Michael Stack

stack

What is Heritrix?

- Goal: open-source, extensible, web-scale, archival-quality, web crawling software
- Collaboratively developed
  - Internet Archive, IIPC, partner libraries, and others
  - Over seven releases since January 2004
- Technical parameters:
  - IpÖ@üPDÒn many other open source libraries
  - Tested and supported on Lin‡n

Last year (IWAW’04)…

- Heritrix introduced at IWAW'04
  - – Version 1.0
  - – Core architecture, basic features in place
- Primarily useful for focused crawls
  - – Hundreds to thousands of specific target web @esÖ•
  - – Over 20 million collected URIs per crawl
  - – Crawls run for up to a week

Heritrix Releases: 1.2

- Heritrix 1.2: November 2004

  Motif: “requested features and fixes”

  – Better session-id handling (“URI canonicalization”)

  – IP politen P ~4T”

I e d f

e r ta t–

i o n P ~•Ő azd

Heritrix Releases: 1.4

- Heritrix 1.4: April 2005

  Motif: “memory efficiency at scale”

  – Much im òvd̀ seingle-machine caq6itPyÖ

    - Big data structures movd̀ teo BDB JE.

  – Customizabe 'decide rule' scoq6oePi6nPsÖ×@

  – Imq6ovPd̀×bealanced- ògress and jVàkcòn-trol (“q `VPbPV@geting”)

  – Exq6iPÖePntal “Adaqi6v Pe×v@ìt”ifrontier (b• Ksìtqinn Sgùrðsson, Ntiònsaal Libary of Iceland)

  – Exq6iPÖePntal ògrammatic remote control (va `“JiMX” Jav ìmanagement extensio`Ösà

Heritrix: Current Wo'''

Heritrix: Future Plans (2.0+)

- Bigger & faster crawling
  - Automatic multi-machine coordination
  - Improved prioritizatza

Das Ende

[illegible]

Thank you

stack@ archive.org