

Peaks and Pitfalls:  
Designing a Large-Scale Repository Workflow for Quality Assurance

LITA 2007 National Forum, Denver CO

Beth Goldsmith, Irma Holtkamp, Frances Knudson, Esha Datta, and Laura Robinson  
Los Alamos National Laboratory Research Library

LA-UR-07-5288

Abstract

The Los Alamos National Laboratory Research Library's aDORe Repository is a standards-based digital object repository that currently holds approximately 80 million metadata records, 1.5 million full-text items, and several million other complex digital objects from multiple data providers, internal publications, and OAI harvests. The Repository differs from many traditional repositories because it is comprised of heterogeneous datasets that are batch-loaded rather than being ingested in small numbers. Metadata for each record or digital object is transformed to a vendor-neutral format prior to packaging in an MPEG-21 DIDL wrapper for ingestion. In addition, the Repository Team is responsible for the ingestion of weekly updates of approximately 50,000 records from a dozen different sources and for analyzing and transforming at least six new data sets for addition to the repository each year.

Throughout every step, Quality Assurance (QA) must be a crucial element to safeguard the integrity of repository contents, with quality encompassing the validity and well-formedness of data at each stage, mapping accuracy, and lossless transformation. Due to the quantity of data and datasets being handled, human QA must, by necessity, be limited; therefore, a suite of independent tools and processes have been developed and adopted to provide assurance of the repository data's quality, of the accuracy of mapping, and of the integrity of transformations. While the Repository Team's processes are focused on its workflow model, the tools and steps behind them are widely available, standards-based, and easily implemented, making them pertinent to any data processing flow, and more broadly to principles of quality assurance in general.

This presentation will focus briefly on the Repository Team's iterative workflow pipeline and the three-dimensional matrix used by the team to manage processing scale and quality. After defining categories of errors and levels of quality, it will step through human- and machine-managed QA processes. The bulk of the presentation will detail the specific tools and strategies used and discuss lessons learned from each.

## Outline

- I. Why
  - a. QA in General
    - i. Pay now or pay later
    - ii. Specifications and standardization
    - iii. Emphasis on QA for the digital library
  - b. QA Necessity at the LANL Research Library Repository
    - i. The aDORe repository  
<http://african.lanl.gov/aDORe/projects/adoreArchive/index.html>
    - ii. Local Loading
      - 1. 80 million records & growing
      - 2. Five vendors... plus
      - 3. Harvests and crawls and scanning
    - iii. Pipelining
    - iv. Quality Matrix: Data sets; Data types; Data processes
- II. What (examples of specific data problems)
  - a. Vendors don't always tell the truth
  - b. Programmers cannot test their own code
  - c. Data in context
- III. How
  - a. Who: The data analyst, generating issues to be studied and resulting in problems getting corrected.
  - b. Quality-assured metadata by design
    - i. Automatic calculation for each record
      - 1. Content validation (DTDs, Schemas, Schematron)
      - 2. XSLT
      - 3. Authorities/Controlled vocabularies
      - 4. Regression testing
      - 5. WYSIWYNTS (What You See Is What You Need To See): Filters, false-color proofs
      - 6. Heuristic-based pattern recognition
    - ii. Meaningful measurements of quality
      - 1. Lossy/Lossless
      - 2. Context
      - 3. Per tag, per dataset, per datatype, per time period
- IV. How now... or then: Bringing it all together
  - a. No data left behind
  - b. Roundtripping
- V. Hindsight: Taking it apart again
  - a. The goal in attaining data quality is not perfection but to produce data that is "accurate enough, timely enough, and consistent enough" (Orr, 67).
  - b. Metadata lifecycle model: optimizing workflow to meet repository metadata and QA requirements with available resources
    - i. Data profiling
    - ii. Local requirements vs. community requirements
    - iii. Defined policies
    - iv. Procedures to compliance

## Bibliography

- Bruce, Thomas R. and Hillmann, Diane I. (2004) "The Continuum of Metadata Quality : Defining, Expressing, Exploiting," In: *Metadata in Practice* / Diane I. Hillmann [and] Elaine L. Westbrook. – Chicago : American Library Association, 2004, pp. 238-256. (WorldCat link: <http://worldcat.org/oclc/55697067>)
- Gross, Mark and Zirilli, Donald. (1996) "The Real Costs of Conversions to SGML: Myth vs. Reality." In: *SGML '96 Conference Proceedings : Celebrating a Decade of SGML*. SGML '96 Conference, Boston, MA, November 18-21, 1996. Sponsored by The Graphic Communications Association (GCA). [Edited by] Conference Co-Chairs: B. Tommie Usdin and Deborah A. Lapeyre. -- Alexandria, VA: GCA, 1996, pp.53-56.
- International Organization for Standardization. *ISO 9000 Standard for Quality Management*. URL: <http://www.iso.ch/iso/en/iso9000-14000/index.html> (Last viewed: July 18, 2007)
- Kelly, Brian; Closier, Amanda; and Hiom, Debra. (2005) "Gateway Standardization : A Quality Assurance Framework For Metadata," In: *Library Trends*, Spring 2005, 53(4). URL: <http://www.ukoln.ac.uk/qa-focus/documents/papers/library-trends-2005/> (Last viewed: July 18, 2007)
- Olson, Jack. (2002) "Data Profiling: The Data Quality Assurance Analyst's Best Tool," In: *DM Direct Newsletter*, Dec. 13, 2002. URL: [http://www.dmreview.com/article\\_sub.cfm?articleId=6156](http://www.dmreview.com/article_sub.cfm?articleId=6156) (Last viewed: July 18, 2007)
- Orr, Ken. (1998) "Data Quality and Systems," In: *Communications of the ACM*, Feb. 1998, 41(2), pp. 66-71. URL: <http://portal.acm.org/citation.cfm?doid=269012.269023> (Last viewed: July 18, 2007)
- Piez, Wendell. (2003) "XSLT for Quality Checking in a Publication Workflow," Presented at: *XML Conference & Exposition 2003*, December 7-12, Pennsylvania Convention Center, Philadelphia, PA, USA. URL: [http://www.idealliance.org/papers/dx\\_xml03/papers/04-04-02/04-04-02.pdf](http://www.idealliance.org/papers/dx_xml03/papers/04-04-02/04-04-02.pdf) (Last viewed: July 18, 2007)
- Rosenblum, Bruce and Golfman, Irina. (2004) "Automated Quality Assurance for Heuristic-Based XML Creation Systems," In: *Extreme Markup Languages 2004 : Proceedings*. URL: <http://www.mulberrytech.com/Extreme/Proceedings/html/2004/Rosenblum01/EML2004Rosenblum01.html> (Last viewed: July 18, 2007)
- Stvilia, Besiki ... [et al.] (2004) "Metadata Quality for Federated Collections," [PowerPoint presentation - University of Illinois Urbana-Champaign, GSLIS, UIUC, Nov. 2004] URL: [www.isrl.uiuc.edu/~stvilia/papers/ASIST04-IMLS\\_poster\\_110104.ppt](http://www.isrl.uiuc.edu/~stvilia/papers/ASIST04-IMLS_poster_110104.ppt) (Last viewed: July 18, 2007)
- UKOLN. (2006) "The QA Focus Web site." URL: <http://www.ukoln.ac.uk/qa-focus/> (Last viewed: July 18, 2007). [Focuses on technical solutions to support structural & syntactical interoperability; takes lead in addressing unresolved issues in the object lifecycle.]
- University of North Texas Libraries, Digital Projects Unit. (2007) "TechTalks: Advanced Metadata Management." URL: <http://www.library.unt.edu/digitalprojects/documentation/techtalks/advanced-metadata.htm> (Last viewed: July 18, 2007)
- World Wide Web Consortium (W3C). (2007) "Quality Assurance Activity Statement," Prepared for: *May 2007 W3C Advisory Committee Meeting* (Members only) per section 5 of the W3C Process Document, Karl Dubost (QA Activity Lead). URL: <http://www.w3.org/QA/Activity> (Last viewed: July 18, 2007)