SIX

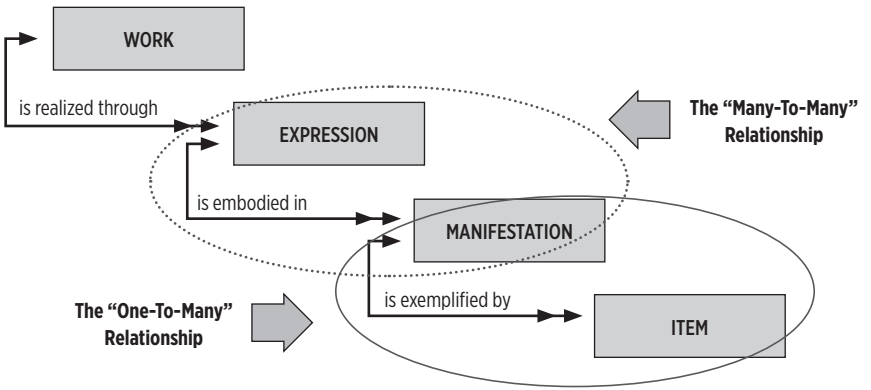
THE ENTITY-RELATION MODEL

Nearly every book, article, or presentation about FRBR has an explication of the primary FRBR entities and their relationships. The entity-relation (E-R) analysis defines the primary structure of the FRBR Final Report, which has chapters for entities, relationships, and attributes. The use of the entity-relation modeling technique was a requirement posed by the Terms of Reference for the study. Because of the great influence that this modeling technique had on the outcome of the study, it is worth examining in some detail.

Entity-relation analysis makes use of particular notation or diagrams to explain what is being expressed. There are many possible notation styles, some using boxes, some ovals, some with many different kinds of arrows and lines, with each notation carrying a specific meaning. The notation used in the FRBR document diagrams dates from the early days of E-R modeling. In this notation there are boxes for entities and arrows for relationships. The arrows can have one head, meaning that only one individual of the entity (e.g., one person) can be related, or they can have two heads, meaning that the relationship can be "many" (figure 6.1).

FIGURE 6.1

Many-to-many and one-to-many relationships in FRBR



Looking at the E-R analysis technique in historical context alongside the work of the Study Group that was developing FRBR helps explain what the diagrams are capable of expressing. Entity-relation modeling and relational databases were developed in the 1970s, and hit their peak in the 1980s. By 1990, the use of relational concepts was overlapping with a new computing paradigm: object-oriented programming and database design. The FRBR group made use of an early version of the E-R modeling concept and notation that was developed in the late 1970s. By 1990, E-R modeling had added design features that allowed the expression of more than just entities and relationships: in these new modeling notations it was possible to indicate inheritance, precise cardinality, processes, and communication paths. These later techniques would have made it possible to indicate whether the FRBR entities were required or optional, something that is not included in the FRBR entity-relation diagrams. Those diagrams show the same relationship between persons and works as between expressions and works, yet we know not every bibliographic description requires that a person be responsible for a work, and the text of the FRBR Final Report states that every expression has a mandatory relationship to a work. These are not distinguished in the diagrams.

By the time the FRBR Study Group provided its first draft in 1994, E-R modeling techniques had been replaced in technical design circles with the Unified Modeling Language (UML), which was developed during the 1980s. UML is a much more expressive language, with fourteen different diagram types, modeling both structures and behaviors. It also is designed primarily for object-oriented analysis, because entity-relation modeling had been superseded by object-oriented design. Had UML been used by the FRBR Study Group the outcome of the

study might have been different, but that also would have required a different skill set on the part of the Study Group members.

For a high-level view, a simple E-R model can still be useful. As its name implies, E-R modeling views one's information domain as entities or things, and defines the relationships between those things. The E-R modeling technique provided a structured approach for the FRBR Study Group, whose task was quite broadly defined. Use of the technique was required by the Terms of Reference document that gave the group its charge. In the Methodology section of the FRBR Final Report, the group explains:

> The methodology used in this study is based on an entity analysis technique that is used in the development of conceptual models for relational database systems. Although the study is not intended to serve directly as a basis for the design of bibliographic databases, the technique was chosen as the basis for the methodology because it provides a structured approach to the analysis of data requirements.

The FRBR Study Group makes clear that the resulting analysis is not a record design, yet there is an acknowledgment that the FRBR Final Report answers some questions that could be applied to bibliographic records:

> The study makes no a priori assumptions about the bibliographic record itself, either in terms of content or structure. It takes a user-focused approach to analyzing data requirements insofar as it endeavours to define in a systematic way what it is that the user expects to find information about in a bibliographic record and how that information is used. The study uses an entity analysis technique that begins by isolating the entities that are the key objects of interest to users of bibliographic records. (FRBR Final Report, 3)

However, the possibility that the analysis could be a precursor to database design was also hinted at in the FRBR Final Report:

> The entity-relationship analysis reflected in the model might also serve as a useful conceptual framework for a re-examination of the structures used to store, display, and communicate bibliographic data.

Barbara Tillett discusses this in her 1994 report on the work of the FRBR Study Group: "We hope this exercise will provide the basics for development of future structured bibliographic databases and future systems that facilitate creation, maintenance, and use of such databases" (Tillett 1994).

At the time, the E-R approach was new to at least some members of the FRBR Study Group and was not part of most catalogers' backgrounds. The FRBR document itself refers to readings in this area that the Study Group members found useful in understanding the entity-relation technology:

> The entity-relationship analysis technique and the conventions for graphic presentation that are used in this study are based in large part on the methodology developed by James Martin and outlined in his book *Strategic Data-Planning Methodologies* (Prentice-Hall, 1982). Graeme Simsion's *Data Modeling Essentials* (Van Nostrand Reinhold, 1994), Richard Perkinson's *Data Analysis: the Key to Data Base Design* (QED Information Sciences, 1984), and Ramez Elmasri and Shamkant Navanthe's *Fundamentals of Database Systems* (Benjamin/Cummings, 1989) were also used in shaping the methodology for the study. All four books are recommended to those who are interested in additional background and more detail on entity-relationship analysis. (FRBR Final Report, 10)

Note the emphasis on database design. Also note the dates on the books cited—the newest is now twenty years old. Relational database design, although still used in business applications, is no longer cutting edge technology. Although modeling of entities and relationships is still common, it has changed significantly from the models in use in these books.

The FRBR entity-relation diagrams show a macro-level model that includes only what are considered to be the primary relationships between entities. Other relationships between bibliographic entities are defined in the text, such as work/work relationships. These do not appear in the diagrams in the FRBR Final Report, thus presenting an incomplete picture of the actual bibliographic model described in the text.

ENTITIES, RELATIONS, AND DATABASE DESIGN

One of the reasons given for the development of an E-R model for bibliographic data was the desire to create a bibliographic data model that was more in tune with current technology. Because MARC was developed as a carrier for printed bibliographic data, and preceded the automation of library catalogs, it wasn't designed with database technology in mind. That doesn't mean that database technology has not been employed in library systems; in fact, they would not function as they do without the storage of data in such systems. Online systems must make use of the efficiencies built into database management systems.

Restructuring MARC data for use in relational databases, as discussed in the chapter on technology, is not an easy task. There are indeed some significant differences between bibliographic data and business data, and there is no question that the main customer base for database technology is the business world, not libraries. Therefore, the database technology that is on the market is optimized for the needs of the majority (and richest) of customers. For example, bibliographic data is primarily textual. Unlike much business data, bibliographic data has few numerical amounts that need computation, and we know that computers are more suited to work with numbers than with text. Also, there is not an even or predictable amount of repetition in bibliographic data; there are some authors or subjects that have high redundancy in a file, but there are even more that exist in a single exemplar. Relational databases are at their most efficient when the same data repeats frequently in the database, but provide less of an advantage for data with a high level of uniqueness.

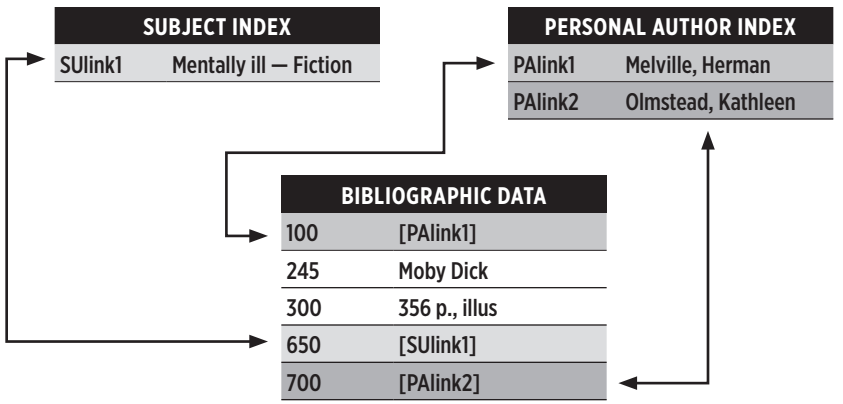
In spite of the fact that bibliographic data isn't what database management system developers had in mind when developing the technology, from the very first every library system has made use of some of the features of a database management system in order to function. It is therefore a misunderstanding to assume that because library data is not easily normalized into ideal relational database forms, library systems do not make use of relational database technology. They do, although the result does not look like the idealized design in database design textbooks.

As part of the development of the FRBR-informed cataloging rules, Resource Description and Access (RDA), Tom Delsey developed the RDA Implementation Database Scenarios, which depict "database structures conventionally used in library applications" (Delsey 2009). Scenario 3 shows library systems working with MARC-based data using a "flat file" approach, which would be similar to data stored in a spreadsheet. In fact, no flat file–based system could produce the kind of retrieval that library systems provide, and most systems today are at least as sophisticated as that document's Scenario 1, which appears to be the preferred model for the management of RDA data. The virtues of the MARC record, with its variability of field and subfield combinations and the unlimited repeatability of most fields, make it unsuited to a flat-file treatment. It would not be possible to provide search or browse on field types, like titles or subjects, without making use of entities and relations. In fact, the database I worked on in the early 1980s definitely used a relational design. A mock-up of its very high-level design is shown in figure 6.2. There were also many other indexes for corporate authors, titles, dates and languages, which I don't show here for reasons of space. An actual

database design is a mass of boxes and arrows that often cannot be reproduced on a single sheet of paper.

FIGURE 6.2

Inside A library system database, circa 1984



There are a couple of points that should be taken away from this. One is that although a data structure that has clear entities and relations defined may be somewhat easy to extend into a database design, many times a database design is derived from data that was initially developed for some other purpose. Although the bibliographic data that is stored in the MARC record still adheres to a structure that originally supported the card catalog, an E-R analysis can be done that results in a database that supports search and display of the data. This database design is primarily based on practical considerations: enabling retrieval of headings and combinations of headings with fast response time, even within large databases. Although the database model for bibliographic data differs considerably from, for example, that of banking, library systems run on the same underlying technology, making use of the features that a database management system provides.

This means that the FRBR E-R model is not the first practical use of E-R modeling for bibliographic data. The opportunity that FRBR afforded was a rethinking of bibliographic data model with entities and relations in mind, which did not adhere to the model of description and headings that has been the form of bibliographic data for centuries. The goals stated in the document— facilitating sharing and decreasing the costs of cataloging on an international scale—may have motivated the FRBR Study Group to develop the entities and relationships in FRBR, although the connection between the goals and the E-R model are not presented clearly in the FRBR Final Report.

SEVEN

WHAT IS MODELED IN FRBR?

An E-R analysis serves to resolve category boundaries and assign attributes to categories of things or functions. That said, for any given data, there can be any number of E-R models developed, depending on the functionality desired, the requirements of your data management system, and the workflow you need to support. The same is true for bibliographic data. The top-level model developed by the publishing industry has three primary "things": people, stuff, and deals (figure 7.1).

This represents a bibliographic model that primarily supports commercial functions around intellectual resources. The library model developed as FRBR could be described as "people, stuff, and subject access." Each model reflects the needs and views of its community.

The simpler your goals, the simpler your data model can be. However, the FRBR Study Group had a rather complex set of goals. One goal had to do with simplifying the bibliographic record for international sharing, with the purpose of cost savings. Another goal required the Study Group to make a connection