Good afternoon ...

It is a pleasure to be here representing Purdue and its Office for Research and Partnerships

and I am glad for the opportunity to encourage you in your effort to pursue the goals of the symposium

in a moment I’ll share with you some thoughts about the growing importance of Research Data Management and collaboration in that endeavor

First let me provide an introduction to Purdue and its sponsored program; activities to give context to the institutional aspects of what follows

Purdue is a state-assisted land grant univ and with a natural focus on STEM research,
40K students, 30K ugrad and 10K grad & prof
10 colleges, no SOM, but Colls of Vet Med & Pharmacy
16K staff, 3K with faculty appointments, 2/3 tenure-associated
Approx. 50% of faculty in Coll. Ag, Eng, and Sci; 75% in STEM
3500 proposals a year
This year expect $400M in awards and $600M in expenditures
Based on expenditures, NSF is typically the top fed sponsor, followed by HHS, DOD, AG, DOE – each ~$30M or more

Research administration bifurcated between sponsored program services under the treasurer and office for research and partnerships.
Let’s turn now to one statement of the goals of this conference

With participation of a variety of stakeholders, the goal is to chart a basic course for universities to follow in lifecycle management of data.

Two key premises are embedded in that statement, and I think they are easy to accept at face value.

First, many different stakeholders have interests: students, faculty, archivists, information and data specialists in multiple domains, research offices, compliance offices and so on.

Second, university-wide solutions are possible and preferred targets.

These two premises imply collaborative broad solutions, and I hope I return to that theme often enough so that it sticks.

I am a stakeholder on at least two fronts, as a faculty member and as a research administrator, and I would like to discuss the situation before us drawing on experience on from both roles.

But I am not an expert in data management. So, there is a very good chance I will say little that you haven’t heard before...that is an inherent risk for talks of this type. Fortunately, wise organizers of symposia of constrain the risk by limiting time.
Let me first reflect on the matter as a scientist from a field of molecular biology now called structural biology, which has drawn scientists & techniques from physics, chemistry, biology, and computing from its beginnings.

Early in my experience, as common is other fields, the outcomes of research projects were documents – dissertations, papers in archival journals – small batch distillations of the mash of data.

The data were stored in notebooks, on punched paper tape or computer cards, 2400 ft reels of magnetic tape, etc. Stored to allow revisitation, but not in a way that encouraged it.

Nevertheless, value of carefully validated collections of data was recognized within individual disciplines – printed Handbooks of Chemistry and Physics come to mind – had copies at home and/or in lab for at least 40 years.

As our field expanded and gained prominence, we, or most of us, embraced the need to disseminate data through international databases, such as the Protein Data Bank and databases of protein and nucleic acid sequences.

**Open access to the data** then became a requirement for publication and funding, a new and important component of the responsible conduct of research, and a primary mechanism to curate and correct the scientific record.

We were lucky because we were early to build databases and because we were in a prominent biomedical field -- the public has paid for our key data management services for a long time.

We were also lucky because the services began and evolved with strong and consistent input from the community. Thus they serve us well.
My other stakeholder role as an associate VP for research leads to different concerns based on different specific responsibilities.

First, my colleagues and I facilitate the research and scholarship of the faculty. We promote their success in funding, access to research tools that support their efforts, and help them achieve recognition of their contributions.

Second, we manage relationships with the sponsors of the research and work to assure compliance with the terms and conditions of sponsorship.

In both domains, the provision of robust and effective data preservation, discovery, and retrieval services is a critical matter experiencing pressure on at least two major fronts.

First, as you well know, the sponsors of university research – the public, congress, the federal agencies – increasingly recognize the value of the original data and insist on its availability – they recognize the value of the mash before it is selectively distilled.

Now, the project life cycle cannot end with a shelved notebook or files on a personal computer. The investigator and the institution have new mandates to preserve the original data and make it accessible.

Second, as you well know, the investigator and the university must fulfill those mandates under the stress of technological advances that can overwhelm us with data – as commonly stated we are dealing with data streams that are ever-increasing in volume, velocity, and variety.

To serve our investigators and our sponsors, we desperately need your best efforts and solutions to data management problems.
I admitted earlier that I am not a data management expert, yet I still wish to encourage solutions with particular characteristics.

At a single institution similar to Purdue, I think the data management operation should be developed and sustained in a manner similar to other central or core university research facilities.

Think of instrumental or animal care facilities and the like.

As in other core facilities, the data management service should operate as a collaboration between users, subject and technical experts, and those with oversight responsibility.

A tightly controlled top-down deployment that places high demands on the users and is poorly attentive to their immediate and changing needs will not persist -- the users will turn away and find another solution.

Moreover, an effective collaboration will produce better research.

Just as it is foolish to design an implantable treatment device without consulting a surgeon, it is foolish to plan a data-rich project without consideration of the downstream data issues, whether it is analysis of the data or its preservation, curation, and dissemination. (Examples)

Pulling data scientists and information experts from the libraries, for example, into the planning and upstream elements of the project will yield better outcomes and help avoid waste or loss of value.
This core facility model also implies that we should avoid unnecessary duplication and fragmentation.

I don’t think we can afford “data management” at the level of the college, department, or lab. Isolated, duplicated services are costly to sustain – Just as we can’t afford multiple, competing proteomic facilities, I don’t think we can’t afford multiple un-federated data management operations.

We must not follow the problematic examples of other fragmented information services.

For example, the support of e-mail servers and data centers in departments and individual labs has proven unsustainable at Purdue. The costs in energy, space, and personnel are simply too high.

Elimination & consolidation of half our data centers (30 of 62) saves energy, personnel, and hardware costs between $500K and $1M a year, liberates space, and prevents upfront costs devoted to providing special HVAC and power services for new, narrowly-focused data centers.

This core facility model also argues that data management operations should have a sustainable financial model that recovers costs from sponsors when possible, but is backed by or subsidized as necessary.

In the realm of information services cost recovery is often considered a difficult problem, but we need to deal with it.

And we need to deal with a broad spectrum of research that includes both well-funded low-demand projects as well as low- or unfunded high-demand projects.
During the symposium you will hear from others about the collaborative efforts, challenges, and progress in the development of a campus-wide data management service at Purdue.

From the onset, the effort depended on central investments of people and funding from the Libraries, Information Technology, Sponsored Program Services, and the Office of Research and Partnerships. It also depended on the involvement of and feedback from faculty and student testers and users.

It is best you hear that story from the people who have invested the most.

I simply ask that you listen for evidence that collaboration and shared purpose were important, and offer suggestions for improvement.

Finally, I am sure you realize collaboration can’t stop at the boundaries of the campus.

You are the ones to think about what associations of scholars, research libraries, and research universities might do to establish and spread good practices and to share resources in this area.

I thank you for listening, and I hope you make outstanding progress toward the goals of the symposium.