

# University role in caring for research data

---

Craig A. Stewart

Associate Dean, Research Technologies

Chief Operating Officer, Pervasive Technology Labs

stewart@iu.edu

*February 27, 2007*



**INDIANA UNIVERSITY**



INDIANA UNIVERSITY

# Who can be trusted to preserve data over the long haul?

Catholic Church ~2,000 years (“How the Irish saved civilization”)

- Universities (Constantinople, founded in 849; Al Karaouine in Fez, Morocco, 859)
- Publishers?
- Federal agencies?
- Google über alles? “Google's mission is to organize the world's information and make it universally accessible and useful” -<http://www.google.com/corporate/index.html>
- Google opened its doors in 1998
- Google's patents expire in 2018



# The essential roles of the University

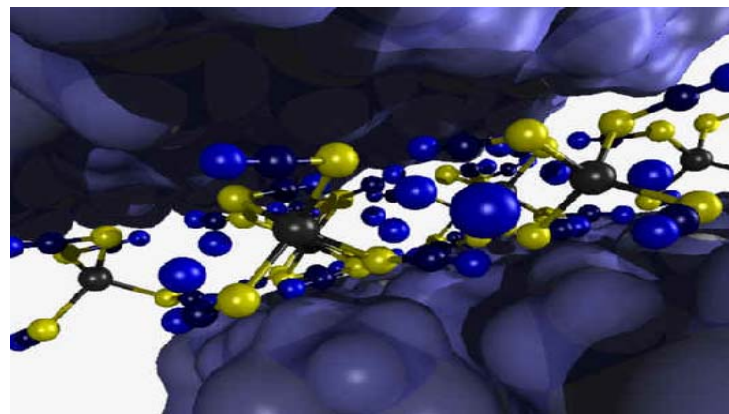
- Provide education
- Provide a forum for open debate and formal mechanisms for peer review
- Serve as a repository of knowledge and information
- The sometimes troublesome but unique multiplicity of roles played by the modern university causes universities to be uniquely suited to be a long term repository of data and information: part public good, part private industry, part structured home for long term debates, part competitor for funds, and part hidebound victim (or beneficiary) of tradition.
- Leading Universities can reasonably be trusted to be (ir)rational over the long haul



INDIANA UNIVERSITY

Data collected now will be of value for the foreseeable future and then some

- Collaborative Initiative for Fetal Alcohol Spectrum Disorder
- IUMSCC
- Insect Genomics





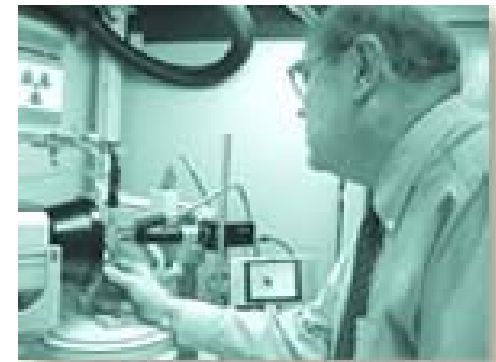
INDIANA UNIVERSITY



Data in general are not really replicable;  
life science data in particular are generally  
not replicable



Works





# Legal/Fiduciary Responsibilities

- HIPAA “...(A) to ensure the integrity and confidentiality of the information; (B) to protect against any reasonably anticipated-- (i) threats or hazards to the security or integrity of the information; and (ii) unauthorized uses or disclosures of the information...”
- 21 CFR Part 11 paraphrased: data collected in drug development may be updated, but never deleted
- NIH guidelines: “...starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking \$500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible....”  
<http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>
- NSF guidelines: “...NSF advocates and encourages open scientific communication. NSF expects ...PIs to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.”

NSF [http://www.nsf.gov/pubs/gpg/nsf04\\_23/](http://www.nsf.gov/pubs/gpg/nsf04_23/)



# Distributed Archival storage infrastructure

- HPSS (High Performance Software System)
- First HPSS installation with distributed movers; STK 9310 Silos in Bloomington and Indianapolis
- Automatic dual commitment to tape in Indianapolis and Bloomington, via I-light.
- Current capabilities: ~ 2 PB data





INDIANA UNIVERSITY

# IU Strategy: self archiving information

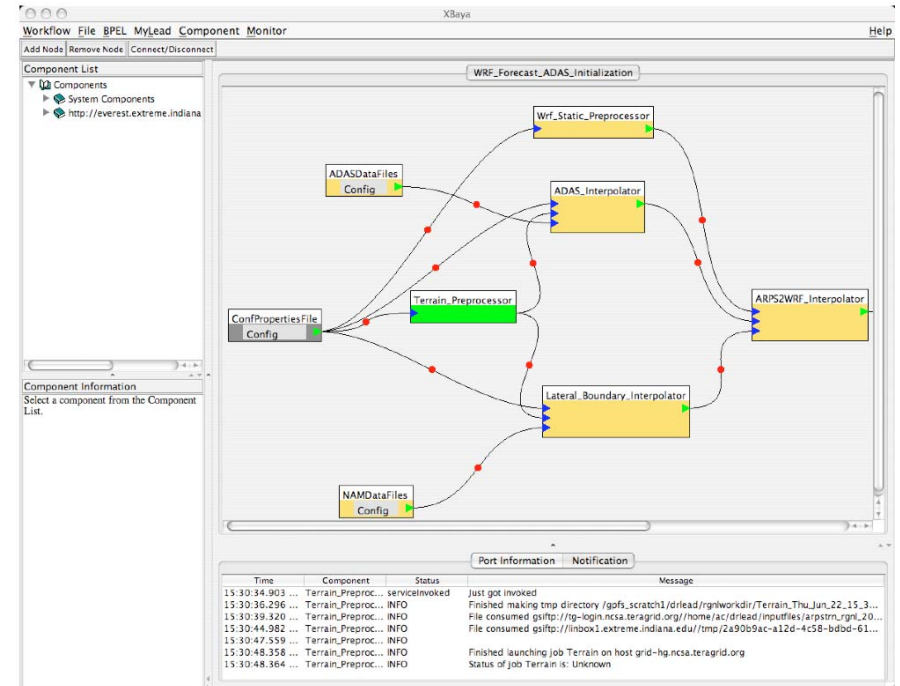
- Long term commitment
- Metadata matters:
  - Discoverability
  - Provenance management
  - Usability
  - Compliance with best available  
metadata standards
  - XML...



INDIANA UNIVERSITY

# Workflow: Composing Computational Tools to build new Tools

- A workflow is then a composition of services in which data moves from one end of an abstract process “pipeline” to another.
- Users “drop and drag” component applications onto a pallet and wire them together
- The graph is compiled into an execution script and loaded into the workflow engine.
- If an execution of a workflow is an experiment, how do you make an experiment replicable?





# Unique partnership

- Libraries
- University Information Technology Services
- Data and Search Institute
- Scholarworks
- Self-archiving data repositories
- Out of region tape copies

The screenshot shows the IU ScholarWorks website. At the top is the Indiana University logo and name. Below that is the title "IU Scholar WORKS". The main content area is divided into several sections: "Browse" with links for Communities & Collections, Titles, Authors, and By Date; "Log On To:" with links for receiving email updates, submitting and managing items, and changing passwords; a search box for "Search IUScholarWorks"; and a "Communities in IUScholarWorks" section listing various departments and programs with their respective item counts. On the right side, there are sections for "About IUScholarWorks", "Join IUScholarWorks and deposit your papers" with several links, "Contact Us", and "Related Links". The date "26 April 2007" is displayed in the bottom right corner of the screenshot.



# National and international entities have a role, but...

- If you want something done...
- Accessibility and usability of data is a competitive advantage for the University and the State
- Why bother producing scholarly works if we are not going to take time to ensure their availability?
- Universities are one of few entities in existence that can be counted on to understand the long term value of data preservation, discoverability, and usability, so Universities have a significant role



INDIANA UNIVERSITY





- Organized garbage is still garbage. Mining garbage might get you a lost diamond, but is more likely to get you garbage.
- Donald Michie, code breaking colleague of Alan Turing at Bletchley Park during World War II and one of the pioneers in research in Artificial Intelligence: "computers cannot create information. This has long been clear. What has only recently become clear is that there is no such barrier to computers creating knowledge." (The Knowledge Machine, with Rory Johnston, William Morrow and Company, 1985, p. 133).
- Advanced software and information technology systems, combined with active management by humans are the only way in which we can hope to maintain data and knowledge and their availability and utility



# Acknowledgments

- This material is based in part upon work related to the TeraGrid supported by the National Science Foundation under Grant No. 0338618, 0504075, and 0451237. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).
- The CIFASD Informatics Core was supported by grant number 1U24AA014818-01 from NIAAA/NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIAAA/NIH.
- This research was supported in part by the Indiana Genomics Initiative and Indiana METACyt Initiative. The Indiana METACyt Initiative of Indiana University is supported in part by Lilly Endowment, Inc.
- This work was supported in part by Shared University Research grants from IBM Inc. to Indiana University.
- The work of the Data and Search Institute is supported by EMC<sup>2</sup>, Attenix, and MUSE
- Thanks most of all to John N. Huffman!!!