edmund_w@mit.edu)



## REFERENCES

Dorr, B. P. Jordan, and J. Benoit (1999), "A survey of current paradigms in machine translation," *Advances in Computers*, Vol 49, M. Zelkowitz (Ed), Academic Press, London, pp. 1—68.

Brock, D.L., (2005), "Data Center Presentation," *Key-Note Speech Halliburton Annual Meeting*, Houston, TX, October.

Brock, D.L., (2005), "Data Center Presentation," *Engineering Marketing Science*, December.

Brock, D.L., Schuster, E.W., and Kutz, Sr., T.J. (2006), "An Overview of the M Language," MIT-DATACENTER-WH-009, January 2006.

Fielding, Roy T. (2000) [Architectural styles and the design of network-based software architectures](#). *PhD Thesis*, University of California, Irvine.

Garrett, Jesse James (2005), "Ajax: A New Approach to Web Applications," *Adaptive Path Journal*.

International Organization for Standardization (ISO) (2002), ISO 639 Representation of Names of Languages.

Lenat, D. B. and R. V. Guha (1990), *Building Large Knowledge Based Systems*. Reading, Massachusetts: Addison Wesley, 1990.

Park, Andrew (2004), "Can EMC Find Growth Beyond Hardware?," *BusinessWeek*, November 1.

Wikipedia, and Wiki (2006), [http://en.wikipedia.org/wiki/Wiki\\_2006](http://en.wikipedia.org/wiki/Wiki_2006).

World Wide Web Consortium (2003), Simple Object Access Protocol (SOAP), <http://www.w3.org/TR/2003/REC-soap12-part0-20030624/>.

World Wide Web Consortium (2003), Web Services Description Language (WSDL) <http://www.w3.org/TR/2006/WD-wsdl20-rdf-20060518/>.

World Wide Web Consortium – W3C (2006), <http://www.w3c.org>.



## NOTES

