First Report of ASCAC-OSTI Subcommittee

Tony Hey

Subcommittee Chair

eScience Institute

University of Washington

tony.hey@live.com

What Is OSTI?

OSTI is a program within DOE's Office of Science, with a corporate responsibility for ensuring access to DOE R&D results.

Since 1947!

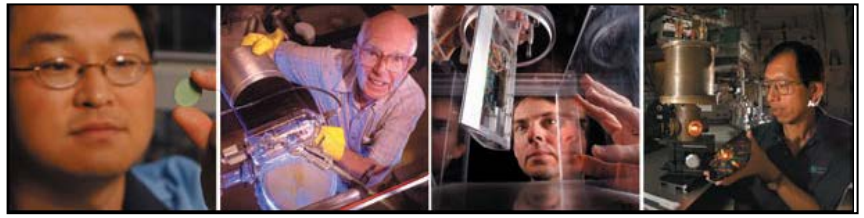
- Public access to unclassified, unlimited
- Restricted access to classified and sensitive

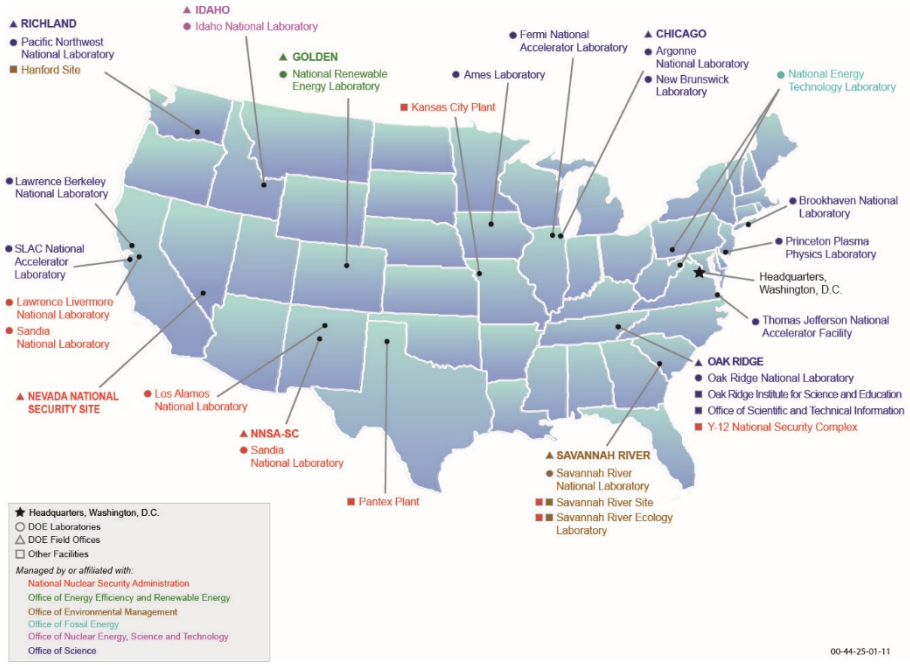
Slide courtesy of Brian Hitson

DOE STI Program

- OSTI manages agency-wide program – STIP
  - Responsibilities defined in DOE O 241.1B
  - Broad network of STI managers across DOE complex
  - R&D results collected electronically from Labs and Grantees
  - STI Submissions by year

FY12	FY13	FY14
20,205	28,793	41,867





Slide courtesy of Brian Hitson

Context: OSTP Memorandum

- Directive requiring the major Federal Funding agencies *"to develop a plan to support increased public access to the results of research funded by the Federal Government."*

- The memorandum defines digital data *"as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens."*

22 February 2013

DOE Public Access Plan, July 24, 2014

Public Access to Scientific Publications

- OSTI is responsible for operating and maintaining DOE's public access system and network
- The submission of accepted manuscripts and publication metadata to DOE will be a condition of funding
- The Department proposes to host a portal, PAGES, which will provide metadata and abstracts for publications resulting from DOE funding
- In cases where the publisher's VoR is publicly available, PAGES will direct the user to this VoR with a direct link to the publisher's website
- In cases where the publisher does not provide public access, PAGES will direct the reader to the accepted manuscript hosted in an institutional repository with a link provided by the author
- In cases where the full text is not publicly accessible through publisher or institutional repositories, OSTI will host the accepted manuscript, submitted by the author, available through PAGES.

The Charge Letter

Requested ASCAC establish a standing subcommittee for an initial period of two years to advise the Office of Science on matters associated with the DOE Office of Scientific and Technical Information (OSTI).

As its first activity, the ASCAC-STI subcommittee was asked to examine the following four questions:

a) Are the current OSTI products and services best in class and are they the most critical for the OSTI mission given the present constrained budget environment?
b) Do OSTI products and services fulfill customer needs now?
c) Are the OSTI products and services positioned to evolve to fulfill customer needs in the future? Has the OSTI strategic plan appropriately addressed the rapid evolution of technologies, research product types, and ways in which research results are communicated and shared?
d) What is the national and international standing of OSTI with respect to similar organizations whether at other U.S. Federal Agencies, DOE Laboratories, or universities? In what areas must OSTI be a clear leader to fulfill its mandated responsibilities to the DOE?

Additional Guidance from the Office of Science

Additional guidance was later provided to the subcommittee by the Office of Science in the form of four succinct questions:

- Is the mission statement sensible in the light of the statutory authorities?
- Is OSTI organized and staffed to accomplish today's mission?
- Are the current and planned OSTI products and services the correct ones?
- What suggestions would the subcommittee make for the next steps?

OSTI Mission and Authorities

Mission

Advance science and sustain technological creativity by making R&D findings available and useful to Department of Energy (DOE) researchers and the public.

Authorities

- **Energy Policy Act of 2005 (P.L. 109-58),** Section 982, called out responsibility of OSTI: "The Secretary, through the Office of Scientific and Technical Information, shall maintain within the Department publicly available collections of scientific and technical information resulting from research, development, demonstration, and commercial applications activities supported by the Department."
- **Department of Energy Organization Act of 1977 (P.L. 95-91)** provided for maintaining a central source of information and disseminating information (42 U.S.C. Sec. 5916, 7112).
- **Energy Reorganization Act of 1974 (P.L. 93-438)** defined responsibilities for developing, collecting, distributing, and making scientific and technical information available for distribution (42 U.S.C. Sec. 5813, 5817).
- **Atomic Energy Acts of 1946 (P.L. 79-585) and 1954, as amended (P.L. 83-703)** established a program for the dissemination of unclassified scientific and technical information and for the control of classified information (42 U.S.C. Sec. 2013, 2051, and 2161).
- **America COMPETES Act of 2007 (P.L. 110-69),** Section 1009, and **America COMPETES Reauthorization Act of 2010 (P.L. 111-358),** Sections 6623 and 6624, required that Federal agencies that conduct scientific research develop agency-specific policies and procedures regarding the public release of data and results of research.
- **Office of Science and Technology Policy Memorandum to Heads of Executive Departments and Agencies, February 22, 2013,** 'Increasing Access to the Results of Federally Funded Scientific Research'
- **DOE O 241.1B**, "Scientific and Technical Information Management"

osti.gov

Slide courtesy of Brian Hitson

ASCAC-OSTI Subcommittee Members

Tony Hey (Chair, UW)
Senior Data Science Fellow

- Deb Agarwal (LBNL)
  - Senior Staff Scientist
- Christine Borgman (UCLA)
  - Professor in Information Studies
- Concetta Cartaro (SLAC)
  - Experimental Physicist
- Silvia Crivelli (LBNL & UC Davis)
  - Research Associate
- Kerstin Kleese-Van Dam (PNNL)
  - Data Scientist
- Rick Luce (U of Oklahoma)
  - Dean of Libraries
- Arjun Shankar (ORNL)
  - Senior Research Scientist
- Anne Trefethen (Oxford UK)
  - CIO and PVC
- Alex Wade (Microsoft Research)
  - Director of Scholarly Communication
- Dean Williams (LLNL)
  - Analytics & Infm Mgmt Project Leader

OSTI Dissemination Products & Discovery Tools

Produce search tools that make DOE R&D results available (see www.osti.gov).

PAGES Beta

(Accepted manuscripts/articles)

SciTech Connect

(All STI)

DOE Data Explorer

(Datasets)

SCIENCECINEMA

(Videos)

DOEpatents

(Patents)

E-PRINT Network

(Pre-prints)

DOE R&D Accomplishments

(DOE's Nobel winners)

E·S·T·S·C

(Software)

Slide courtesy of Brian Hitson

Plus Three Federated Products

-- Searching multiple databases and websites with a single query --

Covers R&D and other energy information from across DOE.

Databases and websites offer over 200 million pages of science information from the U.S. government.

Provides over 400 million pages of science information from databases and portals in 70+ countries, with multilingual search of 10 languages.

Slide courtesy of Brian Hitson

Issues from the 2009 COV Report

The COV report began by commending the leadership of OSTI on its motivated and capable workforce and its spirit of excellence and entrepreneurship. However, the COV also identified a number of concerns and suggestions for improvement:

- The COV was concerned about the *'balance between its mission to provide ready access to DOE R&D results and its more entrepreneurial mission of making all scientific information available to the world.'*
- The problem of a 'leaky pipeline' led to *'the existence of less than optimum capture of DOE R&D output.'*
- A serious concern of the COV was that *'OSTI was not well known within DOE.'* One member recommended that OSTI should undertake a new action: *'listening to what DOE staff do, how they do it, and the challenges they face.'*
- The COV recognized that OSTI had '*demonstrated a clear financially efficient plan to achieve [the expansion of digital access to DOE legacy reports].'* It was suggested that DOE should incur a one-time cost to place these resources in the public domain.
- One COV member noted that *'OSTI is poised to be able to provide access to the primary literature, and to ensure its interoperability with the other publicly accessible databases it currently curates, providing a rich new resource that will facilitate new kinds of search and enable new kinds of computational research to take place.'*

Progress since 2009 COV Report

- **Re-Focus/Re-Balance in 2014**
  - Consolidated/Eliminated seven products
  - Shifted resources to
    a. improve product comprehensiveness and quality
    b. Implement public access "within existing agency resources."
  - Developed FY15-19 Strategic Plan, emphasizing
    - Core functions of collect, preserve, and disseminate; public access
    - Evolution of new and emerging forms of STI (e.g. data, multimedia)
    - Comprehensiveness, quality, and performance metrics
- **Reorganized along three core functions:**
  - Acquisition and Information Programs
  - Preservation and Technology
  - Access and Operations
- **New role in increasing public access to journal articles**
  - Developed DOE's Public Access Plan
  - Launched PAGES^{Beta} – Public Access Gateway for Energy & Science (August 2014)
  - Revised Lab and Grantee submission requirements
  - "Hired" by NSF and DoD to develop or support their public access solutions
  - Working with publishers/CHORUS; SHARE (university/library community)

Slide courtesy of Brian Hitson

Response to the Charge Questions

- Before responding in detail to the Charge questions, it is appropriate here to acknowledge the professionalism of the OSTI organization.
- All of the presenters were knowledgeable and enthusiastic about their subject area.
- From the Director down, all the staff in OSTI were very motivated and committed to delivering excellence.
- It was particularly impressive to see how OSTI had stepped up to the challenge of increased public access to research journal and conference papers.
- The staff had also thought innovatively about future developments and challenges.

Charge question (a): Are the current OSTI products and services best in class and are they the most critical for the OSTI mission given the present constrained budget environment?

- OSTI's products and services are professional and generally well done, and it is welcome that they are now more focused on the search and discovery of DOE STI R&D.
- OSTI is a charter member of CENDI – a US interagency working group of senior STI managers. OSTI operates Science.gov, CENDI's flagship, cross-agency STI product for searching US Government scientific and technical information.
- OSTI was the first US federal agency to be a member of the DataCite organization. With its Data ID Service, OSTI can now provide researchers with DOIs for their datasets. These datasets are then made available to users via OSTI's Data Explorer database. The Data ID Service is clearly a valuable and forward-looking service offered by OSTI.
- Many of OSTI's services do incorporate leading-edge technologies and, in this sense, can clearly be regarded as 'best in class':
  - SciTech Connect offers a powerful semantic search capability
  - ScienceCinema video service integrates automated audio-indexing technology
  - WorldWideScience.org uses automated language translation technology.
  - OSTI also offers an innovative federated search capability in NLE^{Beta}, Science.gov, and WorldWideScience.org.

Charge question (a): Are the current OSTI products and services best in class and are they the most critical for the OSTI mission given the present constrained budget environment?

- The ScienceCinema video indexing service is an interesting addition to the SciTech tool and contains a significant collection of videos and has implemented a nice search capability. However, additional work on improving metadata such as source credits and information is needed. It was also not clear what analysis was done to determine that video was the most critical priority for collection versus images, audio, etc.
- OSTI has made great strides recently and their most recent product releases appear to have elements that are best in class. However, a focus on unifying the product set, reducing redundancy, and improvements in content coverage will be required before they can claim to be fully best in class.
- One OSTI service that was found <u>not</u> to be best in class is the ESTSC software service. This service and its software inventory seems very outdated and out of touch with the leading DOE research software developers. The ESTSC model of charging for software seems unlikely to be an attractive offer given the open source culture of the scientific research community. GitHub is an interesting example of a modern software repository with tools that support the open source community software development process.

DOE PAGES^{Beta} and OSTI's Public Access Role (1)

- In terms of providing public access to the full text of journal articles, the NIH, with its National Library of Medicine and its PubMed Central repository, has a significant head start on all other federal agencies.
- The NCBI PubMed Central site also provides access to a large number of specialist biomedical databases. Their Entrez cross database search engine can then find supplementary information relevant to the PubMed Central article.

➢ At present then, the NIH PubMed Central service must be regarded as best in class among the federal agencies.

- OSTI's rapid development of the DOE PAGES^{Beta} service and the associated further development of their E-Link service for submission has been impressive.

➢ This role in implementing the DOE's Public Access Plan is clearly critical for the fulfillment of OSTI's mission.

The CHORUS Consortium

From the DOE Public Access Plan:

- ‘The publishing community is developing a multi-publisher portal, the Clearinghouse for Open Research of the United States (CHORUS), to provide access to journal articles resulting from government funding.’
- ‘Such an activity offers considerable economies in the integration of article metadata and links for publishers who want to participate in DOE’s public access efforts.’
- ‘PAGES, however, can operate successfully independent of CHORUS.’

- Still much skepticism from academic library community about the DOE collaborating with CHORUS
- Elsevier just announced a change to its long-standing 2004 policy increasing the 12-month embargo period to a 12 – 48 month period
- Institutional repositories are concerned about the confusion this is causing
- The academic research libraries are also developing their own software portal for open science called SHARE

Publisher non-compliance: back-up plans

According to the DOE Plan:

- ‘During an “administrative interval” of up to twelve months, PAGES will not provide access to the full-text manuscripts. During this time, metadata including links to the publishers’ VoR will be discoverable through the PAGES search interface and via PAGES APIs’
- ‘PAGES will automatically reconcile DOIs submitted by DOE authors and by publishers to determine whether the VoR is accessible by the end of the administrative interval. In cases where the VoR is not accessible, PAGES will display a link to the accepted manuscript.
- ‘In all cases, OSTI will maintain a dark archive of manuscripts to be used in the event links become broken or full text access is otherwise interrupted or discontinued. The dark archive will be part of the Department’s Enterprise Data inventory.’

DOE PAGES^{Beta} and OSTI's Public Access Role (2)

- At present the DOE PAGES^{Beta} service must be regarded as a promising newcomer to the public access agenda.
- Perhaps surprisingly, it already seems clear that OSTI's DOE PAGES^{Beta} service could prove to be an attractive alternative solution to that offered by the NIH.
- Both the NSF and DOD funding agencies are working with OSTI and are committed to following the PAGES solution to implement their public access plans.
- The DOE PAGES^{Beta} could emerge as best in class in a few years and the collaboration with NSF and the DOD represents a great opportunity for DOE and OSTI

Charge question (b): Do OSTI products and services fulfill customer needs now? (1)

- In order to answer this Charge question, it is necessary to separate the different types of customers being served by OSTI's products. We note that:
  - From the usage statistics there appears to be significant take-up of OSTI services by the public and by commercial services.
  - Unfortunately, OSTI's services do not appear to be widely used by the DOE researcher community, a community that is specifically called out in OSTI's mission statement.
  - In respect of outreach to the DOE researchers, very little seems to have changed since the COV of 2009.
  - As a result the services are more targeted towards librarians than researchers and the OSTI services seem cumbersome by comparison to existing domain specific solutions (e.g. the INSPIRE system at SLAC).
- OSTI provided comparative evidence for the quality of its services but it sees its natural peer organizations to be other national library services, rather than community archives (e.g. Earth Systems Grid Federation, High Energy Physics INSPIRE/arXiv and Astronomy ADS), or public sharing sites such as ResearchGate, GitHub or NanoHub, which offer very different and more interactive ways of sharing knowledge.

Charge question (b): Do OSTI products and services fulfill customer needs now? (2)

- We note that certain classes of customers (such as major DOE data program managers in ARM and CDIAC) are satisfied with existing OSTI products and services, and see OSTI as one of the few solutions (and the only DOE supported one) for their needs. They also see OSTI as being open to tailoring and improving their services based on their needs.
- Researchers see the need for more integration of the different services and an improved user interface. Using user-interface simplification best-practices to unify and reduce redundancy in the toolset could make the user's experience of OSTI service better for researchers.

Comment

- Further development of OSTI products targeted at DOE researchers must involve dialog with the DOE research scientists.

Charge question (c): Are the OSTI products and services positioned to evolve to fulfill customer needs in the future? Has the OSTI strategic plan appropriately addressed the rapid evolution of technologies, research product types, and ways in which research results are communicated and shared?

- In their presentations, the OSTI staff showed a good awareness of the likely evolution of the services with respect to linking publications to data. The Data ID Service is a useful start and OSTI's involvement with CrossRef, FundRef, DataCite, and the ORCID organizations will be valuable.
- OSTI's vision for providing '*named user*' functionality could clarify the issue of which part of their customer base they are addressing. Adding some *'social functionality'* could also help bring the interactivity of OSTI services up to the level of the best research community tools.
- The emerging challenge of collecting electronic versions of graphs, tables, and images in papers does not seem to be currently addressed in detail in OSTI's plans. Although data, software, images and video are all part of OSTI's larger data vision it must be emphasized that OSTI and the DOE STIP community must work closely with the DOE research community if they are to develop new useful services for today's researchers.

Charge question (d): What is the national and international standing of OSTI with respect to similar organizations whether at other U.S. Federal Agencies, DOE Laboratories, or universities? In what areas must OSTI be a clear leader to fulfill its mandated responsibilities to the DOE? (1)

- OSTI has a leadership role with the CENDI interagency group in operating the Science.gov gateway to government science information. This offers a federated search service across 60 scientific databases and 200 million pages of scientific information.
- The recent development of the DOE PAGES^{Beta} service for access to research journal articles has won the respect of the NSF and the DOD (DTIC) who are considering using the OSTI system for delivering their public access plans. <u>This is an area in which OSTI must be a clear leader to fulfill its mandated responsibilities.</u>
- The OSTI services employ a range of innovative technologies not uniformly available from their peer international scientific information organizations.

Charge question (d): What is the national and international standing of OSTI with respect to similar organizations whether at other U.S. Federal Agencies, DOE Laboratories, or universities? In what areas must OSTI be a clear leader to fulfill its mandated responsibilities to the DOE? (2)

- In terms of international leadership and recognition, OSTI is a founder member of the WorldWideScience Alliance and is responsible for providing novel real-time searching and translation services over globally-dispersed multilingual scientific literature to the other Alliance members.
- OSTI products and services compare well with those delivered by similar organizations in Canada (CISTI), France (INIST-CNRS) and the German National Library of Science and Technology (TIB).
- Because of its recognized expertise in information management, OSTI was invited to chair the Technical Activities Coordinating Committee of the International Council for Scientific and Technical Information (ICSTI).

Response to supplementary questions (1)

1. **Is the mission statement sensible in the light of the statutory authorities?**
   - The OSTI mission statement is entirely appropriate in targeting DOE researchers and the public.
2. **Is OSTI organized and staffed to accomplish today's mission?**
   - The recent re-organization of OSTI in terms of its three core functions has given OSTI clearer focus on DOE research results.
   - It is likely that some changes to the mix of technical expertise at OSTI will be required to design and develop services suitable for modern science environments. This could be acquired either through new hires or by collaboration with existing DOE lab researchers and librarians.
   - If OSTI is to take on a larger role with respect to data it needs to expand its expertise in this area.

Response to supplementary questions (2)

3. Are the current and planned OSTI products and services the correct ones?

- The products and services need to be targeted for at least three different communities – the traditional library and information management community, the DOE research community, and the general public.
- The automated collection of publications and provision of public access versions should remain a top priority for OSTI. Optimization of the publication collection method could significantly reduce the burden on the data submitter (e.g. requesting only the DOI, organization, funding info, and a pdf) and significantly increase collection completeness.
- The start on collecting multimedia content is valuable but improvements in metadata and consideration of the priorities of different types of multimedia should be undertaken.
- The Data ID Service is a critical first step towards making datasets citable and linking data to publications. Other data collection, federating, and brokering services may be the next step.

Response to supplementary questions (3)

4. What suggestions would the subcommittee make for the next steps?

- Initiate some serious two-way outreach and dialog with the DOE Labs research communities to better understand what services they would like and use.
- A more detailed analysis of Google/Bing search results on DOE R&D could help determine in which areas OSTI should focus to deliver complementary functionality.
- Discuss tool usability issues with the DOE research community with a view to developing an integrated 'one stop shop' approach to STI services
- Enlarge the STIP management by 'researcher champions' from each Lab. The lab library staff need to work with researchers to understand the issues of research reproducibility and open science that require linking data and software to research publications.
- With the STIP management, OSTI needs to develop the necessary skills to advise researchers about the required Data Management Plans. This could include discussion about possible data repositories for long-term storage of large data sets as well as how to release sufficient data to support the conclusions of the journal article.
- Discuss approaches to **partner** with the DOE labs and researchers to improve content completeness and help reach the DOE goal.

Recommendations (1)

1. If OSTI is to truly fulfill its mission to create products and services to make '*R&D findings available and useful to DOE researchers*', it needs to initiate a vigorous outreach program with the DOE Lab researchers. This must involve listening to researchers needs and understanding the strong and weak points of existing community sites.
2. To promote a successful implementation of the public access requirement issued by OSTP, OSTI needs top-down support from DOE in clearly communicating that this is not a requirement/burden imposed by OSTI but rather a government-wide and DOE-wide requirement meant to share federal research results and accelerate scientific progress. In this regard, labs, grantees, and their authors need to be incentivized to comply with this requirement, which OSTI is helping them to fulfill, and one such incentive could be a measurable expectation expressed in labs' annual performance plans.

Recommendations (2)

3. OSTI should work with the DOE research community to re-invent the ESTSC software service. With the key DOE scientific software developers on board – such as the linear algebra '*LAPACK team*' and the parallel computing message passing '*MPI team*' - this could be a very popular service internationally. In addition, releasing software in support of a research publication needs to be supported.
4. Work with the Labs to identify 'researcher champions' who can work with the STIP community to strengthen the link to researchers. This could include advice on Data Management Plans and target data repositories.
5. OSTI should work aggressively to continue toward a unified user environment with a limited number of, clearly delineated, non-redundant tools and develop a master plan for future development and areas of expansion through community input.
6. Through partnership with the national lab librarians and researchers identify and address publication content gaps and develop clear instructions and guidelines regarding content submission requirements. Significantly improving the completeness of coverage of the publications collection will require creative solutions.

Towards a Data Management Framework for DOE (1)

Question: Could there be a useful role for OSTI and the STIP management team in developing and advising researchers on DMPs appropriate for their research projects?

Some suggestions:

- OSTI could work with Office of Science Programs (BEM, BER, BES, HEP, …) and the different research communities in the DOE Labs to develop better solutions for linking data and software to publications.
- OSTI could participate in reviews of the data needs by discipline and identify explicit commonalities and differences between disciplines
- OSTI could participate in collaborative pilots that establish the open data and open science end-to-end infrastructures (data provenance, data workflows, experiment integration).
- OSTI may be able to act as a brokering, clearinghouse instead of the core infrastructure manager for some of these services.
- OSTI must become a clear leader in helping the data policy. It needs to actively strategize on partnering with the national laboratories, industry, and academia to accomplish this.

Towards a Data Management Framework for DOE (2)

Suggestions continued:

- OSTI could assist in the development of an evaluation plan to assess how well the DMP and services support the community.
- OSTI could also develop cost models for manageable and cost-effective data solutions.
- Between a new, relevant ESTSC software service and the data ID service there is a clear opportunity for OSTI to pursue a full integration of software and data together with the link to the published article. This would really bring the words '*Public Access*' to a whole new level.
- Finally, could OSTI play a similar role in the US as the UK Data Curation Centre (DCC)? The DCC offers advice on data issues such as metadata, curation, and preservation options. It also developed the first set of DMPs accessible via their DMPOnline tool.

➢ OSTI would need to develop some significant additional expertise to play the coordinating and advising roles suggested above.

Epilogue

- The DOE Public Access Plan commits the Office of Science to leading the way for DOE in the development of a single DOE-wide policy for data management planning
- Phased implementation of the requirements of the OSTP memo:
  - Pilot data management policy for the Office of Science by July 28, 2015
  - Single DOE-wide policy by October 1, 2015

OSTI has analogues . . .

U.S. federal agencies and other countries have OSTI counterparts to collect and disseminate their STI:

Examples:

U.S.

- National Library of Medicine (NIH)
- Defense Technical Information Center (DoD)
- NASA STI Program
- National Transportation Library (DOT)
- National Library of Education

International

- Canada – Canada Institute for Scientific and Technical Information (CISTI)
- China – Institute of Scientific and Technical Information of China (ISTIC)
- France – National Institute of Scientific and Technical Information (INIST)
- Germany – German National Library of Science and Technology (TIB-Hannover)
- Japan – Japan Science and Technology Agency (JST)
- Korea – Korea Institute of Science and Technology Information (KISTI)

(Related to Charge Letter, question D: "What is the national and international standing of OSTI with respect to similar organizations . . .")

Slide courtesy of Brian Hitson