**Title**: Symposium on Integrated Approaches to Integrated Data Publication

**Type of project**: symposium

**Draft task statement**:

Research data is collected in the hope of generating new discoveries, and these discoveries are communicated to the public through scientific publications. While this has been true for centuries, there is a now an increasing recognition that the integration of data and publications can improve the replicability of scientific findings; make publications easier to understand, reuse and evaluate; enhance teaching and education; and support new forms of scholarship.

Stakeholders have approached data publications from different perspectives and with different purposes and requirements. Many of these efforts have been conducted separately, with little integration across systems, and with limited awareness of related efforts. For example:

- Dewald and Thursby’s 1986 study of the replicability of articles published in *The Journal of Money, Credit and Banking* sparked continuing concern about the replicability of research in economics and neighboring social scientists. Within some social science disciplines this has evolved into a practice of archiving “replication” datasets. This line of practice and development continues in systems like the Dataverse Network, Zelig, and ICPSR’s Publication Related archive.

- A decade later, researchers in the image and signal processing, inspired by the examples of Buckheit, Donoho, and Claerbout, developed local practices and tools to create self-contained reproducible electronic documents. In contrast to the replication data approach, the “reproducible electronic development” pioneers developed systems that replicate exact analytic results, but that are less well-adapted to supporting data reuse, documenting data semantics, and providing systematic long-term access.

- Contemporaneously, a number of of statisticians (many of whom were part of the core team developing the R statistical programming language), inspired by Don Knuth’s idea of “literate programming”, lead efforts to develop “literate statistical documents”. These constituted an independent approach to combining text, data, and the code for statistical analysis to generate a final published document. Light-weight statistical document systems such as S/Weave (developed in early 2000) and StatWeave have an active community of users among statisticians, and in many statistically-influenced disciplinary sub-fields (such as psychometrics, and chemometrics).

- Five years later, the increasing use of scientific workflow management systems, which had been developed to marshal computational resources for the analysis of data that had dramatically increased in volume and scale, sparked a cluster of computer science research into models and tools for workflow and data provenance -- which emphasized granular information tracking and change management.

- Recently, scholarly publishers have begun to experiment with enhanced publications that involve embedding or linking data within publications to support interactive analysis. These systems emphasize dynamic interactivity over open platform and long-term access.

The result has been a variety of approaches to enhancing reproducibility and richness of empirical scientific publications that have yielded local successes, but are based on isolated and incompatible approaches. This symposium will bring together stakeholders and researchers from the areas of publishing, government, computer science and curation to identify exemplar projects in different areas; share knowledge and finding. And by making explicit the implicit requirements, design criteria, user communities, and areas of overlap, the symposium will catalyze collaboration across data-publishing approaches.

1. Convene a symposium with leaders of exemplar data publication efforts and approaches by publishers; government; data curators; information scientists; and domain scientists.

2. Facilitate discussions of areas of overlap across independent efforts; emerging good practices; and stakeholder goals and requirements for data publishing systems.

3. Produce a report summarizing exemplar efforts, common themes, areas of overlap, and potential areas for collaboration among stakeholders.

**Significance, and role of BRDI**:

Data publication is rapidly evolving -- and coordinating activities at this stage could have a substantial influence on future practices, norms, and infrastructure. Integrated data publication has the potential to change the dissemination of government information; enhance scholarly communications; and change how scientific information is communicated to the public. The strong connections of the board with stakeholders in the areas of research, curation, publication, along with its reputation for neutrality and scientific excellence make it an ideal convener for a symposium across stakeholders.

**Potential sponsors & specific audience**:

The symposium will identify innovative efforts in data publication and areas of potential collaboration. The results of this symposium will be relevant to leaders in scholarly publishing; researchers in scholarly communications and information science; education researchers; researchers in quantitative scientific fields; developers of data management infrastructure; and funders in related area.