Bug 766009 has been added to the database

Email sent to:

jazzytwocup@gmail.com, halfline+gnome-bug-firehose@gmail.com, asnekrs@gmail.com, hectormendez321@yahoo.com, todd@fries.net, Buster136@gmail.com, alkass1@hotmail.de, nico.behrens@gmail.com, Mendezhctr@yahoo.com, www.rzr.online.fr+bugzilla.gnome.org@gmail.com, bugbot@bugzilla.org, csis@desrt.ca, dimitri@stack.nl, anki.rocking@gmail.com, webmaster.reena@gmail.com, sag1959@hotmail.com, cloos@jhcloos.com, ggchangan@sina.cn, mccannmisty16@gmail.com, huzaifas@redhat.com, s.feltman@gmail.com, hsanand@synopsys.com, burugla_as@hotmail.com, dolphindddd@aol.com, syamsidhardh@gmail.com

Excluding:

keepyourapplesup@aol.com, tel.connect.4688@gmail.com, jsachs@nvidia.com, development@gimp-dev.opentp.org

---

Bug 766009 - Invalid source file characters passed through to XML (edit)

Save Changes

Status: NEW (edit)

Product: doxygen
Component: general
Version: 1.8.7
Hardware: Windows

Importance: Normal normal
Target Milestone: ---
Assigned To: Dimitri van Heesch
QA Contact: Dimitri van Heesch

URL:
Whiteboard:
Keywords:
Tags:

Depends on:
Blocks:

Show dependency tree

Reported: 2016-05-04 23:57:40 UTC by Jonathan Sachs
Modified: 2016-05-04 23:57 UTC (History)
CC List: Add me to CC list
0 users (edit)

See Also: (add)
GNOME target: ---
GNOME version: ---

Attachments	
doxyconfig file and source (.txt) file demonstrating the problelm. (4.38 KB, application/x-obrien-download) 2016-05-04 23:57 UTC, Jonathan Sachs	Details
Add an attachment (proposed patch, testcase, etc.)	View All

Jonathan Sachs [reporter] 2016-05-04 23:57:40 UTC Description [reply] [–]

Collapse All Comments
Expand All Comments
Add Comment

Created attachment 327320 [details]
doxyconfig file and source (.txt) file demonstrating the problelm.

With GENERATE_XML=YES, INPUT_ENCODING=UTF-8, doxygen generates XML files with the XML tag <?xml... encoding='UTF-8' ?>. However, it passes through invalid characters such as the Microsoft Windows copyright symbol, 0xA9, without change. This makes the XML file invalid.

Some XML consumers object to invalid characters. In particular, Python's XML API, xml.etree.ElementTree, throws a ParseError exception when it tries to parse the file, with a message like this:

    not well-formed (invalid token): line 11, column 73

Reproduction: The attached zip file contains a doxyconfig file and a txt file. The txt file contains a copyright symbol in line 2. Run the doxyconfig file, then open ./xml/encoding__file_8txt.xml. The copyright symbol may be found in line 11.

Suggested fix: When doxygen is asked to emit a character that is invalid according to INPUT_ENCODING:
(A) If the character is not in text (e.g. is in an embedded HTML command), issue an error or warning.
(B) If the character is in text, do one of the following:
   (1) Issue an error or warning,
   (2) Transliterate the character to a reasonable equivalent, e.g. a copyright symbol to "(c)", or
   (3) Do (1) or (2) or both according to the value of a configuration setting.

Additional Comments:

Status: NEW
Mark as Duplicate

Save Changes
