Ruby master - Bug #3686

Error in parsing musicbrainz.org with rexml

08/13/2010 12:45 AM - vinc-mai (Vincent Carmona)

Status: Closed
Priority: Normal
Assignee: kou (Kouhei Sutou)
Target version: 2.0.0
ruby -v:
  ruby 1.9.1p378 (2010-01-10 revision 26273) [i486-linux]
Backport:

Description
  =begin
  rexml (ruby 1.9.1) fails to parse this url http://musicbrainz.org/show/puid/?puid=c6a6717f-6d88-4d0e-4c57-d6b949118072.
  require 'net/http'
  require 'rexml/document'
  url='http://musicbrainz.org/show/puid/?puid=c6a6717f-6d88-4d0e-4c57-d6b949118072'
  res=Net::HTTP.get_response(URI.parse(url))
  doc=REXML::Document.new(res.body)
  /usr/lib/ruby/1.9.1/rexml/parsers/treeparser.rb:95:in rescue in parse': #<RuntimeError: Undeclared entity '&raquo;' in raw string "Skip to main content »"> (REXML::ParseException)
  /usr/lib/ruby/1.9.1/rexml/text.rb:165:in block in check'
  /usr/lib/ruby/1.9.1/rexml/text.rb:153:in scan'
  /usr/lib/ruby/1.9.1/rexml/text.rb:153:in check'
  /usr/lib/ruby/1.9.1/rexml/parent.rb:19:in add'
  /usr/lib/ruby/1.9.1/rexml/parsers/treeparser.rb:45:in parse'
  /usr/lib/ruby/1.9.1/rexml/document.rb:228:in build'
  /usr/lib/ruby/1.9.1/rexml/document.rb:43:in `initialize'
  $ ruby1.9.1 --version
  ruby 1.9.1p378 (2010-01-10 revision 26273) [i486-linux]
  ruby 1.8.7 can parse these data.
  =end

History

#1 - 08/27/2010 08:05 AM - meta (mathew murphy)
  =begin
  FWIW, I checked and &raquo; is a valid entity in XHTML 1.1, the page passes validation, and it correctly states the XHTML 1.1 DTD.
  The entity DTD also seems to be included in the modular framework module, a required part of XHTML as per
  http://www.w3.org/TR/xhtml-modularization/schema_module_defs.html#a_character_entities
  i.e. the XHTML DTD includes xhtml-framework-1.mod, which includes xhtml-charent-1.mod, which includes xhtml-lat1.ent, which defines the &raquo;
  entity.
  =end

#2 - 08/27/2010 01:05 PM - shyouhei (Shyouhei Urabe)
  - Status changed from Open to Assigned
  - Assignee set to kou (Kouhei Sutou)

  =begin
  =end

#3 - 10/08/2010 11:21 PM - vinc-mai (Vincent Carmona)

  =begin

08/06/2021
It seems that this bug was fixed. I can parse the page with ruby 1.9.2p0 (2010-08-18 revision 29036) [i686-linux] under ubuntu 10.10.

#4 - 12/07/2010 09:33 PM - kou (Kouhei Sutou)
- Category set to lib
- Status changed from Assigned to Closed
- Target version set to 2.0.0

It had been fixed in trunk.