1. **Problem Statement**
The research data management sub-committee is to develop a roadmap documenting the existing data management practices on campus and possibly a set of best practices for data management. Additionally, the committee will work towards a campus wide awareness of current policies and practices for research data management to be included in a report to be presented to the IT OOC in May 2011.

Project team: Melissa Jane Dark, Patricia Davies, Douglas Foster, Edward (Ned) Howell, Markus Lill, David McKinnis

2. **Steps taken - Chronological**
   - First Meeting – October 22, 2010 – Started framing problem as Research Data Management
   - November 3, 2010 – Invited Peter Dunn, Scott Ksander – started framing the awareness campaign idea – identified a list of others to speak with
   - January 18, 2010 – Came up with a list of people to speak with and questions that we need to address
   - February 2010 – We met with OVPR to discuss our plans and get feedback on who we should talk to. What is the real problem that needs to be solved?
   - February 2010 – We met with Academic IT Liaisons to understand initiatives in this space

3. **Current Status**
   - Data standards for different projects vary widely, spanning human data to engineering data to bird songs.
   - Research collaboration and open sharing of research data need to be supported by sound policies, practices and technologies for information assurance (confidentiality, integrity, and availability of these data).
• Assurance and security of research data (as a part of the institution’s data management plans) is important to: faculty, graduate students, Purdue, sponsors, the public; though the reasons it is important may differ in scope and level across these stakeholders.

• Information assurance and security is a hard problem. Take for example border security where the purpose is to monitor and enforce the traffic across a border established at some point in time to demarcate one territory, state, country from another. Now imagine trying to do this monitoring and enforcement where the borders are not visible, changeable, and without clear jurisdiction. It is a little bit like border security for the air.

• Computing is distributed. We have faculty and graduate students conducting research and collecting research data with a myriad of assurance practices with regard to these data. It is not clear to what extent the faculty and administration of Purdue University understand the assurance risk as well as mitigation strategies. It would seem that in order to solve a problem it would help to define the problem.

The challenge is that there is not a single or just a few problems here. This space consists of multiple problems, some of them technical challenges and some of them human and organizational challenges. Furthermore, the nature of information assurance is such that the problems are emergent. The problem itself is ill-defined, in the sense the there is insufficient information available to enable [us] to arrive at solutions by transforming, optimizing, or superimposing the given information. Presumably, if we had such information, we would use it. But we cannot. And it is not so much that we can simply go get that information somewhere by sheer effort, either. The problem is ill-defined because we cannot obtain enough information, even though we are under obligation to proceed with decisions regardless. The information simply does not exist. Some of it may never exist.

These possibilities do not excuse us from searching for some kind of resolution. Russ Ackoff would argue that what we face is not a problem at all. It is what he would call a “mess.” Understanding the difference is instructive. Ackoff (1981) defined a mess as a “set of two or more interdependent problems (p. 52).” “A mess,” he went on to explain, “like any system, has properties that none of its parts have…. The solution to a mess depends on how the solutions to the parts interact [emphasis provided].” (p. 52). It is this interaction that makes this problem emergent. Ackoff goes on to note that “messes must be [conceived] and understood holistically” (p. 246). We face not one problem, but many.
Based on our investigation, it appears that many of the subproblems are being framed/isolated in an attempt to craft workable solutions. There are efforts in AgIT, Science, ITAP, ECN, etc. There are other working groups that are involved to some degree in this problem space, e.g., the 43 security officers within each of the schools, the Envision Center (we understand there is a project there on export control compliance on data), the group in the OVPR and Libraries working on data management plans to meet NSF requirements, and others we had surely missed.

The main challenge as we see it is Purdue’s ability to understand the systemic and emergent nature of research data information assurance holistically. While the common good is a term that can refer to several different concepts, in the popular meaning, the common good describes a specific "good" that is shared and beneficial for all (or most) members of a given community. That is, it benefits the community and by benefiting the community, it benefits each member of the community.

The tragedy of the commons is a dilemma arising from the situation in which multiple individuals, acting independently and rationally consulting their own self-interest, will ultimately deplete a shared limited resource even when it is clear that it is not in anyone’s long-term interest for this to happen. We see this problem as not only a complex, emergent problem, but as a common good problem. We take this definition to be applicable to the spirit of IT-OOCs and the notion of shared governance.

4. Open Issues

A. Complexity of the task of research data management:
The challenges are myriad, and include IT systems security, lack of human protocols, physical security, emerging regulations, and many other factors. There are pockets of excellence at Purdue but much vulnerability.

B. Grasp of the issue by the three key constituencies:
- **Research administration**: The committee’s (non-scientific) review revealed a mixed understanding of the challenges and vulnerability.
- **Faculty**: Again, a non-scientific sampling indicted significant lack of attention to the research data management, although there are some pockets of excellence.
- **IT professionals**: For the most part, this group has a comprehensive grasp of the IT security issues and is greatly concerned with the
Current state.

C. Need for a university-wide comprehensive review and the development of systems/infrastructure/procedures that support “Discovery to Delivery” as well as compliance with regulatory and ethical requirements.

5. Recommendations

Although an effective, all-encompassing solution to the general problem of research data management and security is not possible due to the status and issues described above, an effort to address immediate requirements and provide functionality and capabilities for the Purdue research community is underway and provides one path going forward. This effort is the Purdue University Research Repository that is under design and development by a working group (PURR WG) consisting of faculty and staff associated with the Libraries, ITaP, and the OVPR. The PURR WG was formed in response to the NSF requirements for research data management plans for all new proposals submitted on or after January 18, 2011. A Steering Committee (McCartney, Mullins, Buckius) is overseeing the efforts of the PURR WG. The PURR WG plans to have the first version of the PURR ready when the initial set of proposals that fall under the new NSF requirements are awarded as early as August 2011.

The following is taken from the PURR Policy and Service Outline, dated April 8, 2011, and summarizes the PURR requirements, design and capabilities:

“Long term data management will occur using central University resources. Purdue is deploying a Trusted Digital Repository for research data utilizing the HUBzero platform, a web-mediated software and hardware platform designed for scientific collaboration and sharing of scientific data and developed under NSF support. The Repository provides tools for ingestion, identification, dissemination, and use of data as well as services to ensure data security, fidelity, backup, and mirroring. Purdue Libraries will provide stewardship and curation of data in the Repository and will supply consultations and tools to facilitate ingestion of data and appropriate metadata conforming to community standards. Industry-standard protocols such as OAI-PMH will be used to expose resources in a variety of contexts, including search engines such as Google Scholar. Data will be accessible for project-appropriate periods and
archived thereafter. Purdue will also assign Digital Object Identifiers (DOIs) and expose documentation about the data sets in the repository.”

The intention is to provide support for data management throughout the lifecycle of a project from proposal/pre-award through conduct of the project through archival activities after project completion. Support includes not only the software/hardware platform but tools and consulting support to help researchers define project-specific data management needs. These include a data management plan self-assessment questionnaire to identify requirements and template data management plans to serve a robust starting point for specific proposals. Implementing PURR for the initial NSF projects awarded under the data management plan requirements will identify needed modifications, enhancements, and processes that must be incorporated to provide solutions to the more general issue of research data management. This is expected to be an iterative and on-going process. Further, the intention is to provide PURR as one option that researchers may employ to satisfy data management needs. Other options include (1) exploiting data management resources already in place for some disciplines [Ex.: data management resources of the Inter-University Consortium for Political and Social Research (ICPSR)] and (2) solutions based on resources that faculty and academic units provide and control themselves. The goal, however, is for PURR to be successful and to emerge as a preferred choice for research data management.

The recommendation of the IT-OOC subcommittee is to follow and fully support the PURR effort. If successful, it will provide significant capabilities in research data management – a first step along the path. These capabilities will form the basis for expansion and enhancement beyond the initial NSF project utilizations to address data management requirements from other funding agencies, for a broader set of data types and volumes, and to resolve emerging security, privacy and regulatory issues. We also recommend convening an informational meeting among the various groups on campus that are attempting to address the data management issues. The idea would be to describe current efforts (requirements, designs/plans, status) and to search for common approaches, identify best practices, surface redundant efforts, and broaden understanding of the issues among all who are working on the problem.