Design - Identifier Generation

Since Fedora 5.0 will be a major release, we have discussed whether we should make changes to the identifier generation configuration and implementation. The main options that have been discussed are:

1. Keep the current PairTree configuration

This would keep the current configuration: when an identifier is generated, 4 levels of PairTree nodes are created to partition the identifier space. These nodes do not behave like normal containers, for example when you list the members of a container, it traverses the PairTree nodes and lists their children.

2. Use AppleTrees

This takes a different approach to avoiding JCR nodes with too many children. Instead of putting it in the externally-visible path, it generates an MD5 hash of the path and creates PairTrees internally to segment this child nodes. This has many of the same benefits of #1, but hides the intermediary nodes from the user. However, it also has the downside of being incompatible with migrating data, so it would require starting a new repository and migrating content.

3. Remove the PairTree configuration, but make it easier to adopt #1 or #2

This would simplify the default configuration, but at the expense of hurting scalability with the default configuration. To help address this, we would need to make it easier to change the configuration to use PairTrees or AppleTrees.

4. Keep Pair Trees under the hood, but use them only for "path segments that don't exist"

In this scenario, Modeshape will create hidden "pairtree" nodes for the creation of any resource whose URI contains path segments that do not correspond to existing resources in the repository. These nodes exist solely to satisfy the needs of Modeshape impl internals, and are not accessible via HTTP. A GET to one of these internal nodes would return a 404

Example:

PUT http://example.org/fcrepo/rest/a/b/c/d/e/myResource

Let's suppose that the repository is entirely empty, and contains only the root resource http://example.org/fcrepo/rest/

The end result is the creation of http://example.org/fcrepo/rest/a/b/c/d/e/myResource as usual

A GET to an intermediate node http://example.org/fcrepo/rest/a/b returns a 404

Internally, modeshape has a hierarchy as usual: root->a->b->c->demyResource, with nodes a, b, c, d, and e being "pair tree" nodes.

When migrating to Fedora 5, existing URI paths would not change. The only observable difference would be that pairtree nodes (path segments that were created as pairtrees) would return 404.

I think this option is complementary with (1) or (3), with the choice of (1) or (3) being reduced to "which ID generator to choose by default". This option (4) is merely saying "keep pair trees in modeshape where necessary structurally, but don't give them representations on the web any more"