Chapter 7
Search Integration with WebSphere Portal:
The Options and Challenges

Andreas Prokoph
IBM, Germany

ABSTRACT

Modern web applications and servers like Portal require adequate support for integration of search services due to user focused information delivery and user interaction, as well as new technologies used to render such information, which is exemplified by two fundamental problems that have long plagued web crawlers: dynamic content and Javascript generated content. Today, the solution is simple: ignore such web pages. To enable "search" in Portals, a different "crawling" paradigm is required to search engines to gather and consume information. WebSphere Portal provides a framework that propagates content and information through "Seedlists"—comparable to HTML based sitemaps but richer in terms of features. This mandates that information and content delivering applications must be "search engine aware", requiring them to enable services and seedlists for fast, efficient and complete delivery of content and information. This is the main integration point for search engines into the portal for Portal site search services for a rich and user focused search experience. This article discusses how such technologies can allow for more efficient crawling of public Portal sites by prominent Internet search engines as well as myths surrounding search engine optimization.

1. PORTAL SITES AND TRADITIONAL WEB SEARCH ENGINES

WebSphere Portal allows people to interact with applications, processes, other people, documents and content in a personalized and role-based fashion. In a day and age when information overload has become commonplace, Portal allows context-relevant resources to be presented via common Web browsers, placing what people need to complete the job at hand at their fingertips, regardless of where they are.

DOI: 10.4018/978-1-4666-0336-3.ch007

Figure 1. Portal and application integration overview

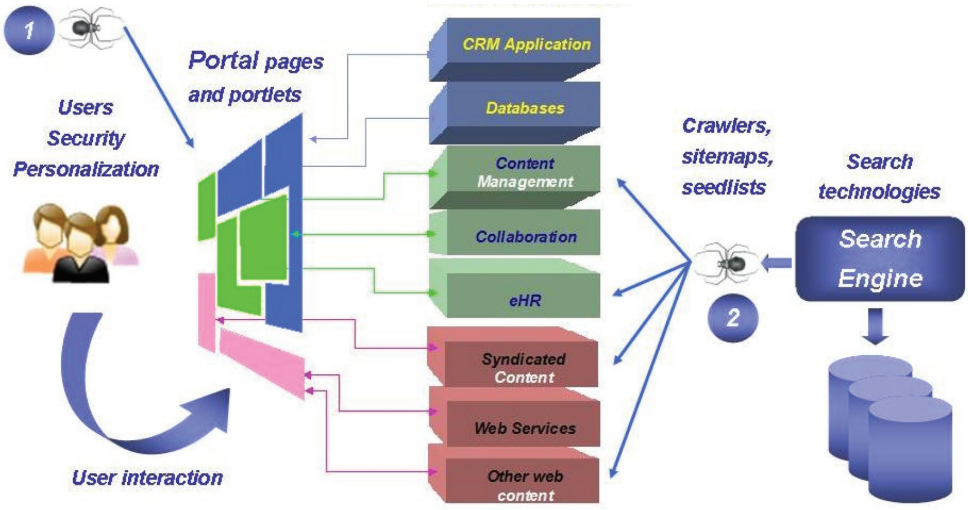

Figure 1 shows the range of applications and services that integrate with Portal, as well as how that information and content is visualized to the end user.

Figure 1 also shows on a very high level how a search engine can integrate with the Portal:

1. Traditionally a regular web crawler starts with the homepage and then follows all links to capture all information provided by the portal site. What options are available to optimize this process will be described in this article. We will also discuss the restrictions and limitations. as well as options to resolve some of these
2. We will show how a new crawling technique can be enabled on the application level, which no longer requires the crawler to communicate and explore applications and repositories (pull technique). The application will rather send the crawler on request the list of what information the crawler is requested to fetch and process (push technique). Technical details as well as advantages of doing so foremost in a Portal environment is also discussed in more depth throughout this article

1.1 Content Centric Portal - Just Like "Any Other Website"

The simplest case to look at first is a Portal site which delivers static content through its pages and portlets. Portal provides a rich user experience in terms of consistent navigation, appearance and information delivery.

Crawling such a Portal site is a matter of pointing the crawler either at the homepage, a sitemap, or a site-directory. The crawler will then identify and record all links to referenced pages of that site, fetch their content and then send it for final processing to the indexing service. And once stored in the index, the user will then be able to perform his search requests.

That said: this is "business as usual" for the web-crawler, no real difference compared to crawling a standard web-server delivering more or less static HTML content.

1.2 What About the URLs?

We just started to look at a portal site and it being crawled by a web-crawler like for example the Google-Bot. To some this might already ring a bell. If not or in addition: others might then ask the