Thesis/Dissertations
Books
Software
Journal Articles
Conferences
Patents
Reports

STI-Finder Pilot

OSTI and JLab on the Hunt for JLab STI

Kim Kindrew, JLab
Kathy Waldrop, OSTI
Kathy Chambers, OSTI

Scientific and Technical Information Program
Working Meeting
Atlanta, Georgia
April 22, 2009

STI Finder

Why the STI-Finder?

- Interested in developing new technology to obtain DOE unclassified unlimited STI
- Assisted GPO in locating Federal documents not found in their system
- The GPO project resulted in the development of a new harvesting technology referred to as the STI-Finder

The Concept and Expectations of STI-Finder

- Assist STI managers in identifying obscure STI posted within their site's scientific programs and research facilities that may be bypassing the STI process
- Allow OSTI to identify and collect additional STI from DOE facilities, laboratories, and program offices that have not been submitted through established STIP channels
- Make DOE STI more easily accessible to the public and the scientific community

Moving from Concept to Development

- OSTI developed functional specifications for STI-Finder Tool
- Wrote software unique to JLab (algorithm, parser)
- Designed and tested data loader

Moving from Development to Implementation

- OSTI Identified JLab's electronic data collections for harvesting potential
- OSTI Analyzed JLab's data collections vs. OSTI's central repository

Initial Findings of JLab STI Resources via JLab Websites

- JLab's Scientific Publications Database (Public)
  - All records in this public database were harvested (1984 – present) by the STI-Finder
- Additional STI Resources Found
  - 28 document resources were identified within JLab's organizations, scientific programs, and/or departments

STI Finder

STI-Finder Harvesting Ready to Harvest from JLab's Publications Database

- What was needed from JLab for OSTI to proceed?
  - Addition of report numbers to their database records
  - Guidance on publication date field format

- What OSTI did to harvest JLab's data
  - Connected to JLab and downloaded data
  - Created XML files, uploaded, validated metadata records
  - Uploaded and cached full-text for archiving
  - Removed duplicates within data and OSTI repository
  - Eliminated non-STI document types
  - Analyzed data elements per E-link rules, resolved problems
  - Processed data through E-link test site
  - Created documentation and management reports
  - Provided files to JLab for review and subsequent release

Data Issues Encountered

- Various formats of publication dates
- Missing releasing official
- Missing site code
- Missing DOE contract number
- Empty author fields
- Authors not delimited in standard way ("; ")
- Journal title field added
- Conference/meeting data was entered in conference information field
- Empty report number fields

STI-Finder Harvesting Statistics

Total records harvested from JLab Publications Database	7,088
Total records found to be duplicates at JLab or OSTI	3,445
Duplicated records within the JLAB records	10
JLAB records found in OSTI's repository	3,435
Total records eliminated that were not STI	2,618
Total records successfully migrated into Elink test site	2,133
Total records failing migration	169
Total records w/o report numbers removed	1,277
Total records provided to JLab for approval to release	856

JLab Review of STI-Finder Harvested Records

- OSTI provided files to JLab for review
  - Good records that migrated successfully into E-link test site
  - Records with missing data in fields preventing successful migration
  - Records with data errors in fields preventing successful migration
- Jlab's next step
  - Research missing metadata fields
  - Add missing data
  - Locate full text or upload doi
  - Release records through 241.1 for public access

STI-Finder Pros and Cons

- Pros
  - Will not replace need for formal STI process
  - Provides an additional tool for STI submission
  - Identifies lab data that OSTI does not have
  - Creates bibliographic records reducing processing effort
  - Provides a way to identify and obtain STI from DOE elements
  - Provides bibliographic data not in lab STI processing systems
  - Provides full-text links for records as required
  - Identifies lab data sources to assist STI Managers with STIP activities
  - Provides reports containing data discrepancies to lab
  - Process can be automated to require little update effort
  - Is a good cost-avoidance measure for individual labs

STI-Finder Pros and Cons continued

- Cons
  - Will not replace need for formal STI process
  - Labs may wish to have more hands-on participation
  - Labs will need to review and release harvested data pulled from a non-authoritative source
  - URL's may become outdated

If the STI-Finder Proves Useful -

How Would it Work?

- OSTI would locate STI resources within lab organizations and libraries
  - An initial harvest would be done at each site as was done with JLab
    - OSTI would load harvested XML data with required fields to E-Link and if it was from an authoritative source the site might approve the data loaded going straight to release.
  - Sites could get STI-Finder harvested data by two methods
    - OSTI would load harvested XML data with required fields to E-Link as skeletal records so Releasing Officials could review, delete or release to OSTI, and/or
    - OSTI would provide labs with XML files to load into their publications/STI management databases as appropriate to enhance their database content

If the STI-Finder Proves Useful -

Then What?

- OSTI and Site would have to determine a frequency for OSTI to harvest from their sites (monthly, quarterly, etc.)
- OSTI's ongoing maintenance would include updating site resource links and routine harvesting to obtain newly posted data

Is the STI-Finder Worth Pursuing?

The STI-Finder would be a tool that could assist sites in the following ways-

- Provide another option for harvesting
- Identify organizations within a site that bypass STI process
- Identify lab STI missing from OSTI's repository
- Identify STI missing from the site's publications/STI management database
- Reduce the burden of STI submission
- Provide bibliographic data to reduce effort
- Provide full-text links for full-text documents
- Save lab resources

STI-Finder Harvesting Tool

Discussion