Office of Scientific and Technical Information

U.S. Department of Energy

The Search for Hidden Treasure:

R&D Results in the Deep Web

Dr.Walter L. Warnick, Director

Office of Scientific and Technical Information

STIP Meeting

April 30, 2003

www.osti.gov

Interagency Partnerships

Leveraging STI Resources for DOE

- Long-standing agency-to-agency partnerships with GPO, NTIS, and DTIC
- DOE representative of CENDI and Science.gov Alliance

Science.gov Is Latest Success Story

Science.gov Launch: December 2002
AAAS Demonstration: February 2003

- **Science.gov Alliance of 14 information offices from 10 major science agencies is similar to the DOE STIP collaboration.**
- **OSTI is DOE's representative, and we are the "home" of science.gov.**
- **Web portal indexes over 1400 resources, with DOE S&T in over 400 DOE URLs.**
- **Deep Web searching of 30 databases is a key feature, developed by OSTI.**
- **Deep Web search tools offer greater access to R&D results.**

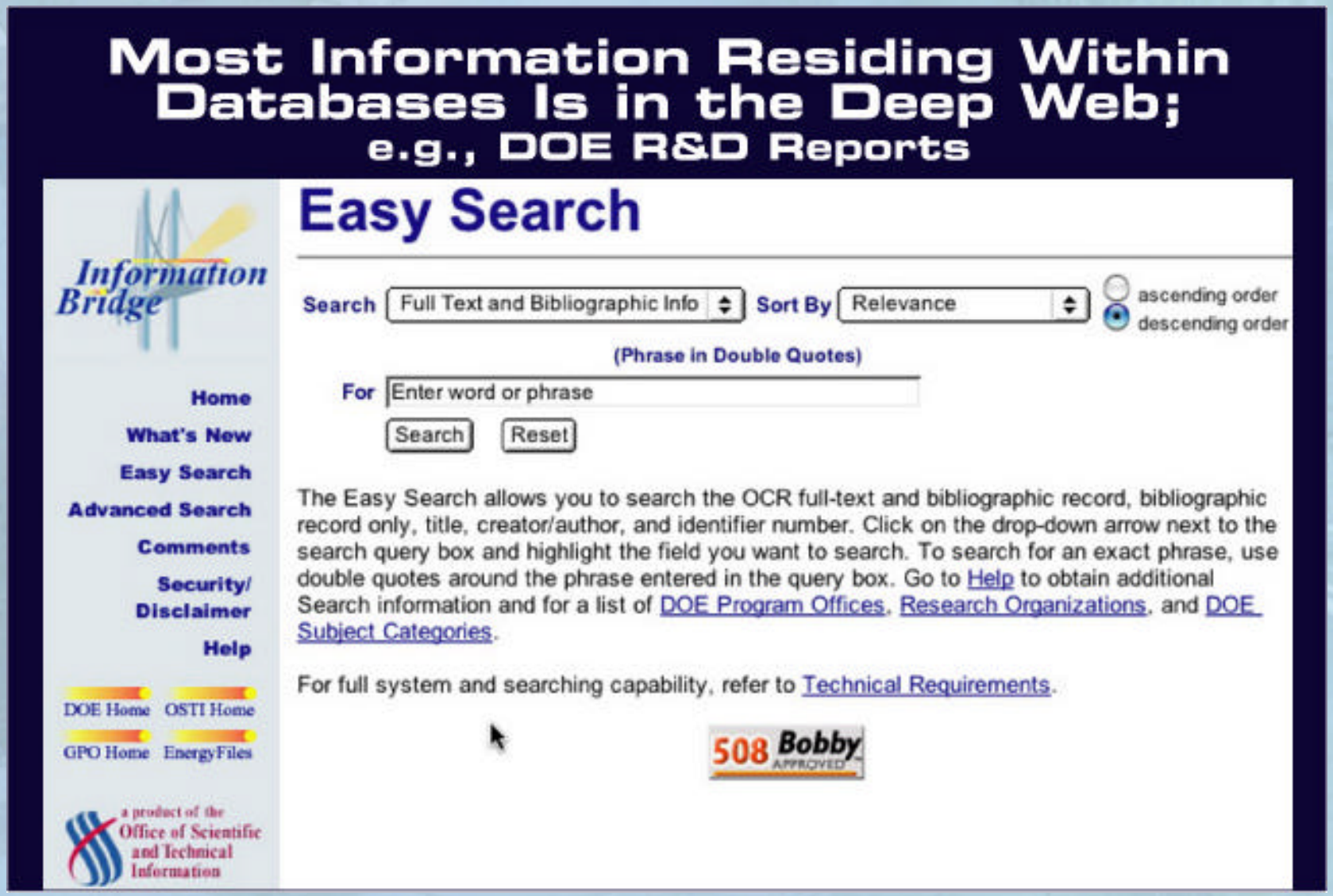
Science.gov aims to bring the substantial resources of the federal science and technology enterprise together, in one place. Working together, federal agencies have assembled countless pages of government research, data, and reports. The site is a great example of e-government in action.

-Dr. John H. Marburger, Director, Office of Science and Technology Policy

Surface vs. Deep Web

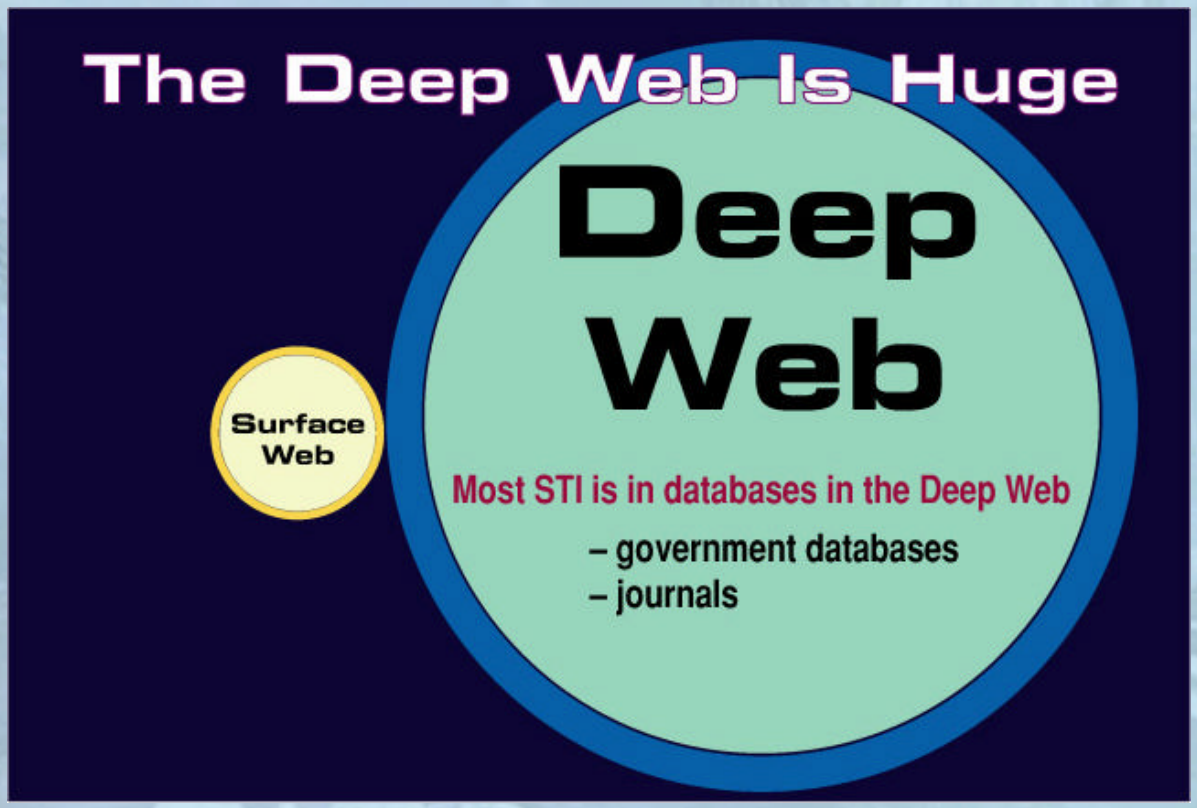
Database Searching

Most Information Residing Within Databases Is in the Deep Web; e.g., DOE R&D Reports



Volume of Information

- The Surface Web has 3 billion pages and is accessible to crawlers of popular search engines.
- The Deep Web is not accessible to such engines, but it is hundreds of times larger than the Surface Web.



Relevance of Results

Among popular search engines, Google is experiencing increasing traffic. Why?

The reason is relevancy ranking –

The item you want is most likely to be among the first ones listed.

To increase traffic in DOE R&D results, make the relevant information easy to find.

Information that is easy to locate is information that is much used.

Approaches to Searching

- For the Surface Web: Web Crawlers
- For the Deep Web: Distributed Search

(The differences are shown in the following charts.)

A Web Crawler Starts by Building an Index (i)

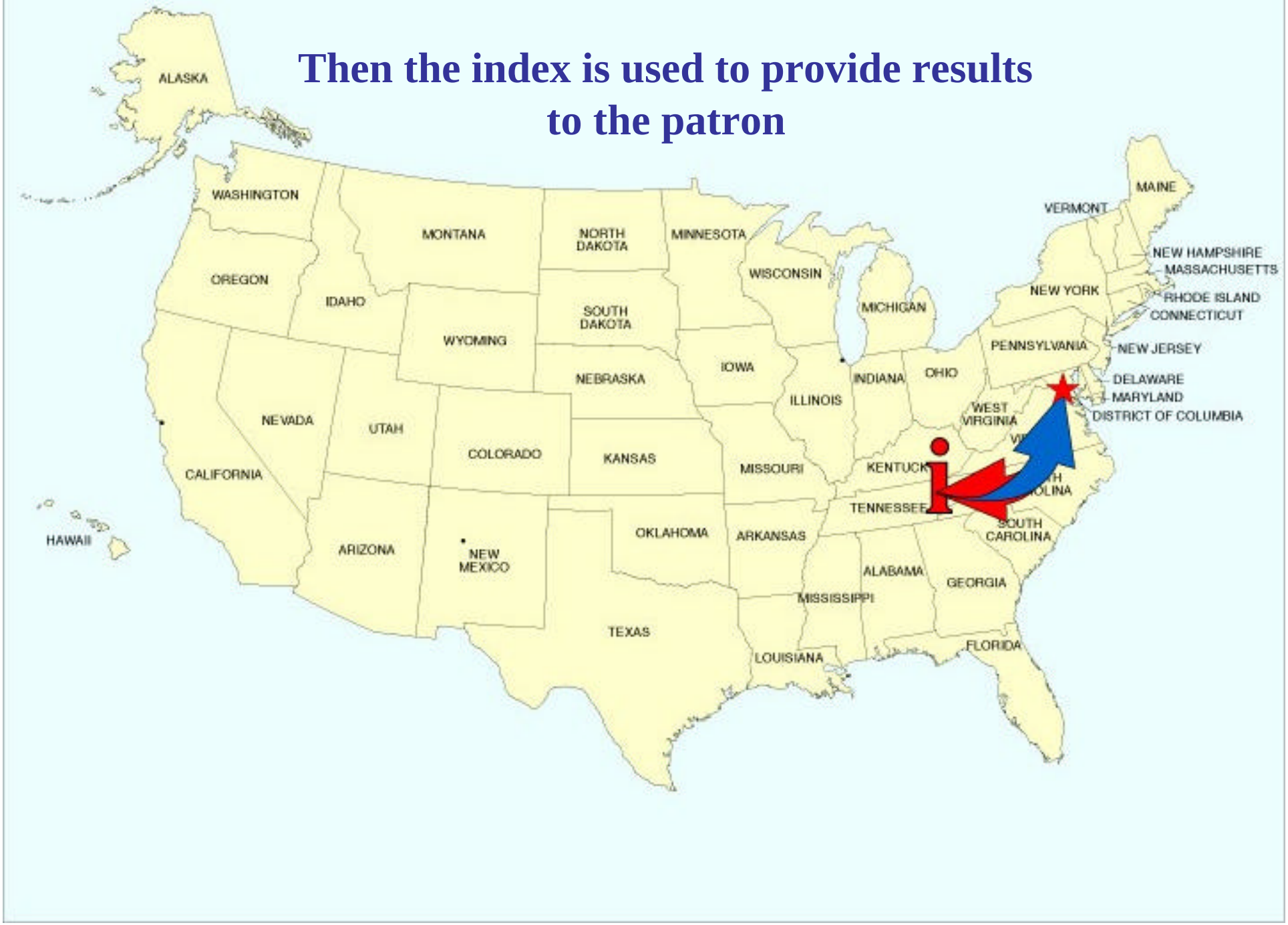
Information is gathered from a site

And added to the index

Then information is gathered from another site

And added to the index

Then the index is used to provide results to the patron



The Problem with Crawlers

Global search tools for the surface web, e.g., Google:

- Seldom allow fielded searches, i.e., searching author, title, etc.
- Do not include databases of the Deep Web.

To search across databases of the Deep Web, distributed search is the only way.

Global distributed search tools for the Deep Web are not available.

The Challenge

Challenge for OSTI:

Make the Deep Web as searchable as the Surface Web

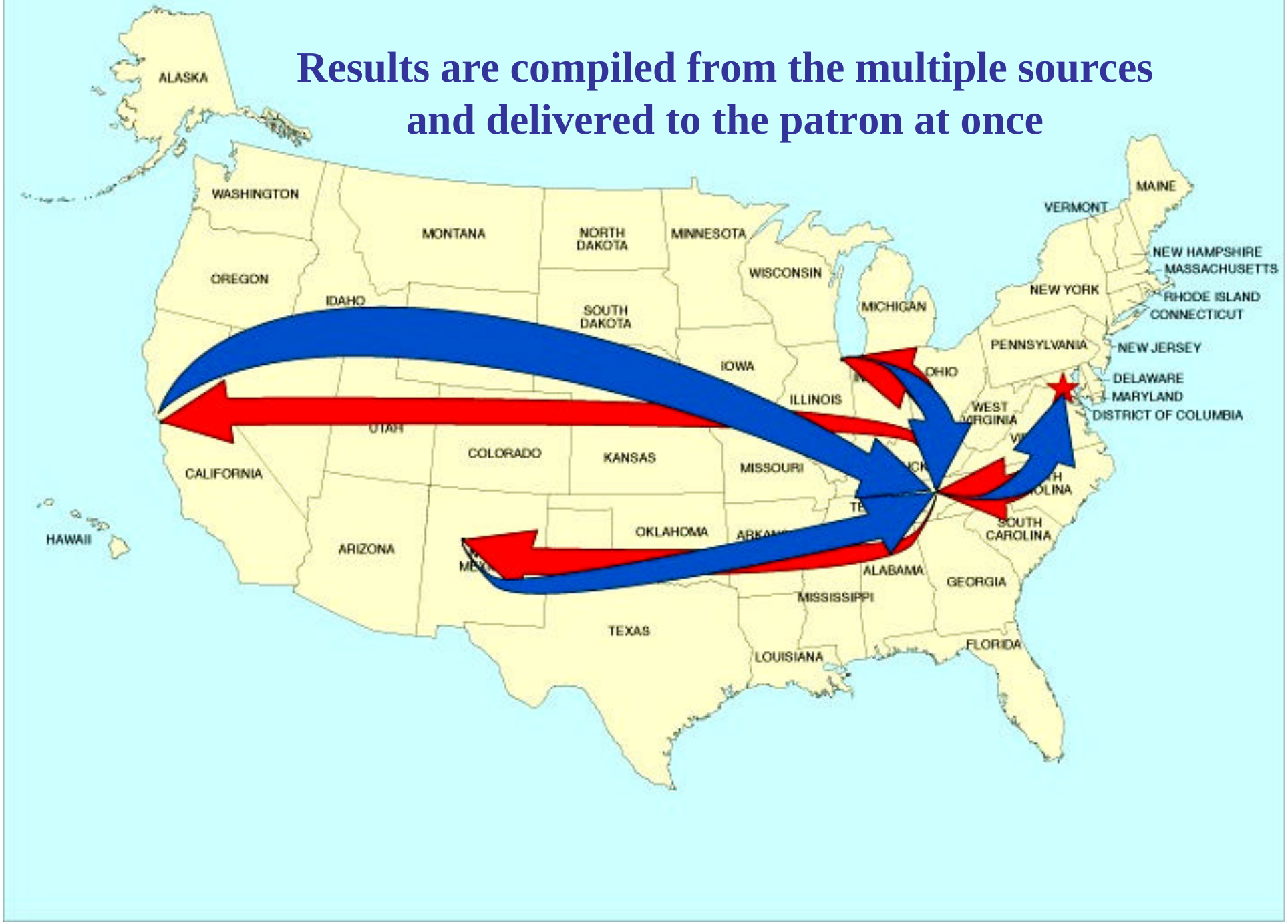
OSTI Strategies:

1. Develop Deep Web search capabilities – described in the remainder of this presentation
2. Create Surface Web Links to Deep Web Resources – deferred to a future presentation

A Distributed Search Starts with the Patron's Query

At time of query, the search goes simultaneously to multiple databases

Results are compiled from the multiple sources and delivered to the patron at once



Current Limitations of Distributed Searching

Distributed searching of the Deep Web has great strengths, but it is not a panacea.

- It is hard to determine if a key database has been overlooked.
- User may receive a glut of information to sort through due to lack of relevancy ranking of search results.

The focus of OSTI in the near term is to strengthen distributed searching tools for the Deep Web.

Benefits from STIP and for STIP

The STIP Collaboration accrues benefits beyond DOE, in the interagency and international STI circles.

- **The distributed electronic environment pioneered by all of us here in DOE is now at work for U.S. science information**
- **The harvesting system initiated with some of you now will soon be tested with our Nordic exchange**

Today's Agenda

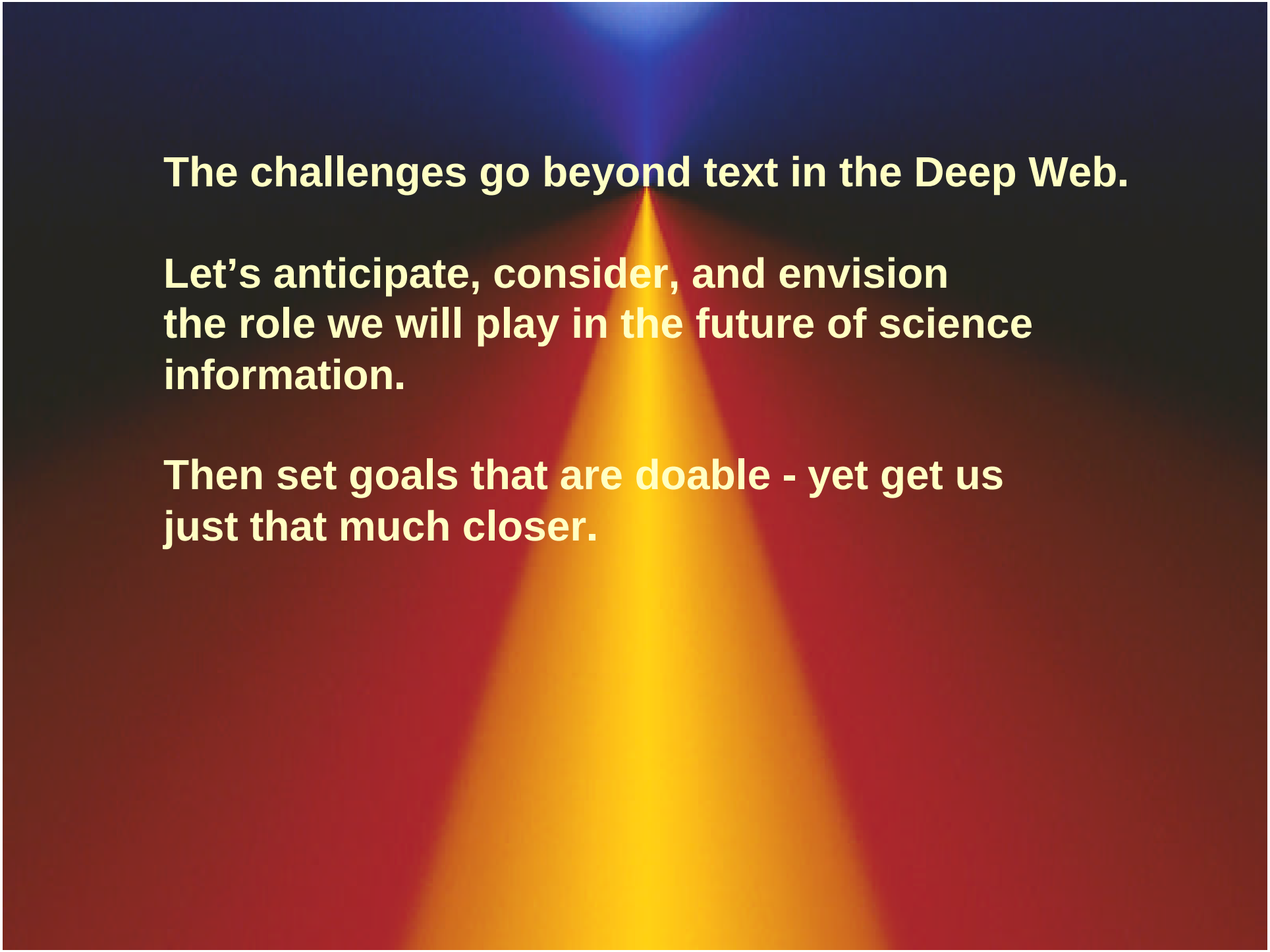
As you work these topics during the meeting:

- Harvesting
- Products and services
- Institutional repositories
- Data and full text processing
- Forms and formats
- Preserving legacy collections
- Homeland security
- Goals and metrics
- And a range of other items, large and small

Keep in mind that each puzzle piece is needed to to complete the picture.

The STIP Partnership Has Come a Long Way

With Electronic Documents conquered, what will the next challenge be?


The challenges go beyond text in the Deep Web.
Let’s anticipate, consider, and envision
the role we will play in the future of science
information.
Then set goals that are doable - yet get us
just that much closer.