The Data Center

AN OVERVIEW OF THE M LANGUAGE

David L. Brock, Edmund W. Schuster and Timothy J. Kutz, Sr.

The Data Center, Massachusetts Institute of Technology, Building 35, Room 234, Cambridge, MA 02139-4307, USA

ABSTRACT

The MIT Data Center is moving into a new stage of development and application of the M Language. This paper gives an overview of the M Language along with applications. Designed to enhance data interoperability, the M Language serves as a base for the intelligent information infrastructure of the future.

ABOUT THE AUTHORS

David L. Brock is Principal Research Scientist at the Massachusetts Institute of Technology, and co-founder and a Director at the Auto-ID Center (now EPCGlobal, Inc. and Auto-ID Laboratories). The Center was an international research consortium formed as a partnership among more than 100 global companies and five leading research universities. 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**

Edmund W. Schuster has held the appointment of Director, Affiliates Program in Logistics at the MIT Center for Transportation and Logistics and is currently working at The Data Center as Co-Director - Administration and researcher. 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 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**.

Timothy J. Kutz, Sr. is the Director of the Information Technology Management practice at MorganFranklin Corporation (MFC) and has responsibility for MFC's partnership with the MIT Data Center. His information technology and management consulting experience includes both domestic and international engagements serving a variety of federal, state, and local governmental clients and fortune 500 commercial clients. He has a bachelor of arts in business administration from Marymount University and an MBA from George Mason University. He also holds a master's certificate in Project Management from George Washington University. Timothy can be reached at **Timothy.Kutz@Morgan-Franklin.com**.

1.0 INTRODUCTION

The MIT Data Center envisions a world in which information flows freely within and across the enterprise, where data from widely divergent sources merge seamlessly into a coherent whole, and where algorithms and software automatically combine with data to form a new *intelligent information infrastructure*.[1,2] By creating an open, global language that communicates between propriety schemas, companies will have the ability to combine, visualize and understand data.[3,4] This paper outlines the vision, approach, application, and benefits of this new initiative.

2.0 M – THE BASICS

The M Language is conceptually simple, consisting of two parts – *words* and *rules*. In M, words take on a new form that allows for easier machine understanding. Rules provide guidelines about how to place words together for representing data or models in a common format that is interoperable. These representations are in the form of messages that can be transferred between computing systems. The next two sections take a closer look at the words and rules of the M Language.

2.1 Words

The words used in the M Language are slightly different from English words. In M, every word has only one definition. This is an extremely important characteristic because computers that communicate using M do not need to understand the context or usage of a word to know its meaning.

English words are ambiguous. For example, the word "cell" might mean "cellular phone," "biological cell," "jail cell," or "fuel cell." Without some idea of the context, it is impossible to know the meaning of the word "cell."

To overcome this issue, the M language includes a number to denote individual words such as:

cell.1

To account for multiple definitions, the M Language allows numeric extensions, one for each definition. Thus, cell.1 is a word in M and cell.2 is a different word. With this method, every word has one and only one meaning.

In English, dictionaries define the meaning of a word. M also uses a dictionary. The M Dictionary serves as a repository for definitions of words used in computer transactions. The dictionary also is a means of storing other important information associated with a

particular word. This provides an effective means of unifying various aspects of a word and forms a base for common computer-to-computer communication.

2.1.1 The M Dictionary

In the M-Dictionary, words and definitions are stored in the following form:

cell.1 -	The basic structural and functional unit of all organisms; they may exist as independent units of life (as in monads) or may form colonies or tissues (as in higher plants and animals).

This, of course, is the only definition for the word cell.1. Other words, such as cell.2, cell.3, and cell.4 all have different definitions expressed using the same format.

In addition to the definition, the dictionary entry also contains three other pieces of important information. These include (1) word relations, (2) data format, and (3) language translations. This information helps in forming and understanding messages composed in M.

Word relations are simply the connections between words. These relationships include *synonyms*, *antonyms*, *types,* and *parts*.[5] Synonyms and antonyms are the same as in English.

Types refer to word generalizations. For example, automobile.1 is a *type of* motor_vehicle.1.[6]

Parts are words that are components of another word. This is often the case when thinking about physical objects, although this could also be the case with abstractions. For example, a wing.4 is a *part of* airplane.1.

Data format provides guidance concerning the forms and patterns of data values that might be associated with a particular word. In many situations, computer-to-computer communication might contain a word such as first_name.1 that has an associated data value such as "John Smith". Other common situations include words like telephone_number.1, account_balance.1, or postal_code.1. In all of these cases, a particular format or pattern of data is attached to the word.

Finally, the *language translation* portion is simply the representation of the word in M in a word (or phrase) in a human language. In most situations, computer-based language translation is very difficult because of a lack of context for the specific communication. Since in M each word has only one definition, the word cell.1 (biological), for example, cannot be confused with cell.2 (telephone). Words with a single definition allow users to specify exact meaning independent of context. This eliminates ambiguity in translation.

2.1.2 Dictionary Development

Developing common definitions for the words, data formats, and translations used in commerce along with the analysis of data across all industries has traditionally been a

source of great debate within business. To build a robust global dictionary containing the words used in the M Language along with other important information, such as relations between words, associated data patterns and translations, requires a different approach.

The "wiki" process has emerged as an innovative application of Internet technology to knowledge management and consensus building. A 'wiki' is a type of website that allows users to add and edit content and is especially suited to collaborative authoring.[7]

Since 2001, Wikipedia has become the largest encyclopedia ever created with over 3 million entries.[8] The M Dictionary uses the wiki approach with several important modifications including improved security through user registration, maintenance of the integrity of word relations, a monitoring function to reduce the chances of near identical definitions, and administrative controls to ensure accuracy.

However, having a robust dictionary is just a part of the M Language. To form messages, computers need a set of rules that give instructions on how to glue the words together. The next section discusses the rules of the M Language.

2.2 Rules

Language is more than just a collection of words defined by a dictionary. For most languages, grammar gives explicit rules on the order of words to give meaning to a sentence. In English, the simple sentence "Threw ball the Jack" is nonsensical. Establishing correct word order is essential. Thus the sentence "Jack threw the ball" formed by rearranging the words makes sense. From this example, it is clear that word order, sometimes called syntax, has an important role in communicating meaning. If words are in the correct order, instant recognition takes place.

Just as English has rules of grammar for word order, the M Language also has rules establishing the order needed for machine understanding of messages.

The initial version of the M Language contains three simple rules. These three rules, however, represent a significant portion of computer-to-computer communication. The three are (1) phrases, (2) key-value pairs, and (3) tables.

A phrase is a sequence of machine-understandable words representing a single idea. A phrase in M is just like a phrase in English. The syntax is such that the last word in the phrase is the root and all the others are modifiers. As an example, the phrase "initial account balance" appears in M as:

initial.1_account.1_balance.1

In this phrase, balance.1 is the root word, while initial.1 and account.1 are modifiers. Phrases within the M Language represent a unit of meaning that is extremely useful in increasing the precision of data element descriptions.

Key-value pairs are simply a list of words with associated data values. Tax forms, medical records, and financial statements are all representations of key-value pairs. A key-value pairs example follows:

```
                Name.1 – "John Smith"
Telephone_number.1 – "(703)-459-1234"
```

Key value pairs in the M Language are useful in making data interoperable within and external to the firm. Interoperable data opens a number of possibilities for combining data posted on the Internet with internal company data.

The final rule involves tables, which are the most common way to store data on the computer. There are many different ways to represent tables, comma separated values (CSV), Excel spreadsheets, HTML tables, and others. In the M Language, a table takes on the pattern of repeating sets of key value pairs, each with identical keys. The following is an example:

```
  patient.1
            name.1  – "John Smith"
telephone_number.1  – "(703)-459-1234"
  patient.1
            name.1  – "Robert Williams"
telephone_number.1  – "(703)-457-1234"
```

Subsequent versions of M will include additional rules, such as rules for spatial data, equations, and mathematical models. Integrating spatial information, for example, will be valuable in marketing, demographics, transportation, and logistics, while rules for mathematical equations and algorithm will ultimately allow the integration of data and models.

3.0 APPLICATION

The M Language is a tool that enables the free flow of data and models across the network. Achievement of this vision will result in a number of practical applications in industry.

3.1 Interoperable Data

Perhaps the most obvious application of the M Language is as an intermediary between proprietary data systems. In this application, data from one database is translated into M *before* it is communicated to another, as shown in Figure 1. Here data from a source system, is translated *at the server* into M. This data is sent over the network – either as human readable text or compact binary – to the target system. The data can then be

used in the native M format or translated again into another proprietary schema and stored in the local database.

In broad terms, the M Language serves as a common transport between distributed, incompatible data systems. The advantage of this approach is that data providers do not need to know the format or content of every possible target application. Providers need only expose data in the standard M language for their data to be interoperable. Using M as a common carrier, translation takes place only once at the server instead of many times for every possible consumer.

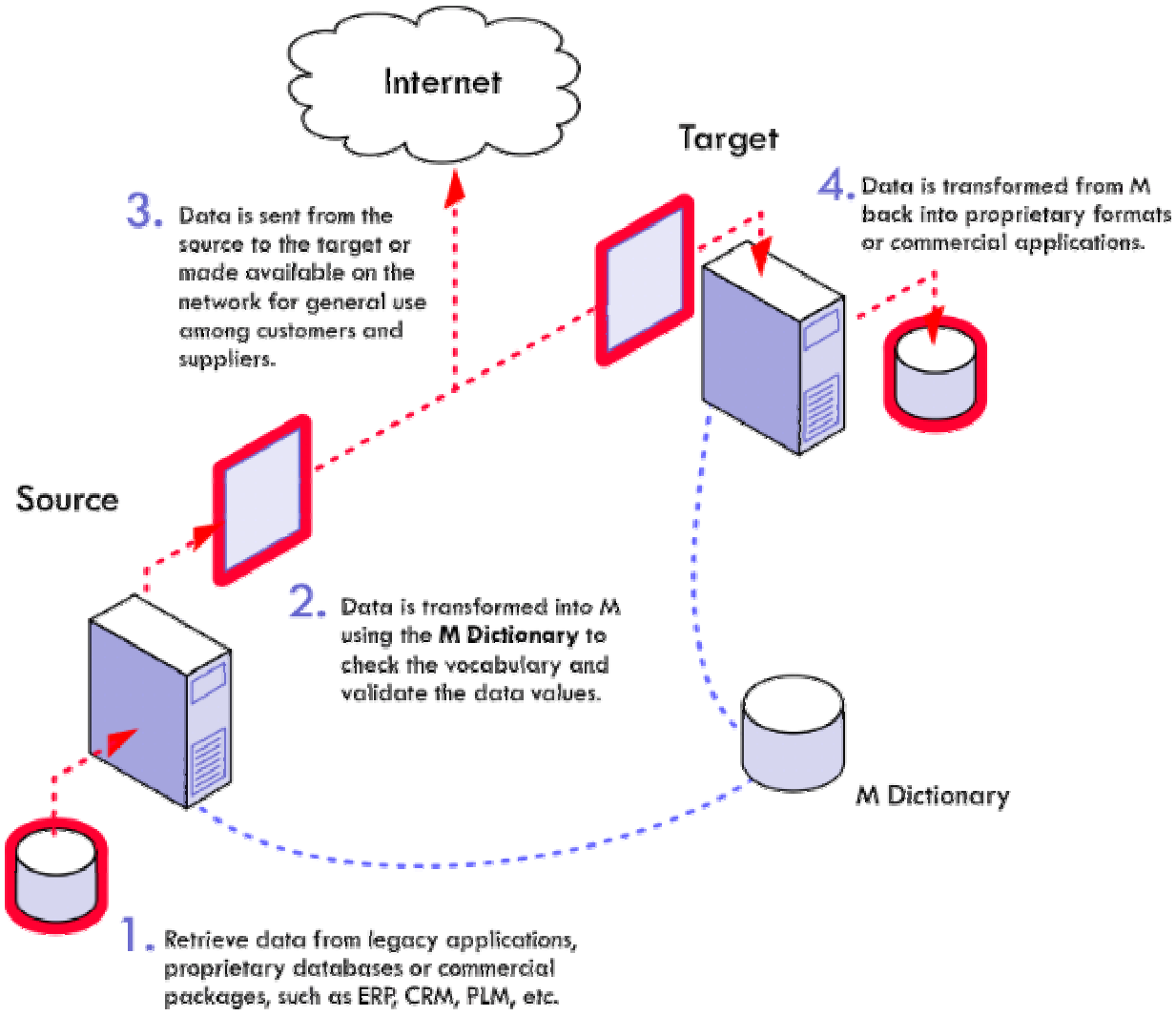

Figure 1. The M Language can serve as an intermediary between disparate, proprietary data systems, as well as a general interface for internet communication.

3.2 Browser

The M Language provides more than just a translation service. As a common language, M encourages third-party developers to create a wide range of software tools and applications. As a first step, the MIT Data Center has created a browser to view, edit, and manipulate data directly in M. The browser presents data in an easy-to-understand format *without* the need for additional styling information. Sections are indented, headings aligned, data color-keyed and tables displayed properly as spreadsheets. The application also provides data plotting functions, definition based search, data validation, language translation, and model integration. Data plotting functions refers simply to charts and graphs common in spreadsheet applications. The other features, however, are novel and are described in more detail below.

3.2.1 Definition Based Search

Since words in M have only one meaning, a key-word search in M yields matches based on the definition, not on the character string. This allows for search based on meaning rather than keyword. Further, the dictionary proposed for M provides even more powerful search capabilities as compared with current approaches.

Every definition in the M Dictionary includes provision for word relations; that is relationships between words based on *types* and *parts*. From our previous example, automobile.1 is a *type-of* motor_vehicle.1 and wing.4 is a *part-of* an airplane.1.

Combined with definition-based search, these relations provide a powerful tool. Search can span types of a word. For example, a query for types-of flower.2 would return rose.4, violet.3, or marigold.1. A search for parts-of automobile.1 would return fender.1, muffler.3, or engine.1.

Searching on a definition and using the word relations from the dictionary transform search on the internet from frustration to production. In addition, keyed words and phrases from the dictionary can be used in traditional HTML code for web pages as meta-tags to significantly increase the accuracy and performance of web-based search tools such as Google, MSN Search, and Yahoo.

3.2.2 Data Validation and Quality

Given the rapidly increasing amounts of data available within business, data integrity takes on ever greater significance. In addition to definitions, the M dictionary also specifies the structure and format of any attached data value.

For example, the entry telephone_number.1 specifies that data values should conform to those of a specific country or internationally recognized patterns. Thus, an entry (617) 258-123 would be recognized as incorrect since United States domestic telephone numbers require a four digit extension. Likewise, the entry +128(0) 2522 1111 is also improper because a country code of 128 does not exist.

This capability provides an opportunity for data sharing and information validation within and across the enterprise, with the goal of improving data quality.

3.2.3 Language Translation

Another benefit of the M Language is translation to different languages. Since words in M have only one meaning, translation to a human language is much easier. It is not necessary to use the context or meaning of a passage to discern the definition of a word. For example, cell.2 does not mean biological cell, fuel cell, or jail cell, it means cellular telephone. Thus cell.2 can be translated into "cell phone" in American English, "mobile" in British English, "handy" in German, "portable" in French, "手機" in Chinese, or "джиесем" in Bulgarian.

3.2.4 Model Integration

The most far-reaching feature of the M Language is the ability to share not only data across the network, but models. Mathematical models all contain three basic features – inputs, outputs, and an algorithm.

By describing the inputs and outputs in the M Language, a model can be linked to data *semantically*. In other words, because the data and interface to the model are described in the same language, they can be connected together using the words and word relations from the common dictionary.

Models can also be connected to one another. In this case, outputs from one model can be matched semantically to the inputs of another model, forming a model pair. Connecting more models in the same way can build up the simple pair into a larger *model network*. These model networks can then function as single units, operating on data and performing complex mathematical analysis.

4.0 CONCLUSION

The MIT Data Center is now in its initial stage of development. The concepts presented here have already been recognized as immediately beneficial in industry and global data exchange. We are beginning to work with our partners in industry, academics, trade bodies, and government (international and domestic) to craft the standard to have both short-term practical utility and long-term industry-wide adoption. The language and dictionary developed here will be open and distributed under an open license agreement. The M Language – and its associated tools and technologies – is the first step in the foundation of a new architecture for worldwide data exchange and information integration – and provides the groundwork for a *new* Intelligent Information Infrastructure.

REFERENCES

1. **Brock, D.L.**, "The Intelligent Data Network Proposal for Engineering the Next Generation of Distributed Data Modeling, Analysis and Prediction," *The MIT Data Center*, Cambridge, MA, MIT-DATACENTER-WH-001, April 2003.

2. **Brock, D.L.**, "The Data Center Vision - Making Sense of the Data," *The MIT Data Center*, Cambridge, MA, MIT-DATACENTER-WH-002, December 2004.

3. **Brock, D.L., Schuster, E.W., Allen, S.J. and Kar, P.**, "An Introduction to Semantic Modeling for Logistical Systems," *The MIT Data Center*, Cambridge, MA, MIT-DATACENTER-WH-003, December 2004. and *Journal of Business Logistics*, Vol. 26, No. 2, 2005.

4. **Schuster, E.W., Brock, D.L., Allen, S.J. and Kar, P.**, "Prototype Applications for Semantic Modeling," *The MIT Data Center*, Cambridge, MA, MIT-DATACENTER-WH-004, December 2004.

5. **WordNet®, Princeton University, http://wordnet.princeton.edu/**.

6. The M Dictionary allows 'compound' words such as motor_vehicle.1, though a legitimate question would be whether this should be a phrase using two words from the dictionary (i.e. motor.1_vehicle.1). In general there is no rule, but as a guideline compound words with unique meaning, such as 'public relations' or 'accounts receivable', or very common pairings, such as 'telephone number' should be 'compound' words in the dictionary (i.e. public_relations.1, accounts_receivable.1, and telephone_number.1). Conversely, phrases whose meanings can be inferred from their parts, such as 'red ball' or 'patient name', should be formed as phrases (i.e. red.1_ball.1 and patient.1_name.1). This allows complex communication without the resulting combinatorial explosion that would occur from storing all word sequences.

7. **Wiki, http://en.wikipedia.org/wiki/Wiki**.

8. **Wikipedia, http://en.wikipedia.org**

NOTES

