

# IATUL

Creating an institutional repository for massive data sets -- a statement of the problem and an assessment of the challenge and opportunities.

James L. Mullins, Ph.D.  
Dean of Libraries  
Purdue University  
May 30, 2005



# The Problem

Exponential Growth in the creation and size of massive (mega) datasets generated by research in genomic, chemical, molecular, demographic, environmental, drug and medical research among many others.



## The Challenge

- Access to, searching and preservation of mega datasets straining capacity of information technology resources.
- Along with equipment and infrastructure -- need for classification structure and protocol to retrieve, compare, and integrate the data.



# The Opportunity

Opportunity for collaboration among the academic departments and research centers, the information technology unit and the libraries at the university.



# Scope

One data source - Doppler radar –  
generates data at the rate of  
one Terabyte per day



# Scope

One data source - Doppler radar –  
generating at the rate of terabyte per day

Just how much data  
is that?



## Scope

- 3 ½ inch floppy = 1.44 Megabytes
  - 100 Megabytes - several encyclopedia volumes
- 1 Gigabyte = 1000 Megabytes – content of 360 books
- 1 Terabyte = 1000 Gigabytes - 1000 copies of the Encyclopedia Britannica, ten Terabytes could hold the printed collections of LC



# Scope

- 1 Petabyte = 1000 Terabytes =
  - 20 million 4 –drawer filing cabinets or 500 billion printed text pages



## Scope

- 1 Petabyte = 1000 Terabytes = 20 million 4 – drawer filing cabinets or 500 billion printed text

Exabyte ... Zettabyte ... Yottabyte ...  
Brontobyte

From an article written by R.L. Stock, PC's 'n' Dreams, 2005



# The Environment – Purdue University

## **History**

Founded 1869 in West Lafayette, Indiana, and named after benefactor John Purdue

## **Student Body**

38,653 total students (Fall 2004); 7,906 graduate – from all 50 states and 130 countries.

International student enrollment – 4,921 3rd largest in US

## **Faculty & Post docs**

1,767 faculty and 740 post docs (Fall 2004)



# The Environment – Purdue University

*U.S. News & World Report*, Spring 2005, ranked Purdue's [College of Engineering](#) - 10th overall nationally; Recruiters ranked Purdue Engineering 8th (tied with Cornell and Michigan)

Purdue University has the best university [work environment](#) in the country, according to a survey of researchers in the Oct. 20, 2003, issue of *The Scientist*.



# The Environment – Purdue University

The first man to walk on the moon (Neil Armstrong) and the last man to walk on the moon (Eugene Cernan) are Purdue graduates. More astronauts (22) have graduated from Purdue than any other university.

Dr. Arden Bement, distinguished Purdue professor, on leave to serve as director of the National Science Foundation (NSF)



## Process & Options @ Purdue

- Late summer – 2004
- Purdue Libraries approached ITaP (Information Technology at Purdue)
- Immediate interest and response
- Identification of roles and contributions
- Development of ‘White Paper / Road Map’
- Assessment of scale



## Progress

- Two Prong Approach
  - ITaP – negotiating equipment, space, infrastructure
  - Libraries – promoting, consulting, spreading the word, continue to develop digital document repository
- Assessment of Operating Systems
  - DSpace (open source, developed at MIT)
  - Collaboration with third party software and hardware vendors



## Progress

Slow, very, very slow

- Competing demands on staff time
- Defining concept within Libraries
- Buy in by labs, departments and centers
- Funding



# Progress

By End of Summer 2005:

- Purchase of hardware
- Final testing of software
- Agreement (if required) with third party vendors
- Promotional Plan in Place



# Progress

Stay Tuned for 2006  
Update in Portugal!!

James L. Mullins, Ph.D.  
Dean of Libraries  
Purdue University  
May 30, 2005





