XML validation and entity resolution

DTDs, catalogs and whatnot

Table of contents

1 Introduction....2
2 XML validation....2
3 Validating new XML types....3
3.1 Creating or extending a DTD....3
3.2 Associating DTDs with document types....4
4 Debugging the Catalog Entity Resolver....5
5 Referring to entities....6
6 Validating in an XML editor....6
7 Validation using RELAX NG....6
8 Validation using Cocoon sitemap and the Cocoon Validation block....6

Copyright © 2002 The Apache Software Foundation. All rights reserved.

1 Introduction

When a source document uses a DTD to define its structure, then various abilities are enabled. Forrest handles the annoying side of this and takes advantage of the benefits.

The xml parser must locate the DTDs and other entities. Forrest uses the XML Catalog Entity Resolver to associate DTD references with local copies.

Forrest also uses the XML Document Type Declaration to effect the Content Aware Pipelines. Remember that you are not required to use DTDs (for example Forrest can also respond to a namespace) so RELAX NG or W3C XML Schema can be used.

If you do use DTDs, then forrest is automatically configured to do XML validation.

This document explains both the validation and entity resolution aspects of the Forrest XML framework.

2 XML validation

By default, Forrest will validate your XML before generating HTML or a webapp from it, and fail if any XML files are not valid. Validation can be performed manually by doing '`forrest validate`' in the project root directory.

For an XML file to be valid, it *must* have a document type declaration at the top, indicating its content type. Hence by default, any Forrest-processed XML file that lacks a DOCTYPE declaration will cause the build to break.

Despite the strict default behavior, Forrest is quite flexible about validation. Validation can be switched off for certain sections of a project. In validated sections, it is possible for projects to specify exactly what files they want (and don't want) validated. Forrest validation is controlled through a set of properties in `forrest.properties`:

```
###############
# validation properties

# This set of properties determine if validation is performed
# Values are inherited unless overridden.
# e.g. if forrest.validate=false then all others are false unless set to true.
#forrest.validate=true
#forrest.validate.xdocs=${forrest.validate}
#forrest.validate.skinconf=${forrest.validate}
#forrest.validate.sitemap=${forrest.validate}
#forrest.validate.stylesheets=${forrest.validate}
#forrest.validate.skins=${forrest.validate}
#forrest.validate.skins.stylesheets=${forrest.validate.skins}

# *.failonerror=(true|false) - stop when an XML file is invalid
#forrest.validate.failonerror=true

# *.excludes=(pattern) - comma-separated list of path patterns to not validate
# e.g.
#forrest.validate.xdocs.excludes=samples/subdir/**, samples/faq.xml
#forrest.validate.xdocs.excludes=
```

For example, to avoid validating `${project.xdocs-dir}`/slides.xml and everything inside the `${project.xdocs-dir}/manual/` directory, add this to `forrest.properties`:

Copyright © 2002 The Apache Software Foundation. All rights reserved.

```
forrest.validate.xdocs.excludes=slides.xml, manual/**
```

> **Note:**
> The `failonerror` properties only work for files validated with Ant's <xmlvalidate> and not (yet) for those validated with <jing>, where `failonerror` defaults to `true`.

3 Validating new XML types

Forrest provides an OASIS Catalog [see tutorial] `forrest/main/webapp/resources/schema/catalog.xcat` as a means of associating public identifiers (e.g. `-//APACHE//DTD Documentation V1.1//EN` above) with DTDs. If you add a new content type, you should add the DTD to `${project.schema-dir}/dtd/` and add an entry to the `${project.schema-dir}/catalog.xcat` file. This section describes the details of this process.

3.1 Creating or extending a DTD

The main Forrest DTDs are designed to be modular and extensible, so it is fairly easy to create a new document type that is a superset of one from Forrest. This is what we'll demonstrate here, using our earlier download format as an example. Our download format adds a group of new elements to the standard 'documentv20' format. Our new elements are described by the following DTD:

```
<!ELEMENT release (downloads)>
<!ATTLIST release
version CDATA #REQUIRED
date CDATA #REQUIRED>

<!ELEMENT downloads (file*)>

<!ELEMENT file EMPTY>
<!ATTLIST file
url CDATA #REQUIRED
name CDATA #REQUIRED
size CDATA #IMPLIED>
```

The document-v20 entities are defined in a reusable 'module': `forrest/main/webapp/resources/schema/dtd/document-v20.mod` The `forrest/main/webapp/resources/schema/dtd/document-v20.dtd` file provides a full description and basic example of how to pull in modules. In our example, our DTD reuses modules `common-charents-v10.mod` and `document-v20.mod`. Here is the full DTD, with explanation to follow.

```
<!-- ============================================================

Download Doc format

PURPOSE:
This DTD provides simple extensions on the Apache DocumentV11 format to link
to a set of downloadable files.

TYPICAL INVOCATION:

<!DOCTYPE document PUBLIC "-//Acme//DTD Download Documentation V1.0//EN"
"download-v10.dtd">

============================================================ -->


<!-- ============================================================ -->
<!-- Include the Common ISO Character Entity Sets -->
<!-- ============================================================ -->
```

Copyright © 2002 The Apache Software Foundation. All rights reserved.

```
<!ENTITY % common-charents PUBLIC
"-//APACHE//ENTITIES Common Character Entity Sets V1.0//EN"
"common-charents-v10.mod">
%common-charents;

<!-- ============================================================= -->
<!-- Document -->
<!-- ============================================================= -->

<!ENTITY % document PUBLIC "-//APACHE//ENTITIES Documentation V2.0//EN"
"document-v20.mod">

<!-- Override this entity so that 'release' is allowed below 'section' -->
<!ENTITY % local.sections "|release">

%document;

<!ELEMENT release (downloads)>
<!ATTLIST release
version CDATA #REQUIRED
date CDATA #REQUIRED>

<!ELEMENT downloads (file*)>

<!ELEMENT file EMPTY>
<!ATTLIST file
url CDATA #REQUIRED
name CDATA #REQUIRED
size CDATA #IMPLIED>

<!-- ============================================================= -->
<!-- End of DTD -->
<!-- ============================================================= -->
```

This custom DTD should be placed in your project resources directory at `src/documentation/resources/schema/dtd/`

The <!ENTITY % ... > blocks are so-called parameter entities. They are like macros, whose content will be inserted when a parameter-entity reference, like `%common-charents;` or `%document;` is inserted.

In our DTD, we first pull in the 'common-charents' entity, which defines character symbol sets. We then define the 'document' entity. However, before the `%document;` PE reference, we first override the 'local.section' entity. This is a hook into document-v20.mod. By setting its value to '|release', we declare that our <release> element is to be allowed wherever "local sections" are used. There are five or so such hooks for different areas of the document; see document-v20.dtd for more details. We then import the %document; contents, and declare the rest of our DTD elements.

We now have a DTD for the 'download' document type.

> **Note:**
>
> Chapter 5: Customizing DocBook of Norman Walsh's "DocBook: The Definitive Guide" gives a complete overview of the process of customizing a DTD.

3.2 Associating DTDs with document types

Recall that our DOCTYPE declaration for our download document type is:

```
<!DOCTYPE document PUBLIC "-//Acme//DTD Download Documentation V1.0//EN"
          "download-v10.dtd">
```

Copyright © 2002 The Apache Software Foundation. All rights reserved.

We only care about the quoted section after `PUBLIC`, called the "public identifier", which globally identifies our document type. We cannot rely on the subsequent "system identifier" part ("download-v10.dtd"), because as a relative reference it is liable to break. The solution Forrest uses is to ignore the system id, and rely on a mapping from the public ID to a stable DTD location, via a Catalog file.

> **Note:**
> See this article for a good introduction to catalogs and the Cocoon documentation Entity resolution with catalogs.

Forrest provides a standard catalog file at `forrest/main/webapp/resources/schema/catalog.xcat` for the document types that Forrest supplies.

An additional system-wide catalog can be configured for use by multiple forrest-based projects. See the "local-catalog" parameter in `main/webapp/WEB-INF/cocoon.xconf`

Projects can augment this with their own catalog file located in `${project.schema-dir}/catalog.xcat` to use it you must specify either the path (full or relative) to your `catalog.xcat` in the `CatalogManager.properties` file. If you provide a relative path you must set the property `relative-catalogs` to "yes".

When Cocoon starts, it reads the `CatalogManager.properties` file from your `project.classes-dir`. This is usually src/documentation/classes/ but you can change this in `forrest.properties`. When you seed a new site using `forrest seed` a sample catalog file is placed in the site structure, you can use this as a template for your own files.

Forrest uses the XML Catalog syntax by default, although the older plain-text format can also be used. Here is what the XML Catalog format looks like:

```
<?xml version="1.0"?>
<!-- OASIS XML Catalog for Forrest -->
<catalog xmlns="urn:oasis:names:tc:entity:xmlns:xml:catalog">
  <public publicId="-//Acme//DTD Download Documentation V1.0//EN"
    uri="dtd/download-v10.dtd"/>
</catalog>
```

The format is described in the spec, and is fairly simple and very powerful. The "`public`" elements map a public identifier to a DTD (relative to the catalog file).

We now have a custom DTD and a catalog mapping which lets both Forrest and Cocoon locate the DTD. Now if we were to run `'forrest validate-xdocs'` our download file would validate along with all the others. If something goes wrong, try running `'forrest -v validate-xdocs'` to see the error in more detail.

4 Debugging the Catalog Entity Resolver

If you suspect problems with your project catalog, then raise the "verbosity" level in `src/documentation/classes/CatalogManager.properties` and re-start forrest. This should show your project catalogs being parsed and loaded.

However this configuration does not show your DTDs being resolved. So raise the verbosity level in the central configuration at `main/webapp/WEB-INF/properties/core.properties` and re-start forrest. This also shows the main catalogs being loaded and shows the resolving of every DTD and entity set.

When a DTD is successfully resolved you should see the message: `Resolved public:`

Copyright © 2002 The Apache Software Foundation. All rights reserved.

When debugging such issues, a network monitoring tool (e.g. ngrep.sf.net) is useful to ensure that all resources are being locally resolved and not wandering onto the network to find remote copies.

5 Referring to entities

Look at the source of this document (`xdocs/docs/validation.xml`) and see how the entity set `"Numeric and Special Graphic"` is declared in the document type declaration.

ISOnum.pen	½	½

6 Validating in an XML editor

If you have an XML editor that understands SGML or XML catalogs, let it know where the Forrest catalog file is, and you will be able to validate any Forrest XML file, regardless of location, as you edit your files. See the configuration notes your favourite editor.

7 Validation using RELAX NG

Other validation is also conducted during build-time using RELAX NG. This validates all of the important configuration files, both in Forrest itself and in your project. At the moment it processes all skinconf.xml files, all sitemap.xmap files, and all XSLT stylesheets.

The RNG grammars to do this are located in the `main/webapp/resources/schema/relaxng` directory. If you want to know more about this, and perhaps extend it for your own use, then see `main/webapp/resources/schema/relaxng/README.txt` and the Ant targets in the various build.xml files.

8 Validation using Cocoon sitemap and the Cocoon Validation block

The content of pipelines can be validated at various points in the Cocoon sitemaps using RELAX NG or W3C XML Schema. See notes.

Copyright © 2002 The Apache Software Foundation. All rights reserved.