1

INTRODUCTION

- What You Will Learn
- What a Semantic Search Is
- What the Semantic Web Is
- How It All Works

The Semantic Web. The Social Web. Social search. Metadata. Linked data. To most people these words have little meaning despite the importance of what they are and what they will become. Tired of wading through millions of search results looking for the "right" thing? Think that there must be a "better" way to find things for both you and your patrons?

Search engines return millions of results, which no group of searchers of any size, never mind one person, can possibly view and evaluate. Social media sites expand the concept of friends to anyone you wish to add to your online network, increasing your ability to get advice, learn, network, find others for social media gaming, and, yes, even find life partners. Social search is leveraging the collective knowledge of the world to find what you need. Social search encompasses both the Social Web and what exists of the Semantic Web today.

This practical primer will focus on leveraging new search and Semantic Web tools to create better user experiences. It will provide the steps that those who create data and websites can take to make your library's resources both more accessible on today's web and ready for the future. The projects we provide all include ideas that you can put into place now and encompass both the Social Web and the Semantic Web. They are easy to implement and require only minimal technical know-how.

WHAT YOU WILL LEARN

We set out with several goals for this book. By the time you're done reading it, you should:

1. have better ways to find what you are looking for, tapping into the newest and most innovative search strategies, including mining social media and hidden content across the web;
2. have a solid understanding of the concepts of the Semantic Web and understand why it is important to the next stages in the evolution of online resources; and
3. be knowledgeable about practical applications for the Semantic Web and have tools you can use to align your resources for the Semantic Web of the future.

▶ WHAT A SEMANTIC SEARCH IS

A semantic search is a new way of searching that takes advantage of connecting data. In many cases, for a semantic search to work, the underlying concepts of the Semantic Web must first be in place. Some of them, such as location-based data, are already in wide use, while others, such as microformats (covered in Chapter 5), are still in the early stages of implementation. A few examples of new semantic search technologies include location-based searching (Bing Maps), real-time searching (Twitter's Real Time Search), and social searching, such as Google's Social Search features, which are still listed as "experimental."

▶ WHAT THE SEMANTIC WEB IS

The notion of the Semantic Web has been around at least since 2001, but the notion of the evolving web has been around much longer. The concept involves us adding context to our data so that computers can do more of the work for us.

The Semantic Web is customizable and personalizable; it provides a better means of filtering your search results and relies on the idea that giving context to data makes that data more rich and the searches "smarter." The Semantic Web relies on metadata and linking data together (linked data) and encompasses a variety of formats, such as images and video. Another feature of the Semantic Web is the ability to weave social media (blogs, Twitter feeds, tumblogs, etc.) into search results. This next stage of connecting and interconnecting data on the web is the Semantic Web. Being able to search that content and to take advantage of all those connections is semantic search. The Semantic Web is not a radical change but an evolution. Additionally, it is not solely data driven, as much of the metadata is created by users through practices such as tagging.

We are not suddenly going to wake up one day to find that all of the resources in the world have been made available via the Semantic Web.

Considering many resources are still in analog formats, there is a lot of work to be done before much of that content is converted to a digital format, let alone have good metadata and linked data associated with it. Even more work is needed before everything has metadata in a form that can be used by search engines. However, we are already seeing Semantic Web features and functionalities showing up in online search tools and even in some library catalogs. With these features we can better customize our searches for our specific needs. We can interact with data in new ways, whether through tagging library catalog records or creating reading lists.

Much of the data generated by libraries, museums, and other information-based institutions is siloed. In other words, each kind or group of data sits in its own container (typically a database), often with a proprietary structure around it, making it nearly impossible to be searched via a web-based search engine. One of the biggest examples of this kind of data is the library catalog; in some cases records can be harvested by search engines, but even then they are extraordinarily hard to find.

As libraries move toward new standards and rules for creating semantic library catalog metadata records, specifically FRBR (http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records) and the forthcoming RDA (http://www.rdatoolkit.org/) (discussed later in this chapter), there is much hope that this information will become less invisible to the web. Knowing more about the Semantic Web means that librarians will understand why their data works (or doesn't) with search engines. The goal of the Semantic Web is to take all of the "stuff" on the web, including library content, and put it together in a way that is accessible and usable to the general public and to provide a better means of filtering and controlling that information.

Of course, there is more to the web than just searching; the Semantic Web will allow even more leveraging of mobile devices. For example, a typical smartphone today allows you to search for a restaurant based on your current location, sort the results by best review and/or distance, book a reservation for you (assuming the restaurant has an online presence), provide directions (driving or walking), and send all of that information to your friends all in a matter of minutes. In the not-so-distant past, this scenario would have seemed like science fiction, but it is a reality today. The ability to do these kinds of things using a mobile device is based on data and the ability of the device to communicate with the web. Bits of data and the relationships among them (e.g., reviews linked to a restaurant's website) are examples of how the Semantic Web can (and does!) work.

Another example of the Semantic Web is Amazon.com's recommendation service. Using semantic markup and semantic search, Amazon analyzes what

you've looked at and/or ordered and makes a recommendation. Providing more relevant results for your search is a big part of the Semantic Web, but it goes beyond that. It is facilitating the dialogue between devices and people, making our lives easier. In the future, your "smart" refrigerator may keep an inventory of its contents and add items to your shopping list that you are low on. Then, when you're driving home, your phone will alert you to the fact that you need to stop to pick up more milk. You and your friends or colleagues might collaborate on a novel online, publish it automatically to a variety of sources, and automatically link reviews to it. The Semantic Web doesn't care about where the data comes from.

Wow, is there a lot of stuff to wade through on the web! Images, podcasts, statistical data, websites, documents, maps, blog posts; the list is practically endless. How can we possibly find what we really need? If the goal of the Semantic Web is to take all of the "stuff" on the web and pull it together in a way that is usable, how does this happen? What makes it work? Do you need to know how it works in order to use it? Ultimately, the hope is that the answer to that last question becomes "no." However, given that the Semantic Web and semantic search are still developing, knowing more about how they work behind the scenes may save a headache or two down the road when trying to decide what kinds of projects to do or even how to do them. Knowing a little bit about how they work means that you can better use the resources that exist and explain them to friends, patrons, a child who needs homework help, and others. You will be more information literate.

▶ HOW IT ALL WORKS

Metadata is the key to the Semantic Web. It is like a translation service, allowing programs to talk to each other and "understand" the data they find. "Metadata" is a scary word to a lot of people; it conjures up confusion with its definition of "data about data." True, this is what metadata *is*, but it doesn't really tell us what metadata *does* or how we can use it. We already have machines that talk to the web—desktops, laptops, netbooks, MP3 players, streaming television devices, tablets, e-readers, phones, even cars and cameras. Metadata is simply the language used to communicate among devices, databases, items, and objects.

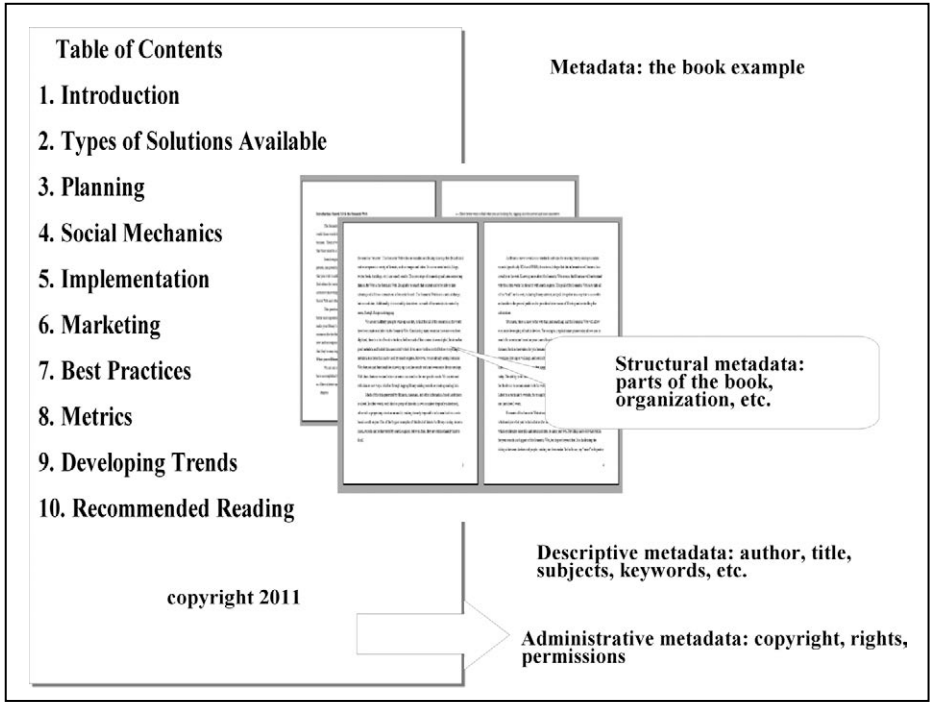
How about a real-life example of metadata? Let's consider my dog. When I got a new dog, one of the first questions people asked is what kind of dog is it. My dog is an Australian shepherd. Other questions were how old is it, its name, and its gender. Each feature about the dog can be thought of in terms of metadata. These features are bits of descriptive metadata, because they describe the dog. My dog is a five-year-old female Australian Shepherd

named Roxy. All of these features are coded into a database at my vet's office, and this database could share those bits of metadata, even with another vet's office. Going further into the metadata analogy, my dog is assigned an ID number, which is a unique number, a **unique resource identifier**. On the practical side of things, this keeps all of her information together and ensures that her medical history is not mixed up with another dog's. From a metadata standpoint, she has a metadata record with a unique number.

Libraries have been creating metadata for a very long time, although traditionally the largest focus has been on descriptive metadata. Descriptive metadata describes what an item is, with a goal of identifying the object and making it findable. This type of metadata is the foundation of library catalog records. Looking at the catalog record of a book, its title, author's name, notes, and subject headings are all descriptive information. There are three major types of metadata (see Figure 1.1):

1. Descriptive
2. Structural
3. Administrative

▶ **Figure 1.1: Types of Metadata**



Source: Graphic by Robin Fay, 2011.

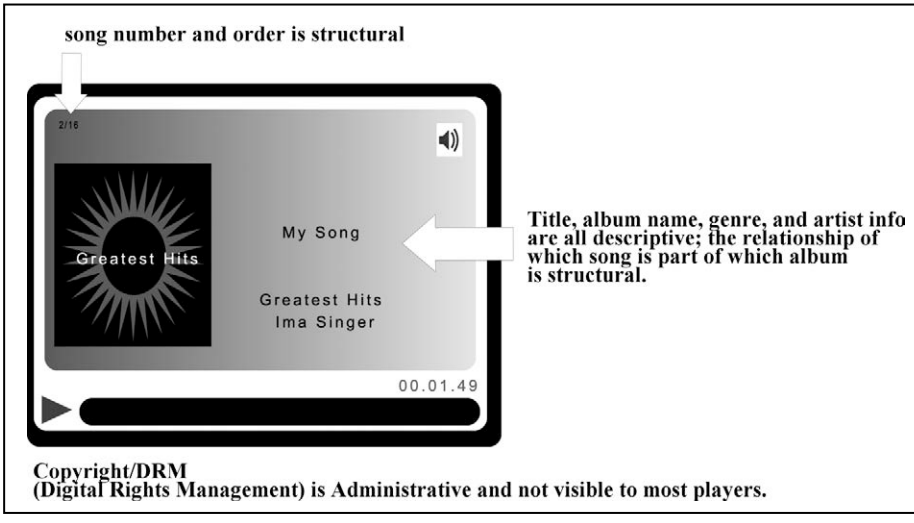
Different communities organize metadata into different groupings; however, all metadata depend on a schema, a set of rules for the elements and attributes, and the rules for how that information should be arranged. Descriptive metadata describes the item itself, using titles and topic description (keywords, subject headings, and tags). Structural metadata explains the nature of the object, such as how many parts are included in the overall object. In a book, for example, this might be the pages, how the pages form chapters, or format. Administrative metadata includes information about how to handle or process the object. In the example of Roxy's vet record, administrative metadata might include her owner's information; for a letter or painting, it might include the chain of ownership (i.e., the provenance). One of the most commonly used types of administrative metadata on the web is ownership rights metadata. One format that many are familiar with these days is Digital Rights Management (DRM), the licensing that allows the use of commercially produced digital files, such as music, movies, and e-books.

Let's take a look at a common type of audio file, .mp3. These files generally have a variety of metadata associated with them (see Figure 1.2):

- Title > Descriptive
- Song no./album info > Structural
- Licensing/limits > Administrative/rights

Metadata is the data that describes an object (photo, book, your dog, etc.) in a database (such as your vet's patient database; to an even larger

▶ **Figure 1.2: Metadata on an MP3 Player Display**



Source: Graphic by Robin Fay, 2011.

extent, the World Wide Web), explains how to display it, and who has access to it.

How do creators of metadata know how or where to add it? Users at social sites such as Facebook, Flickr, and YouTube create metadata through a guided system. By typing in the title of a video or photo, the user is creating descriptive metadata. Some metadata can actually be embedded into the file itself by the device that creates it. For example, many digital cameras now record the type of camera and the file format into the file itself. When a user uploads a photo file to Flickr, Flickr will recognize it as a .jpg format and often will be able to retrieve (harvest) the camera make and model from the file itself. In both cases, the metadata that is created follows instructions for displaying the information, as determined by its schema (rules). We'll look into this example further in Chapter 5.

Linked Data

In the Semantic Web world there is no limit to number and diversity of schemas as long as there is some way (usually through a link pointing back to the schema file) for a search engine or other semantic tool to access the schema. Just making all of the data in the world accessible doesn't necessarily create a way for people to find it.

The web has always allowed us to create links between webpages and files, and that's a great first step. When we add a link to a webpage, post a link to Facebook or Twitter, or embed a video in a blog post, we have created a link from one resource to another. However, that link has no context other than what we provide with it. For example, users can only assume that a link that reads "White House" will lead them to the White House's website. The contexts of links are created manually by site content editors, and we need to trust them to actually link us to the content they've implied they're taking us to.

The Semantic Web uses linked data to do just what its name implies—link individual items of data together, creating points of connection, relationships, and/or context so that we can find what we're searching for. Linked data is created by computers based on data, instead of just a link (or links) to files, that we create when we link to a website. For example, recommendation services ("If you bought this, you might like this . . . ") rely on creating a connection between like items to make that recommendation.

RDA

Resource Description and Access (RDA) is a new code of cataloging rules, designed to replace the *Anglo-American Cataloguing Rules*, Second Edition (AACR2), as the law of the cataloging land. RDA was released by the Joint

Steering Committee for Development of RDA in June 2010, and the rules were subsequently evaluated in a testing process by three U.S. national libraries (the Library of Congress, the National Library of Medicine, and the National Agricultural Library). The result of the test was the decision that RDA will be adopted but no earlier than January 2013. Even though it has not been officially implemented yet, RDA has the potential to greatly change library data and how that data interacts with other information on the web.

What can RDA do for your library data, and how can it help libraries get ready for the Semantic Web? For one thing, RDA has the potential to break library records down into smaller pieces of information, some of which can be provided in a machine-actionable format. Right now, bibliographic description according to AACR2 is based on eight areas of description. For example, one area of description deals with information relating to the publication of the item being cataloged, and, in a catalog record, this area in AACR2 could look like this: Chicago, Ill. : American Library Association, c2011. A number of different pieces of information are expressed in this one area: place of publication, publisher, and copyright date. In RDA, each of these three smaller pieces of information is its own element.

RDA also clarifies how different types of information should be recorded within an area. To return to the example of publication information, some AACR2 records have publication dates: Chicago, Ill. : American Library Association, 2011. Others have copyright dates: Chicago, Ill. : American Library Association, c2011. Right now, the presence of the letter "c" before the date is the only thing that indicates to the human reader that the date is a copyright date and not a publication date, and there is no way at all for a computer to understand the difference between these dates. According to RDA, publication date and copyright date are two different elements. Because it explicitly indicates particular elements for specific pieces of information, RDA is the first step toward creating data that can be recognized by computers as particular types of information and therefore integrated into Semantic Web searches.

In addition to breaking down catalog records into smaller pieces of data, RDA also specifies a great deal more data to be included in authority records to represent the creators of the items in library catalogs. New authority elements include occupation, gender, and associated places, and authority records created according to RDA are much richer sources of information than the ones currently used. This is a good thing, because the more information that libraries have, the more opportunities they have to link their data to other sources in the Semantic Web.

A third change that RDA will bring to cataloging is an emphasis on the relationships between pieces of data in catalog records and between catalog

records for different items. RDA's rules are based on a conceptual model called the *Functional Requirements for Bibliographic Records*, or FRBR. FRBR identifies the entities that are represented in library catalogs and their relationships to each other, as well as their relationships to the creators of these entities and the subjects of these entities.

To understand the effects of FRBR on library catalogs, think about searching for an item like *Romeo and Juliet*. When searching a so-called "FRBR-ized" catalog for *Romeo and Juliet*, patrons would be able to find all of the various editions of that work in an easily navigable display rather than having to view different records for each edition. In addition, catalog users could find movie adaptations of the play and even related works, such as *West Side Story*.

The changes introduced by RDA and FRBR could have a great effect on libraries, even within the confines of their own catalogs. However, the real power of RDA is its potential to integrate library information with other information on the web. If library records are broken down into pieces of data, the principles of linked data can be used to create connections to other sources of information on the web. If this happens, it will be possible to retrieve library data through searches that do not start with the library's catalog. A search like the one for *Romeo and Juliet* could result in information from a number of sources being brought together without having to search the sources independently.

RDA is the library profession's way of preparing metadata to be ready for use by Semantic Web search engines; other information communities are developing their own ways to prepare metadata for their materials so that they can be harvested. Libraries have exciting times ahead—ready to get started?