The Archive-It Not-so-Secret Open Source Sauce

Gordon Mohr

October 19, 2007

Archive-It Internals

- 3 open source software projects at IA:
  - Heritrix: Crawling
  - Wayback: Browse and search-by-URL access
  - NutchWAX: search-by-text access
- On top of other open source infrastructure:
  - Linux
  - Apache/Tomcat
  - MySQL
  - Lucene-Nutch-Hadoop

Open Source?

- Open Source Initiative says:
  "Open source is a development method for software that harnesses the power of distributed peer review and transparency of process. The promise of open source is better quality, higher reliability, more flexibility, lower cost, and an end to predatory vendor lock-in."
- More than access to source code:
  Right to change, reuse, extend
- Wins:
  - Harmonize formats, practices
  - Avoid duplication of effort
  - Reduce costs

Heritrix – the beginning

- Project Inception – 2003
  - Aim: open source crawler with archival focus
    - Perfect records (“ARC format”)
    - Highly configurable and extensible
    - Excellent discovery/depth
  - Assistance of IIPC libraries in kickoff
- First release: “0.2.0” January 2004

Heritrix – evolution

- 17 releases since
- Improvements:
  - Scale: we do >500 million URL contract crawls, > 2 billion URL research crawl
  - Configuration: driven by partner needs, fine-grained scope control
  - Administration: remote-control as used by Archive-It and othr projects

Heritrix – latest

- Current public release: 1.12.1 (May 2007)
  - Theme was “duplicate reduction options”
  - Other fixes, improvements
  - Archive-It now on 1.12.1+

Heritrix – elsewhere

- Web Curator Tool
  - New Zealand, British Library
- NetArchive Suite
  - Denmark
- Web Archives Workbench
  - OCLC
- Other commercial (usually search) businesses

Heritrix – future

- ‘Smart Crawler’ work in progress
  - Sponsored by LoC, BL, BnF
  - Reduce storage, improve prioritization, optimize revisit schedules
  - WARC format – revision of ARC
- Other upcoming priorities
  - Rich media improvements
  - Spam/trap/mirror suppression
  - Automate ever-larger crawls

Heritrix – more info

- Project website
  - http://crawler.archive.org
- Source code
  - Sourceforge ‘SVN’
- Discussion
  - http://tech.groups.yahoo.com/group/archive-crawler/
- Issues/Bugs
  - http://webteam.archive.org/jira/browse/HER
- Key IA staff:
  - Paul Jack, Gordon Mohr

Wayback – the beginning

- Inception in 2005
  - Aim: URL-based browsing ‘as if’ at previous dates
  - Contrasts with classic:
    - Open source, diverse installs
    - Java vs. Perl
    - Refactored:
      - Many extension points
      - Basis for new features & experiments
- First release: “0.2.0” December 2005

Wayback – evolution

- 4 releases since
- Improvements
  - UI: inline timeline, proxy mode
  - Deployment: distributed for large collections
  - Exclusions: administrative, automatic
  - Content: better handle aggressive design, diverse character encodings

Wayback – latest

- Current public release: 1.0 (last week!)
  - Access control, discrete collections
  - Other fixes, improvements
  - Archive-It on 1.0

Wayback – future

- Accessibility – deployment options avoiding need for Javascript
- Expert modes – to handle rich media, aggressive Javascript design
- UI – better indication of changes, new ways to explore large collections

Wayback – more info

- Website

  http://archive-access.sourceforge.net/projects/wayback/
- Source code

  Sourceforge ‘SVN’
- Discussion

  https://lists.sourceforge.net/lists/listinfo/archive-access-discuss
- Issues/Bugs

  http://webteam.archive.org/jira/browse/ACC
- Key IA staff:

  Brad Tofel

NutchWAX – the beginning

- Inception in 2005
- Nutch Web Archive eXtensions
  - Based on Nutch, Hadoop, and Lucene
    - Lucene: full-text search
    - Nutch: web specializations
    - Hadoop: cluster-sized scaling
  - Read ARCs, add time dimension
- First release – “0.2.1” – July 2005

NutchWAX – evolution

- 6 releases since
- Improvements:
  - Track Nutch changes
  - Time-based queries
  - Scale: use Hadoop
- Latest release: 0.10.0, January 2007
  - Archive-It on 0.10.0+

NutchWAX – future

- Move functionality:
  - To Nutch where possible
  - To Wayback where appropriate
- Ranking improvements
- Incremental indexing
- Improved duplication-suppression
- Driven by big in-house R&D work (1.5 billion -> 30 billion)

NutchWAX – more info

- Website

  http://archive-access.sourceforge.net/projects/nutchwax/
- Source code

  Sourceforge ‘SVN’
- Discussion

  https://lists.sourceforge.net/lists/listinfo/archive-access-discuss
- Issues/Bugs

  http://webteam.archive.org/jira/browse/ACC
- Key IA staff:

  John Lee