



Drill Down: Who Wrote the Books for DW?

by Elliot King

Richard Hackathorn clearly remembers the moment the central idea for his new book *Web Farming for the Data Warehouse: Exploiting Business Intelligence and Knowledge Management* (Morgan Kaufman Publishers, Inc.: San Francisco 1999) came to him. He was sitting through a tedious presentation at a conference for chief information officers in 1997 when he jotted on his notepad that "the Web is the mother of all data warehouses." Hackathorn, a former professor who worked on fundamental concepts of enterprise systems, database management, decision support and data warehousing, is a well-known industry innovator and consultant. He defines Web farming as the systematic refining of information resources on the Web for business intelligence. To achieve that goal, relevant content found on the Web must be refined into a form that is compatible with a data warehouse.

Although the Web is a dynamic and expansive information space, it is a database designer's worst nightmare, Hackathorn notes. It is a free-form combination of text, images and virtually any kind of information object for which somebody can develop a browser. It has no structure at all, just a series of links and pointers, many of which no longer work. Finding information on the Web, Hackathorn observes, is like trying to find a needle in a haystack while people are constantly adding and subtracting from the pile.

And while the Web has been productively used to distribute information from data warehouses to users for analysis, nobody has seriously addressed the possibility of using Web data as input for a data warehouse. Hackathorn makes just that argument in this book. While significant barriers must be overcome to make Web content suitable for a data warehouse, the benefits, he suggests, outweigh the costs.

Hackathorn has divided his book into four parts. Acknowledging the negative reaction many IT professionals have to even trying to incorporate Web content into a data warehouse, his first section is primarily motivational. He provides the business argument for Web farming -- noting that the efficient processes to turn data into information and then knowledge are essential to the well-being of any enterprise.

The second section of the book lays out a strategy for initiating Web farming efforts inside an enterprise. Adopting a structure first articulated in the 1970s by Richard Nolan, a professor at the Harvard Business School, Hackathorn lays out a four-step process. First, the business case based on the objectives and business environment of the enterprise must be made for Web farming. Then, the concept has to be accepted and an infrastructure built. Next, pipelines to users have to be established. Finally, the Web content must be structured for the warehouse.

The final two sections of the book look at some of the tools and sources of content available to implement Web farming projects -- there currently is no single solution for Web farming -- and the social and cultural ramifications of these efforts. In an extremely interesting section at the end of the book, Hackathorn proposes a code of ethics for Web farming.

Clearly, Hackathorn has a vision. As he notes, Web farming is not about technology or the Web. It is about basic business practices in the contemporary environment. Moreover, he has the certainty that a visionary needs. Indeed, he writes, "The development of Web farming is certain. It will become a standard function within data

warehousing systems as companies strive in desperation for their next competitive advantage." Doing business as usual, he writes, will no longer be a viable strategy.

If Richard Hackathorn has developed a provocative argument for the marriage of Web and data warehousing technology, then Dorian Pyle's *Data Preparation for Data Mining* (Morgan Kaufman Publishers, Inc.: San Francisco 1999) offers data modelers and other insiders an in-depth look at data preparation. Data mining tools have focussed almost exclusively on building models -- the aggregation of data points that could, when analyzed, yield useful information. But effective data preparation is essential for effective and cost efficient modeling.

A partner with Naviant Technology Solutions, a consulting company, Pyle argues that too frequently people discuss data mining applications only in terms of what different data mining tools can do. That, he says, is analogous to trying to determine what you can build because you have a power saw. You may not want anything you can build with a saw, so why bother?

Instead, Pyle says, data mining projects must start with a definition of the problem. After the problem is clarified, potential solutions must be explored. Next, implementation methods must be specified. Only at this point can data mining techniques be applied.

He divides data mining into three steps: First, data must be prepared; next, data must be surveyed; and then finally, data can be modeled.

Data preparation involves manipulating and transforming raw data so that the information content contained in the data set can be most effectively used. Appropriate data preparation depends on two factors -- the requirements of the solution and the data mining tool available.

According to Pyle, data preparation is routinely overlooked in most data mining projects. The majority of the book consists of a set of approaches to data preparation. The book also includes a CD-ROM with the software Pyle uses or describes to perform data preparation tasks.

Pyle's book is directed more specifically at the practitioner than Hackathorn's book, which lays out a vision for the marriage of the Web and the data warehouse. Both, however, attest to the vitality of the data mining/data warehousing arena. New vision, new approaches and new tools will enable the ongoing development of the discipline.

About the Author: *Elliot King is an Associate Professor of Communications at Loyola College in Maryland. He can be reached at (410) 356-3943, or by e-mail at eking@loyolanet.campuswix.net.*

[To November Table of Contents](#)

[to ESJ Home Page](#)