Metadata & Scientific Data: Integrating DDE, STTR, and ICSTI Initiatives

Presented by
Franciel Azpurua Linares
Information International Associates, Inc.
29 October 2008
The Initiatives

• DOE Office of Scientific and Technical Information (OSTI)
  – DOE Data Explorer

• Ila with DOE Office of Scientific and Technical Information (OSTI)
  – STTR Phase I: Building Blocks for an Ontology to work with an Automated Tool for Locating, Harvesting & Storing S&T Data
  – STTR Phase II: Automated Tool for Locating, Harvesting & Storing S&T Data – Automated Concepts for creating linked data infrastructure and Ontology

• International Council for Scientific and Technical Information (ICSTI) Technical Activities Coordinating Committee (TACC) Project. Participants:
  – DOE OSTI
  – Committee on Data for Science and Technology (CODATA)
  – German National Library of Science and Technology (TIB)
  – Canada Institute for Scientific and Technical Information (CISTI)
  – Univ. of Columbia Center for International Earth Science Information Network (CIESIN)
We already know...

- Since 1945 the US alone has spent over $4.2T on R&D
  - The ROI is in the use of the knowledge and information generated

- Advancing Information Technologies has transformed the scientific landscape

- The Sheer Volume of Scientific Data is overwhelming

- Digital Data is Fragile and not Always Accessible
The Current Landscape

• The products of science and the starting point for new research are increasingly digital and increasingly “born-digital”;
• Exploding volumes and rising demand for data use are driven by the rapid pace of digital technology innovations;
• All sectors of society are stakeholders in digital data management and access
What’s the Challenge?

• Scientific and technical data sets distributed in data repositories.
• Internet search engines can help but results are uneven and unreliable.
• URLs alone are not reliable locators of electronic objects.
• Data retrieved from these repositories are more accurate but such databases cannot be crawled easily by search engines and are less likely to be represented in search engine results.
• A federated search interface based on a reliable object identifier, a well-designed metadata framework and ontology can assist with this challenge.
Digital Data Collections & Research

• Digital data are increasingly important as a primary mechanism for scientific output and as a resource for new research

• Digital data collections are a powerful force for inclusion, removing barriers to participation
The Opportunity

- The Office of Scientific and Technical Information (OSTI) created and hosts interfaces that search accurate data repositories from U.S. federal agencies and worldwide organizations -such as the DOE Data Explorer (DDE).

- These efforts will be expanded to provide a tool that supports a more automated, streamlined process for creating and maintaining scientific and technical data repositories.

- Related efforts will be leveraged to provide mechanisms for annotating datasets with relevant metadata.
OSTI - DDE

http://www.osti.gov/dataexplorer/
Launched June 2008
Data collections
Fielded search & browseable
242 Collections
IIa & OSTI – STTR Phase I & II

- Generic basic STI Ontology
- Building blocks for STTR Phase II
- Prototype system that, through Digital Object Identifiers (DOI), automates search and retrieval of numeric datasets used in gray literature.
  - Sample gray literature that cites numeric data sets contained in the identified data sets in DDE.
  - Work on a process for assigning DOIs to numeric data sets and assign DOIs to the DDE data sets using this process.
  - Create DDE version 2 prototype containing the gray literature collection identified in bullet 1 including links to cited numeric data sets per document.
- Integrate basic S&T Ontology and concepts for federated searching
ICSTI-TACC

• Analyze current limitations in data search and citations
• Demonstrate state-of-the-art data citation practices already in existence
• Demonstrate the integration of numeric data sets through electronic textual information
• Recommend actions for scientific, publishing, and search communities to improve access to numeric data
The Integration

Gray Literature Documents

OSTI

Metadata and Data Repository

ICSTI TAAC

DOI

TIB
Conclusion

• These are 3 projects addressing the management of Scientific data.
• There are many others by discipline.
• Challenge is to bring the community together.
• The integration of these 3 is a start.
• We welcome additional input and cooperation.
Conclusion/Contact Information

Information International Associates
1055 Commerce Park Dr. Suite 110
Oak Ridge, TN 37931
865-481-0388 (Main Office)
865-481-0390 (Fax)
http://www.iiaweb.com

Franciel A. Linares
falinares@iiaweb.com
Office  (865) 298-1226
Mobile (865) 363-8632