Infrastructure - Task #2096

Feature # 1764 (Closed): Finalize dataoneTypes schema for public release

Identifier should not be able to have any whitespace

2011-12-07 03:57 - Roger Dahl

Status:	Closed	**Start date:**	2011-12-07
Priority:	High	**Due date:**	
Assignee:	Chris Jones	**% Done:**	100%
Category:	Documentation	**Estimated time:**	0.00 hour
Target version:	Sprint-2011.51-Block.6		
Milestone:	CCI-1.0.0	**Story Points:**	
Product Version:	*		

Description

Identifier should not be able to have any whitespace, but that is currently not being enforced in the schema. Identifier is a NonEmptyString800, which is a NonEmptyString with length <= 800. NonEmptyString is a string with length > 0 and which matches "[\s]*[\S][\s\S]*". The regex allows any string that has at least one non-whitespace character.

[\s\S] allows newlines (unlike "."). In most strings, I don't think we want to allow newlines, and in the strings where we do want to allow whitespace, shouldn't it be the opposite -- that it's allowed within the text, but not in the start or end (where it may be invisible).

I think we may want to add another string type that is used only for blocks of text that may contain newlines.

So,
Identifier (and other names) would be len < 800 and one or more non-whitespace: "\S+"
Names that allow whitespace only within text and does not allow newlines would be: \S|\S.*\S
Names that allow whitespace only within text and does allow newlines: \S|\S[\s\S]*\S*

History

#1 - 2011-12-07 04:13 - Dave Vieglais

The problem description needs some clarification.

Identifiers are non-empty strings of limited length that do not have leading or trailing whitespace or non-printing characters. The restriction SHOULD be enforced by the schema, but MUST be enforced programmatically.

#2 - 2011-12-07 06:58 - Matthew Jones

the \s and \S patterns work for the common ASCII whitespace characters, but in XML schema they don't catch all of the non-printing whitespace characters that are available in other character codesets that could be used in UTF-8 strings. I felt it was non-trivial to write a pattern that properly excluded all of the non-printing/whitespace characters in Unicode. I still think this is the case. We could write a pattern that excludes the \s characters, but we still wouldn't know if it follows the spirit of the no whitespace, no non-printing characters specification.

#3 - 2011-12-12 15:36 - Dave Vieglais

- Assignee changed from Matthew Jones to Chris Jones

- Target version changed from Sprint-2011.49-Block.6 to Sprint-2011.50-Block.6

Decision: Identifiers should not have whitespace.

- Limit as best possible through regexp in the dataonetypes schema the structure of identifiers to prevent whitespace appearing anywhere in an identifier. Such a test may be limited to ASCII because of hte implementation of regexp evaluations by xml schema processors.

- Ensure that programmatic checks are made on identifiers to verify that no white space is present in new identifiers added to the system (reserve, create)

#4 - 2011-12-12 23:35 - Chris Jones

- Status changed from New to Closed

Added the NonEmptyNoWhitespaceString800 type to the schema.