



---

## **Chapter IV**

# **Managing Document Taxonomies in Relational Databases**

Ido Millet  
Penn State Erie, USA

### **ABSTRACT**

*This chapter addresses the challenge of applying relational database technologies to manage taxonomies, which are commonly used to classify documents, knowledge and websites into a hierarchy of topics. It first describes how denormalizing the data model can facilitate data retrieval from such topic hierarchies. It then shows how the typical data maintenance difficulties associated with denormalized data models can be solved using database triggers.*

### **INTRODUCTION**

The need to maintain classification and retrieval mechanisms that rely on concept hierarchies is as old as language itself. Familiar examples include the Dewey decimal classification system used in libraries and the system for classifying life forms developed in the 1700s by Carolus Linnaeus. A more recent example is Yahoo's subject taxonomy.

Information technology has led to an explosive growth in digital documents, records, multi-media files and websites. To facilitate end-user access to these resources, topic hierarchies are frequently maintained to allow intuitive navigation and searching for resources related to specific categories. This chapter deals with the challenges of using relational database technology to maintain and facilitate queries against such topic hierarchies.

In a chapter written for another book (Millet, 2001), I discuss the generic issue of managing hierarchies in relational databases. This chapter focuses on applying these techniques to the specific needs of managing document, website and knowledge management taxonomies. For example, this domain is complicated by the typical need to classify a single document or website under multiple topics.

Relational databases and the current SQL standard are poorly suited to retrieval of hierarchical data. After demonstrating the problem, this chapter describes how two approaches to data denormalization can facilitate hierarchical data retrieval. Both approaches solve the problem of data retrieval, but as expected, come at the cost of difficult and potentially inconsistent data updates. This chapter then describes how we can address these update-related shortcomings via database triggers. Using a combination of denormalized data structure and triggers, we can have the best of both worlds: easy data retrieval and simple, consistent data updates.

## THE CHALLENGE

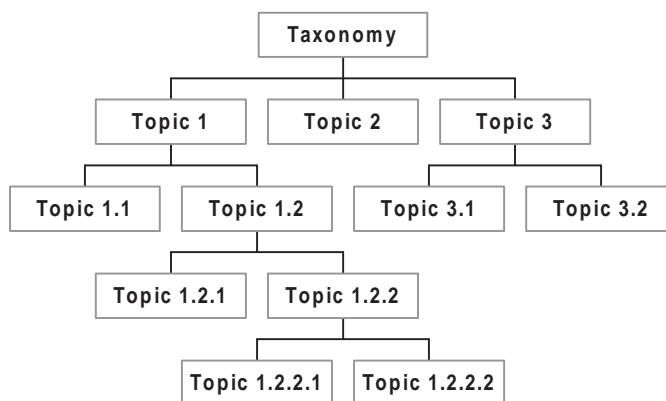
To demonstrate the data retrieval difficulties associated with topic hierarchies, consider a document database where each document is classified into a hierarchy of topics shown in Figure 1.

First, let us discuss how the topic hierarchy itself is stored in a relational database. Since each subtopic has at most one parent topic, we can implement this hierarchy via a recursive relationship. This means that each topic record maintains a foreign key pointing to the topic record above it. Figure 2 shows a data model for this situation. Note that the classify table allows us to assign a single document to multiple topics.

To demonstrate the difficulty of hierarchical data retrieval against the normalized data model in Figure 2, consider the following requests:

- Show a list of all Topics (at all levels) under Topic 1
- Show a list of all Documents (at all levels) under Topic 1
- Show how many Documents (at all levels) are classified under each Topic at Level 1 of the hierarchy

Figure 1. A Topic Hierarchy



15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/chapter/managing-document-taxonomies-relational-databases/9204](http://www.igi-global.com/chapter/managing-document-taxonomies-relational-databases/9204)

## Related Content

---

### Knowledge Management in Tourism

Daniel Xodoand Héctor Oscar Nigro (2005). *Encyclopedia of Database Technologies and Applications* (pp. 319-329).

[www.irma-international.org/chapter/knowledge-management-tourism/11167](http://www.irma-international.org/chapter/knowledge-management-tourism/11167)

### E-R Approach to Distributed Heterogeneous Database Systems for Integrated Manufacturing

Hemant Jainand Mohammed I. Bu-Hulaiga (1992). *Journal of Database Administration* (pp. 21-29).

[www.irma-international.org/article/approach-distributed-heterogeneous-database-systems/51107](http://www.irma-international.org/article/approach-distributed-heterogeneous-database-systems/51107)

### Disclosure Control of Confidential Data by Applying Pac Learning Theory

Ling He, Haldun Aytugand Gary J. Koehler (2010). *Journal of Database Management* (pp. 111-123).

[www.irma-international.org/article/disclosure-control-confidential-data-applying/47422](http://www.irma-international.org/article/disclosure-control-confidential-data-applying/47422)

### An Emergent Model of End-users' Acceptance of Enterprise Resource Planning Systems: A Grounded Theory Approach

Fiona Fui-Hoon Nahand Xin Tan (2015). *Journal of Database Management* (pp. 44-66).

[www.irma-international.org/article/an-emergent-model-of-end-users-acceptance-of-enterprise-resource-planning-systems/153517](http://www.irma-international.org/article/an-emergent-model-of-end-users-acceptance-of-enterprise-resource-planning-systems/153517)

### Assessment of Students by a Teacher with a Handheld Device and a Networkable Database

C. Paul Newhouse (2009). *Database Technologies: Concepts, Methodologies, Tools, and Applications* (pp. 1309-1319).

[www.irma-international.org/chapter/assessment-students-teacher-handheld-device/7974](http://www.irma-international.org/chapter/assessment-students-teacher-handheld-device/7974)