XML Data Snapshots

Tomasz Finc
Software Developer
Wikimedia Foundation
August 27, 2009

What are they?

- Archived snapshots of wiki content encoded in XML or raw SQL
  - Most public tables are exported
    - revisions, categories, user group assignments, page links, image data, etc ..
  - Available in a number of storage formats and text combinations
    - Full page history, just current article, expanded templates and many more
      - Bz2 and 7z archives posted

Why do they exist?

- Backups of our content
  - Credibility
  - Disaster recovery
  - Transparency
- Statistical research
  - http://stats.wikimedia.org/
- Allow for bulk data re-use
- Available at http://download.wikimedia.org/

Who uses them?

- DBPedia
- OpenZIM
  - Wikipedia 1.0
- Yahoo
- Bing
- OLPC
- ...
- And many more

What was broken?

- Undocumented
- No dedicated operations staff for maintenance
- Lack of dedicated hardware
- No commitment to regular release cycle
- Frequent failure
- Long run time

What's been solved

- Documentation has steadily been compiled and added to our repositories
- We've set aside budget for a cluster of machines to provide this as a service
- We've been releasing a new snapshot every week for all 800+ projects for over two months now
- Numerous software bugs fixed and even more qa checks planned

What's been solved .. continued

- Archives are being saved off site at our new data store in Belgrade
- Monthly data snapshots available in Amazon S3 public data sets
- Live mirrors are being added as we set up more partnerships

What's still to do?

- Fix EN Wikipedia History Snapshot
  - Current runtime is in months
  - Biggest bottleneck is external storage
- Continue adding consistency checks
  - Release with confidence
- Increase awareness and usage of snapshots
  - Push snapshots to more clusters that do statistical analysis

How you can help

- Test and report back when the snapshots are inconsistent
- Log issues in bugzilla
- Let me know if any data might be missing
  - We recently added redirect tagging
  - Flagged revisions is pending to be added
- Contribute to the discussions on xmldatadumps-l@lists.wikimedia.org

Thank You

- Tomasz Finc
  - tomasz@wikimedia.org
  - http://wikitech.wikimedia.org/view/Presentations