LIBRARY OF CONGRESS

Wikipedia Link Bot

Project Charter

Version 1.0
Last Updated: April 22, 2013

Document History

Date	Author	Version	Changes

Project Title

Wikipedia Link Bot

Author

David Brunton, Ed Summers

Project Purpose / Description

Citation activity on Wikipedia generates many external links out to the Web. There is a body of literature that examines the rate at which link rot occurs on the Web. Wikipedia's Articles With Dead Links report demonstrates that the same degradation occurs on the various language Wikipedia properties as well. Wikipedia's policy recommends that dead links not be removed from articles, but instead be marked as dead. This prevents temporary routing failures from resulting in a link being removed. It also allows the URL itself to be used as metadata in locating the new location for the resource. There are also various Wikipedia bots that attempt to repair links, however most are inactive.

Meanwhile members of the International Internet Preservation Consortium including the Library of Congress are actively crawling and archiving portions of the Web. One of the deliverables of the NDIIPP-funded Memento project is the MementoProxy which provides federated access to the contents of Web Archives from around the world.

The idea of this project is to create a simple bot that could examine links in Wikipedia for two purposes: to examine dead links and see if they are available somewhere in a Web Archive, and to examine currently active links for possible inclusion in a seed list for Web Archives.

Overview of Scope

The goals:

- write a program that parses Wikipedia dumps looking for external links
- check each URL to see if it is currently available
- check each URL to see if is available at Internet Archive (and other? Web archives)
- create reports of the programs activity
- make code available so that others can run it

If results are good it is expected that data gathered would be collected and reported in a report or article suitable for publication in the Signal or elsewhere.

Content Characterization

Wikipedia Dumps made available under CC-BY-SA license. http://dumps.wikimedia.org/

Approval Signature & Date

Approved by Leslie Johnston on *June 4, 2013*