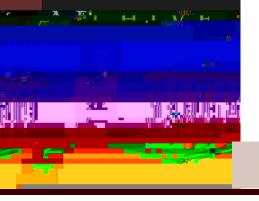
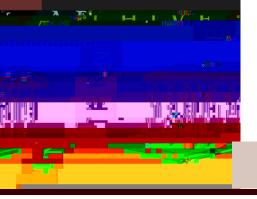


Michael Stack stack



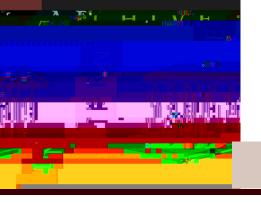
#### What is Heritrix?

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



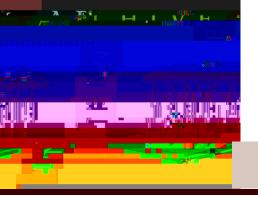
## Last year (IWAW'04)...

- 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 @sÖ•
  - Over 20 million collected URIs per crawl
  - Crawls run for up to a week



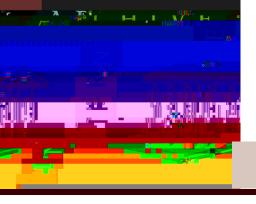
### Heritrix Releases: 1.2

- Heritrix 1.2: November 2004
  - Motif: "requested features and fixes"
  - Better session-id handling ("URI canonicalization")
  - IP politen P ⊸4T"

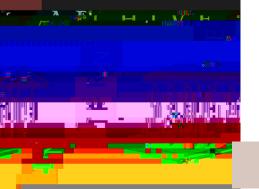


#### Heritrix Releases: 1.4

- Heritrix 1.4: April 2005
  - Motif: "memory efficiency at scale"
  - Much im ovd single-machine caquatlyÖ
    - Big data structures movd to BDB JE.
  - Customizabe 'decide rule' scoq@₽ÔPS×@
  - Imqovitybealanced-`ogress and jVàkcòntrol ("q `VVIIV Geting")
  - Exquirite Radaque Raceit "ifrontier (b. Kstinn Spùrðsson, Niòmal Libary of Iceland)
  - Exqfirflöffal `ogrammatic remote control (va`"IMX" Jav management extensio`Öà

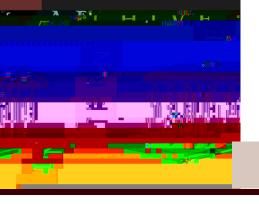


## Heritrix: Current Wo'"



# Heritrix: Future Plans (2.0+)

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



### Das Ende



Thank you stack@archive.org