The research data ecosystem:
Stakeholders with potential to ensure shared data could be reused in the long term

General data repositories: Developed as data publishing services and historically have not shared the curatorial mission of domain specific data archives who are involved in data preparation and review prior to publicly sharing data. They do provide secure storage, persistent identifiers, and they support varying degrees of file inspection. May offer some metadata creation tools, and useful guidelines but typically put responsibility for “data quality review” on researchers. These platforms have enormous potential to change how data are prepared for publication and sharing, but they have not yet taken on that role.

Academic libraries: Some libraries have a history of including data files in their collection policies, and they support tools for data reuse and analysis, but most have only partnered with individual local researchers to provide guidance on data acquisitions or research data management. Institutional repositories are making progress in taking on the role of stewardship of data outputs by their affiliated researchers but typically do not review data beyond basic bibliographic-level information, and they rely upon data being properly prepared for sharing prior to submission. One solution is for IRs to partner with data archives. Libraries could also do more to promote author identification systems such as ORCID and data citation (note recent developments in data citation).

Scholarly journals: Currently, legacy journals that require data demonstrate uneven oversight of the data in terms of how they manage it, where they put it, how they enforce compliance, and what type of review they conduct. Many offer guidelines and checklists for researchers and hold researchers responsible for what they deposit. Avenues for implementing “data quality review” might be stricter enforcement of policies (stick) and replication audits in which a random sample of articles are subjected to closer scrutiny and get recognized if they pass inspection, see Dafoe (carrot). New ways of publishing data, such as data journals, data descriptors, and data papers (e.g., Scientific Data, Journal Open Psychology Data, PLOS) may hold promise for incorporating review activities.

Researchers: Researchers are best positioned to implement “data quality review” practices, and ideally they would do so as part of the research workflow. The problem is that they currently view this as extra burden. The curation community has taken a huge step forward by making it easier for researchers to share data, but we also know that it comes with the cost of poorly documented and sloppy datasets and code. Guidelines for researchers can help, but they are often not enforced. Data management plans, required by many funders, ask researchers to describe how they will store and share their data but not how they will ensure the data are usable and understandable. The peer review mechanism is also problematic, as many have noted. Most scientists want to use the data, not review it; if the data are not usable and independently understandable, researchers will be able to do neither. We see potential in collaborative environments, such as GitHub and Open Science Framework (OSF), for incorporating data review practices into the research workflow, but as of yet they are not focused on such curatorial tasks.

Based on “Committing to Data Quality Review,” Peer, Green & Stephenson, 2014