Performance tuning OAI-PMH services

There are numerous factors that can affect the performance of any computing system, and systems that utilize the Open Archive Initiative Protocol for Metadata Harvesting are no exception. The performance of computing systems in general is a complex topic upon which much has been written, both general and specific to certain subfields of computer science. Therefore, in this chapter we will attempt to discuss those factors that are common to OAI-PMH systems with as little digression into the field of computing performance at large as possible.

This chapter is divided into two sections: one for data providers and one for service providers. This division is mostly based on which of the data provider or the service provider has the most control over a particular aspect of performance. However, the performance characteristics of each are closely related, for example the good or poor performance characteristics of a data provider can have a large impact on the performance characteristics of service providers attempting to harvest from that provider. Likewise, the characteristics of a service provider can have a large impact on the performance of a data provider from which it is harvesting. These interrelationships will be noted in the appropriate sections below.

It should also be stressed that your level of control over many of these factors might be limited. If you are a programmer developing a new data or service provider from scratch, you will have control over many of these factors. However, if you are simply running a turnkey system that utilizes OAI, you will probably have little control other than buying faster hardware or tweaking a few configuration parameters of the turnkey system. We will also note these issues in the following sections.

Data Provider Performance

Quite often data providers are seen as a primary bottleneck for service providers. This should be fairly obvious because service providers are entirely dependent upon data providers for their records, and a few poorly performing data providers can radically slow the update cycle for a service provider. These slowly performing data providers can be the difference between an update cycle of days or weeks and an update cycle of hours. In some cases a service provider may decide to entirely drop a poorly performing data provider from the service. In other cases service providers may relegate poorly performing data providers to a slower or lower priority update cycle. The OAIster service at the University of Michigan currently has such a two-tiered update cycle where good performing data providers are harvested once per week and poorly performing data providers are harvested once a month.[1]

Poor performance may be related to one or more of several general factors, including: incorrect implementation of the protocol, inefficient implementation of the protocol, poor metadata quality, slow hardware or network, or overburdened server or network. These problems can be ameliorated in various ways. The following sections will address some of the common methods for affecting performance.

Protocol Implementation

As obvious as it may seem, it is difficult to overstress the importance of correctly implementing the OAI-PMH. The first inclination of data providers is to assume that if their implementation was not correct they would quickly discover this either because prospective harvesters would report the failure or because they have tested their implementation with one of the online test suites, such as the Repository Explorer.[2] This assumption is generally true for gross protocol errors, such as those that would be detected by XML Schema validation. However, for subtle errors, some of which can have a large impact on performance, it fails for a number of both technical and social reasons.

First, as highly recommended as the Repository Explorer and other similar tools are, they typically do not or cannot test for every idiosyncratic behavior that could be exhibited by a data provider. These tools usually do not do a complete harvest of all records using all combinations of sets and metadata formats. Therefore, errors that would occur toward the end of a large harvest go undetected. These tools also do not look at the behavior of a data provider over time. For example, they will not detect if date stamps are being correctly updated when records are modified. Finally, there are aspects of the protocol which if misused or not used are not strictly errors, but can affect performance, such as failure to use resumption tokens or flow control.

A couple factors come into play to explain why a harvester may not detect or may not report errors with your data provider. The first is a social issue. Unless there is a strong relationship between the data and service provider or the service provider has a strong desire to harvest a particular data provider, it is often easier for a service provider to simply not harvest a misbehaving data provider than to report the problem. On the opposite side, data providers do not always assign actively monitored email addresses to the <oai:adminEmail> field of the Identify response, so even if problems are reported they can go unnoticed.

The second factor is more technical but has a social aspect as well. In keeping with the spirit of the OAI-PMH which places a much higher burden on service providers than data providers, many developers of OAI harvesting software have gone to lengths to accommodate errant data providers even if it means employing inefficient processes or harvesting strategies. This is essentially a manifestation of the Robustness Principle or Postel's Law[3], "be conservative in what you send, liberal in what you accept." The affect is that inefficient or misbehaving OAI data providers can often go undetected, even if this means that service providers are spending an inordinate amount of processing time dealing with them.

In the next few sections, several of the subtle protocol errors that can affect performance will be discussed.

Date Stamps

The correct usage of date stamps is crucial to one of the protocols most important performance enhancing strategies: incremental harvesting. …

Related to this is the deleted status…

Service Provider Performance

Performance tuning OAI-PMH services

OAI-PMH response compression

Scalability & robustness issues

Considerations in provider metadata update procedures

Validation options when harvesting

Considerations in setting harvesting frequencies

[1] Kat Hagedorn. “Information for Potential Data Providers,” OAIster, University of Michigan Digital Library Production Service. http://www.oaister.org/o/oaister/dataproviders.html

[2] Hussein Suleman, “Open Archives Initiative - Repository Explorer,” version 1.46a. Advanced Information Management Laboratory at University of Cape Town, July, 2005. http://re.cs.uct.ac.za/

[3] Wikipedia contributors, "Robustness Principle," Wikipedia: The Free Encyclopedia, http://en.wikipedia.org/wiki/Robustness_Principle (accessed August 27, 2005).