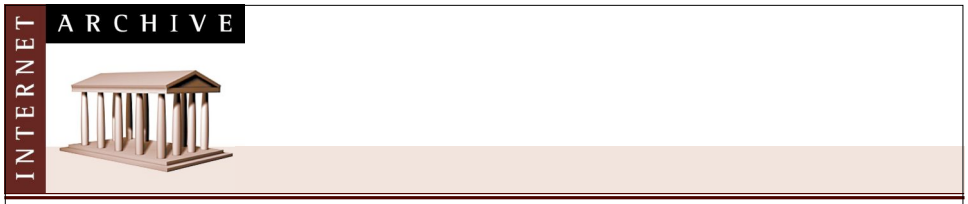



# IA/IIPC Open Source Tools Update

IWAW'06

Gordon Mohr, Internet Archive

September 22, 2006


## Heritrix Crawler

- Open source, extensible, web-scale, archival quality web crawling software
  http://crawler.archive.org
- Collaboratively developed
  - Internet Archive, IIPC, partner libraries, and others
  - Ten releases since January 2004; three in last year
- Technical parameters:
  - Java, built on many other open source libraries
  - Tested and supported on Linux
  - Web-based control console
  - Highly configurable and customizable


## Heritrix Releases, Last Year

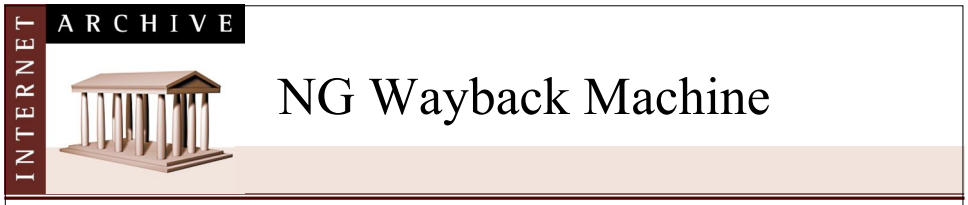
- 1.6, December 2005
  - Expanded remote control and monitoring via JMX
  - Crawl checkpointing
  - Improved performance/stability in large crawls
  - Experimental support for:
    - bloom filter already-included testing
    - partitioning a crawl across multiple independent crawlers
    - per-host/domain/queue-grouping collection quotas
- 1.8, May 2006
  - Improved checkpointing, stability, seed-centric reporting


## Heritrix Latest Release

- 1.10, September 2006
  - Requires Java 1.5/5.0
  - New configuration options:
    Scriptable modules, new DecideRules, added ordering/breadth/quota/split options, etc.
  - Web UI tweaks & improved default security
  - ExtractorImpliedURI
  - Experimental support for:
    - FTP fetching
    - WARC/0.10 handling (format design testing)
  - 43 other bugfixes


## Heritrix Futures

- Scale & Shape:
  - Targetting 2 billion-page crawl in early 2007
  - Automatic multi-machine distribution
  - Better avoiding traps and spam
- "Smart":
  - Sophisticated deduplication at scale
  - Rich prioritization by national/topical/expert input
  - Adaptive contiuous crawling at scale
- Release names, dates TBD

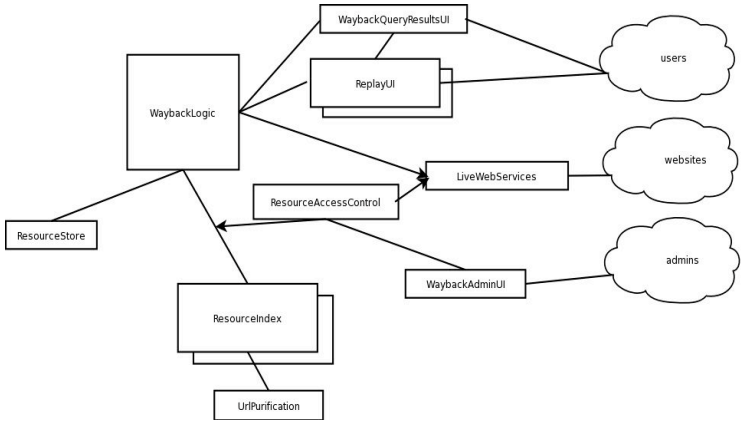



## NG Wayback Machine

- Improved classic Wayback Machine
  http://archive-access.sourceforge.net/projects/wayback
  - Open source, all Java
  - Self-contained
  - Refactored for experimentation/extension
- Core features:
  - Start with set of ARCs (index)
  - Lookup by URL/date (query)
  - Browse from lookup (redisplay)

INTERNET ARCHIVE

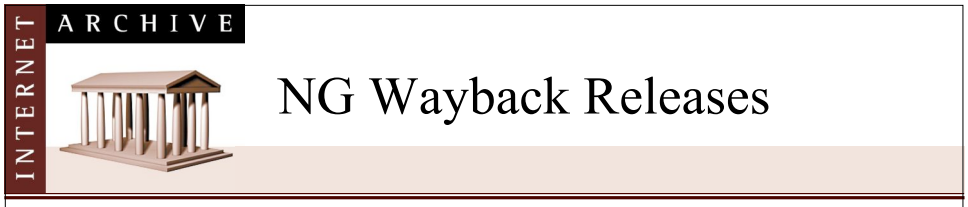
# NG WM Architecture




INTERNET ARCHIVE

# NG Wayback Limitations

- Classic WM features not yet matched
  - request missing documents from live web
  - highly distributed content index
  - document comparison

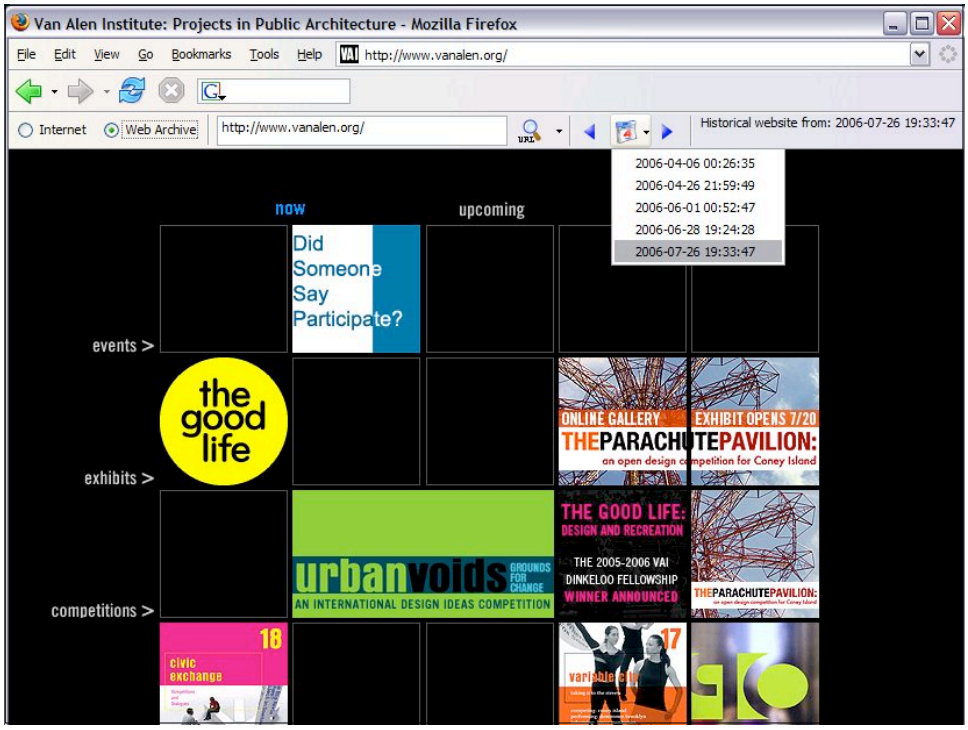



## NG Wayback Releases

- NG Wayback 0.2.0 (December 2005)
  - Skeletal functionality on single machine
  - Two “replay” modes:
    - Archival URL (http://host/DATE/URL)
    - HTTP Proxy
- NG Wayback 0.4.0 (March 2005)
  - Distributed ARC storage, retrieve via HTTP
  - Improved JS/doc rewriting in Archival URL mode
  - ResourceIndex options: Remote BDB, NutchWAX
  - live web robots.txt caching and retroactive compliance.
  - "Classic" Wayback Machine query UI


## NG WM Latest Release

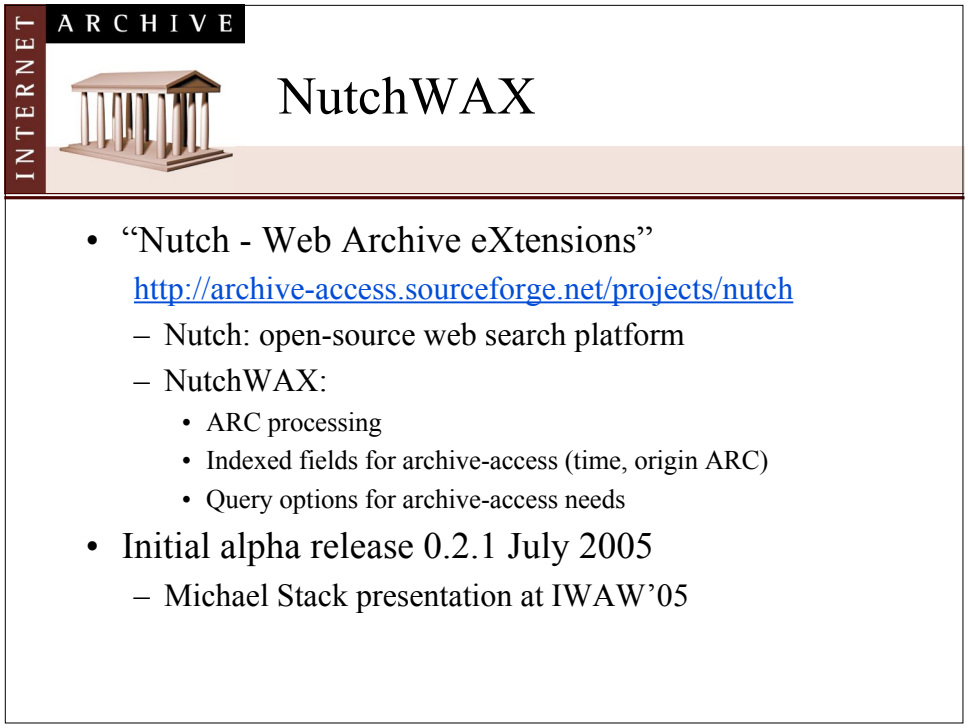
- Release 0.6.0 (July 2006) “enhanced UI”
  - Manual excludes UI
  - In-page info floater + timeline (WERA-style)
  - Proxy time-control via WaxToolbar
- WaxToolbar (April 2006)
  - Firefox extension by Oskar Grenholm, KB-SE
  - Proxy mode toggle, time-controls
  - Query box

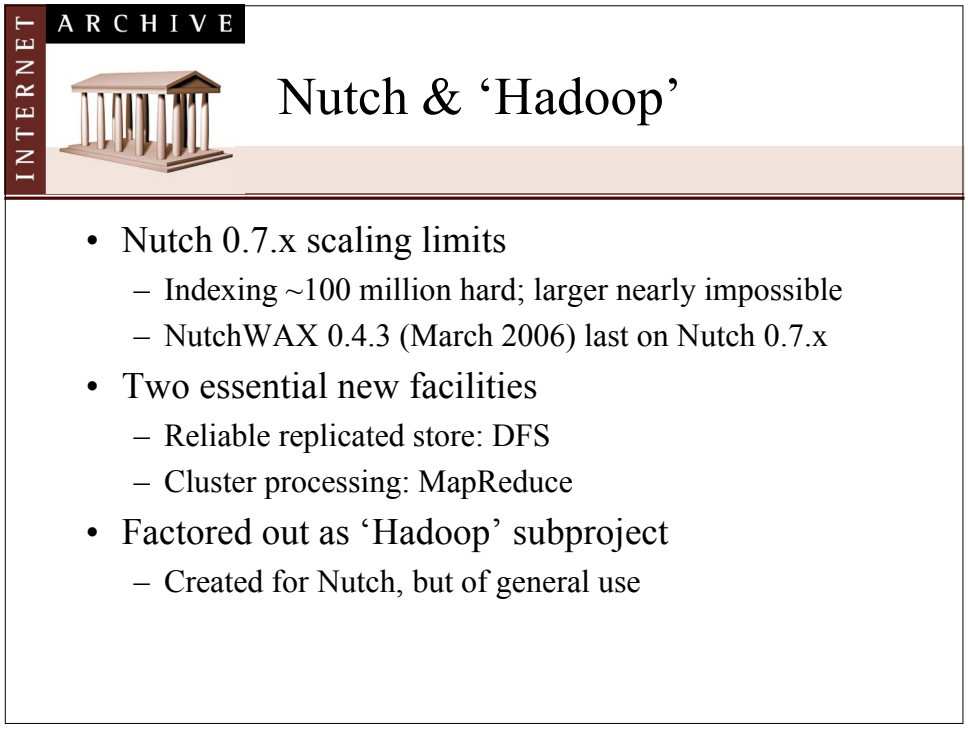



INTERNET ARCHIVE

# NG WM Future

- Release 0.8.0 (expected October 2006)
  - Distribute index over many machines
  - UI, administrative improvements
- Replace Classic WM everywhere
  - Including ‘Worldwide Wayback Machine’
  - Targeted by end of year
  - Then, 1.0 release


INTERNET
ARCHIVE
NutchWAX
• "Nutch - Web Archive eXtensions"
http://archive-access.sourceforge.net/projects/nutch
– Nutch: open-source web search platform
– NutchWAX:
• ARC processing
• Indexed fields for archive-access (time, origin ARC)
• Query options for archive-access needs
• Initial alpha release 0.2.1 July 2005
– Michael Stack presentation at IWAW'05

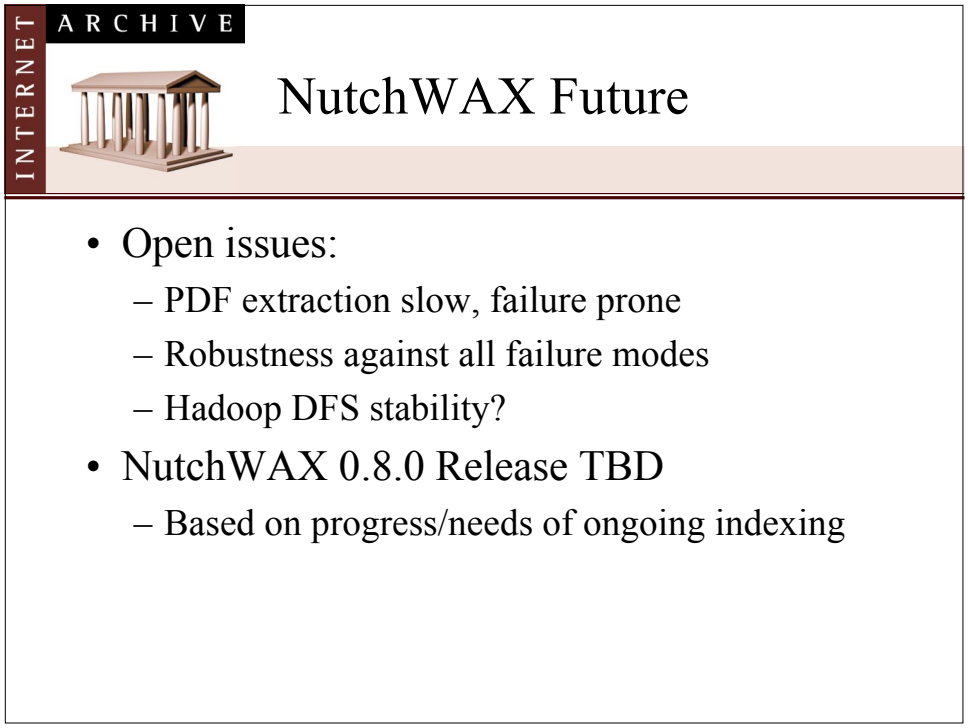

INTERNET
ARCHIVE
Nutch & 'Hadoop'
• Nutch 0.7.x scaling limits
– Indexing ~100 million hard; larger nearly impossible
– NutchWAX 0.4.3 (March 2006) last on Nutch 0.7.x
• Two essential new facilities
– Reliable replicated store: DFS
– Cluster processing: MapReduce
• Factored out as 'Hadoop' subproject
– Created for Nutch, but of general use


## Hadoop-based NutchWAX

- Development progressing
  - Doug Cutting et al @ Yahoo
  - Michael Stack @ IA
  - Nutch community
- Latest releases
  - NutchWAX 0.6 (1st on Hadoop; May 2006)
  - Nutch 0.8 (1st on Hadoop; July 2006)
  - Hadoop 0.6 (September 2006)


## Hadoop-based indexing

- Clusters & example throughput:
  - (up to) 35 puny nodes
    - Via C3 512MB
    - NARA index: 9 days, 61MM docs, 38MM unique
  - (up to) 35 “power nodes”
    - Athlon64 dual-core 4GB
    - Fast, but has had many HW/OS issues
    - E04 index: 4 days, 141MM docs, 30MM unique
    - NLA05 in progress: ~250MM docs
    - NLA06 coming up: ~500MM docs


INTERNET
ARCHIVE
NutchWAX Future
• Open issues:
– PDF extraction slow, failure prone
– Robustness against all failure modes
– Hadoop DFS stability?
• NutchWAX 0.8.0 Release TBD
– Based on progress/needs of ongoing indexing


INTERNET
ARCHIVE
Multi-billion page search?
• Everyone wants Google-style search of the IA's global archive
• But, at 60 billion captures, it's too big
– Beyond even latest Nutch software
– Massive hardware requirements
– Not even Google does a 60-billion URL index
• Idea: useful subset
– 1/2 of archive by years = 1/10th of archive by size!