What Every Early Career Researcher Should Know about Research Data Management

Alex de Sherbinin, Elaine Faustman, & Rorie Edmunds

Introduction

The Scientific Committee of the ICSU World Data System (WDS-SC) believes that all Early Career Researchers (ECRs) require a basic set of data-related skills. The following presents essential areas of Research Data Management (RDM) that are relevant to budding scientists and that cover the range of issues they are likely to encounter as they collect, analyze, and manage data over the course of their careers. It has been formulated with the assumption that ECRs play an important role for future data sharing, and must take an interest in data stewardship and best practices in data management, including how to make data openly accessible and reusable.

The Essentials of RDM

Open Data. Almost all science funding agencies require that research results, including data, be made publicly available. Journals, too, are requesting that authors of scientific articles post their data, and even the code used to generate results. Data sharing and open data are important to the advancement of science, and data reuse has resulted in important scientific discoveries. ECRs need to be familiar with the FAIR principles—that data need to be Findable, Accessible, Interoperable, and Reusable—and work towards data sharing and research transparency in their own work.

Big Data. The term 'Big Data' arose to describe the Volume, Variety, and Velocity (the *three Vs*) of data being generated almost continuously by a range of sciences, from Biomedical to Earth Sciences. An ECR should have an understanding of what is meant by Big Data, and how they are increasingly important to a variety of scientific fields. Familiarity with tools and approaches to analyzing Big Data is also an important requisite for career advancement.

Definitions and Jargon. An ECR must know some of the terminology in the data arena, such as 'ontologies', 'informatics', 'metadata', and 'knowledge networks'. A critical element for data sharing is common definitions, and particular attention should be paid to understanding ontologies, thesauri, and controlled vocabularies: what ontologies are, where to find them, and how to create them, as well as ways for integrating ontologies and using them to support metadata and data disambiguation efforts.

Funder Requirements and Writing Data Management Plans (DMPs). Funders increasingly require that scientists articulate through a DMP how they will ensure the open availability of their data for the long term at the onset of a project. An ECR should know how to thoughtfully prepare a DMP that will also increase the odds of them obtaining funding. Awareness of the domain-specific data repositories where their data may be archived is also important (see below). A conceptually ideal DMP is extensible, interoperable, and machine readable, and an ECR must understand why these aspects are needed and how to address them.

Data Organization and Storage. Organization and long-term preservation of data is an increasingly daunting task. An ECR should know methods of sustainability to ensure the continuance of databases as they begin to generate data. Documenting versioning, choice of technology and standards, and archiving also need to be understood. The principle that data have several end uses throughout their lifecycle—each with its own requirements—is fundamental within this, and the concepts of 'Analysis-ready' and 'Publication-ready' (data with quality assurance, citation, and metadata) data should be familiar to an ECR.

Metadata Formats, Usage and Data Discovery. Metadata are critical for data discovery and reuse, and are the *bread-and-butter* of catalogue services. Metadata standards are strongly format and discipline dependent, but common elements are increasingly captured in efforts by DataCite, DCAT, and others. The International Organization for Standardization (ISO) has also developed a number of domain specific standards, such as ISO-19115 for geospatial information. An ECR should recognize the importance of proper metadata development, and be aware of a number of the standards that are available.

Data Documentation. To be of use to other researchers, data need to be carefully documented: to describe how they were developed, their limitations, and to what use they may be put. Incomplete and cursory documentation often renders data unfit for future use. An ECR should have knowledge of the different approaches taken to data documentation in various fields of science, as well as of the increasingly important practice of properly referencing protocols, methods, and samples.

Data Formats and Interoperability. Data formats and applicable standards for data and metadata are largely dependent on the scientific discipline and the type of software used. There are data formats that are common across disciplines, but this is not the norm. An ECR should support open formats and well-entrenched standardized services (e.g., CSV files, DDI, OGC services, and OPENDaP, to name a few), and having an overview of their scope is a useful starting point for an ECR to make appropriate choices. For a discussion of data standards and interoperability in the health domain, visit AHIMA.

Choosing a Long-term Repository. An ECR must have an understanding of not only which disciplinary repositories are best suited to the domain in which they are working, but also the 'trustworthiness' of these data repositories, and how this is underpinned by a hierarchy of certification standards (e.g., the CoreTrustSeal). By examining the strengths and weaknesses of different repositories in terms of data access, documentation, and so on, it helps an ECR to conceptualize what makes for a successful data service.

Standardization, Licences, and Intellectual Property Rights. To aid in their reuse, data should ideally be made available in standardized schema and using standardized services. Each 'data family' has its own set of such standards, and an ECR should know which are relevant to their discipline. Moreover, with Open Data an increasing norm in the scientific community (see above), an ECR should be aware of the different types of licensing and copyright arrangements under which data are often disseminated, in addition to the importance of machine-readable licensing arrangements.

Data Ethics. While primarily salient for ECRs working in the Health Sciences, Social Sciences, and Humanities, ethical issues that arise throughout the data management lifecycle should be a topic of broad interest to all researchers likely to engage with disclosive data (e.g., research on rare biodiversity, where there may be commercial interests in their exploitation). Areas that an ECR should have knowledge about include, but are not limited to: data ownership and stewardship, handling sensitive data, consent, privacy and confidentiality, reconciling ethical and legal norms impacting data sharing and exportation, constructing equitable partnerships and data sharing agreements, and navigating the complexity of ethics review.

Data Publication, Citation, and Persistent Identifiers. An increasing number of data journals, such as the Nature Group's *Scientific Data*, are now available for the publication of datasets. In addition, proper citation of data using persistent identifiers is becoming the norm in the scientific community. An ECR should be aware of the approaches to data publication and citation and the importance of doing these properly.

Research Translation and Societal Benefits. To facilitate use of data collected and stored within archives, an increasingly wide range of software has been developed for decision analytics and support. In addition, there is a great deal of work on integrating data across disciplines to support new discoveries. An ECR should understand the value that well-curated and sustained data management provides to the scientific community and larger society, and have some understanding of data indicators, decision-analysis techniques; and the graphical interfaces that can simplify exchanges. Linked ontologies and robust metadata can facilitate these possibilities.

Citizen Science and Crowdsourced Scientific Data. Citizen science and crowdsourced data have already proven to be of tremendous scientific value. However, the modest budgets of these initiatives typically mean that systems are lacking for the curation and long-term stewardship of their data. An ECR should know what citizen science is, and how to design an initiative that engages citizens in improving scientific data collection and use: addressing issues of data stewardship, validation, confidentiality, dissemination, and licensing from the beginning. SciStarter provides a good introduction to Citizen Science, and an example of pointers for the design of citizen science can be found at the Cornell Lab of Ornithology.