Saxon - Bug #4814

Error reported by XML parser: "unknown protocol: classpath" when the MathML3 DTD is referenced

2020-10-29 23:57 - Michael Kay

Status:	Closed	**Start date:**	2020-10-29
Priority:	Low	**Due date:**	
Assignee:	Norm Tovey-Walsh	**% Done:**	100%
Category:		**Estimated time:**	0:00 hour
Sprint/Milestone:		**Spent time:**	0:00 hour
Legacy ID:		**Fix Committed on Branch:**	10, trunk
Applies to branch:	10, trunk	**Fixed in Maintenance Release:**	10.5

Description

When the source document DTD references the MathML3 DTD, XML parsing fails with error "unknown protocol: classpath".

History

#1 - 2020-10-30 00:18 - Michael Kay

The MathML3 DTD is being fetched locally by Saxon's StandardEntityResolver. Because it is fetched using the classpath ResourceLoader, Saxon gives it a base URI using the "classpath" scheme. (`StandardEntityResolver#789). Xerces is objecting to this URI.

A stacktrace shows that the failure occurs in Xerces' org.apache.xerces.impl.XMLEntityManager.setupCurrentEntity() method, which is calling a java.net.URL constructor with this supplied URI.

The puzzling thing is that the application isn't obviously doing anything unusual.

#2 - 2020-10-30 12:44 - Norm Tovey-Walsh

- Assignee changed from Michael Kay to Norm Tovey-Walsh

#3 - 2020-11-30 17:15 - Norm Tovey-Walsh

I've worked out why this error occurs, but I can't (yet) offer any ideas about why it ever worked or what might have changed that caused it to stop working.

The document's DTD makes a request for the MathML 3.0 DTD. That goes to the standard URI resolver which finds it in the table of built in identifiers, it calls fetch to get it from our resources on the classpath, adjusts the system identifier to begin "classpath:...", so we know where it came from, and returns it.

Subsequently, the MathML DTD attempts to load a module. That module *isn't* in the table of built in identifiers so we fall back to trying to load the resource with the URI. At this point, we notice the "classpath:" scheme and attempt to find it in our system resources by calling getResource.

And therein lies the problem, fetch calls Configuration.locateResource which adjusts the filename so that it begins with net/sf/saxon/data but getResource just tries to find the "raw" name. When it fails to find the resource, processing falls back to the caller (deep in the guts of Xerces) where the classpath: URI scheme is, quite reasonably, reported as invalid.

I can think of several ways to fix this, the simplest seems to be to call fetch instead of getResource in the second case. Alternatively, the path that's added to classpath: could be fixed so that the additional prefix has already been applied.

A cursory examination of the history of fetch, getResource, and Configuration.locateResource suggests that this aspect of their behavior hasn't changed in at least a year so I can't say if/when this ever worked.

#4 - 2020-11-30 17:19 - Michael Kay

I think the theory is that if we have a copy of a DTD module in the Saxon JAR, then we ought to also have all its dependencies, expanded recursively. It's always been difficult to verify that this is actually the case (is there a tool that will analyse DTDs to give us the dependency tree?).

#5 - 2020-11-30 17:53 - Norm Tovey-Walsh

That's not quite the issue here. We *have* the resource, it's just not getting found.

"-//W3C//DTD MathML 3.0//EN" resolves to w3c/mathml/mathml3/mathml3.dtd which is located in the class path at net/sf/saxon/data/w3c/mathml/mathml3/mathml3.dtd

The system identifier is changed to classpath:w3c/mathml/mathml3/mathml3.dtd

MathML 3.0 includes a qnames module, mathml3-qname.mod.

We have the qnames module in net/sf/saxon/data/w3c/mathml/mathml3/mathml3-qname.mod, but it isn't part of the catalog. (That'd be another way to solve the problem, but it would make the catalog a lot bigger).

The attempt to resolve classpath:w3c/mathml/mathml3/mathml3-qname.mod fails to find the file because getResource doesn't append net/sf/saxon/data/ the way Configuration.locateResource does.

#6 - 2020-12-03 15:44 - Norm Tovey-Walsh

- Status changed from New to Resolved

- Fix Committed on Branch 10, trunk added

Fix commited to 10.x and trunk.

#7 - 2021-04-14 17:45 - O'Neil Delpratt

- Applies to branch 10, trunk added

#8 - 2021-04-14 17:49 - O'Neil Delpratt

- Status changed from Resolved to Closed

- % Done changed from 0 to 100

- Fixed in Maintenance Release 10.5 added

Bug fix applied to Saxon 10.5 maintenance release.