Long-term ecological research programs represent tremendous investments in human labor and capital. The amount of data generated is staggering and potentially beyond the capacity of most research teams to fully explore. Since the funding of these programs comes predominately from governmental institutions these data should be considered public goods. As researchers involved in these programs we have an ethical, and, increasingly legal, responsibility to ensure the data are properly curated, archived, and ultimately made available to the larger research community. Fortunately, new tools are available to help data managers with this task. Data can now be stored in relational databases that are archived online and processed with open source software such as R, Rstudio, and GitHub. R is a fully functional, open source programming language that can be used for statistics, GIS, and data management. Rstudio, also open source, extends the capabilities of R. Github, an online version control system, is fully integrated into Rstudio and facilitates code sharing and data documentation. In this talk we will demonstrate how these tools are used to effectively manage and analyze one of the worlds largest small mammal capture-mark-recapture databases.

As part of a large scale experimental manipulation in the semiarid zone we initiated a small mammal capture-mark-recapture study in Parque Nacional Bosque Fray Jorge (Chile) in 1989. Our protocol calls for monthly small mammals inventories on a minimum of 16 0.54 ha. grids. During 4-day censuses, small mammals are captured in 50 large Sherman traps, marked (if new), and standard population and condition data are recorded. During each census there are up to 500 traps in operation leading to ~4000 trap-nights of effort/month. When we began the project we had no idea that work would last over 25 years (and counting) leading to more than a half million captures of over 81,000 individuals. Such a dataset poses special challenges for quality control and analysis due to its enormity.

Data management procedures have evolved during the course of the study as the complexity of the data has increased, and new technology became available. Initially capture records were stored in spreadsheets, but eventually we moved to a SAS database. Currently, data are stored in a relational database that allows easy retrieval by statistical programs such as SAS and R. We have developed strategies for data entry, quality control and quality assurance, version control, error handling, documentation, analysis, and data sharing typical of capture-mark-recapture studies. Particular problems include reuse of tag numbers, tag changes, and observer errors. The importance of these problems has increased over the years with the complexity of the database. In this talk we give a historical view of our data management and analysis approaches, provide examples of problems and solutions, discuss lessons learned, and provide insights into how to work with datasets of this size.