EVOS data project - Plots

Contents

Background

This script produces graphics for a publication reporting the results of a two year data archiving effort by a group of researchers and students at the National Center for Ecological Analysis and Synthesis at UC Santa Barbara. The Exxon Valdez Oil Spill Trustee Council (EVOSTC) was formed following the Exon Valdez oil spill in Alaska in 1989. Since then, the EVOSTC has funded hundreds of projects and in 2012 we initiated an effort to recover and archive the data collected through these EVOSTC funded projects. The recovery effort spanned two years.

For this paper we ask 3 main questions about the data collected from the Exxon Valdez Oil Spill Trustee Council funded projects:

1. Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?
2. Are there differences in data reporting based on characteristics of the data?

- Research field
- Sector of researching body
- Year data projects ended

3. Which of these characteristics are most *important* in determining if a dataset will be successfully recovered and how do the important characteristics influence the output (success)?

We were also interested in why data were not recovered and plotted reasons why recovery was unsuccessful.

Status definitions

Data.status	Definition
outreach attempted	Contact information found and outreach attemtped via email and/or phone
contact established	At least one reply was received by the archiving team from the target researcher. Confrims contact information was correct
data received	Data was received from the researcher, regardless of data quality or level of documentation
published	Data were received and docemented well enough to be archived by the data team or the researcher cooperated in data clean up and documentation
unrecoverable	Data were unrecoverable either due to inability to find contact information or for reasons confirmed by the researcher/data owner

Hurdles definitions

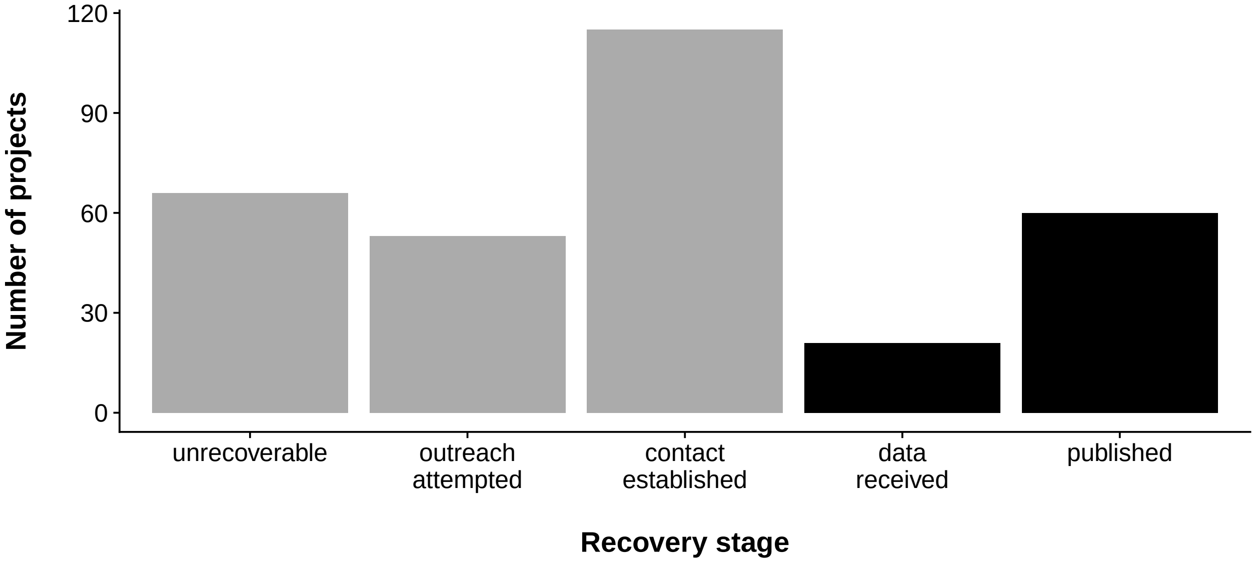
Hurdle	Definition
no contact info	No contact information found or contact information was never confirmed because outreach attempts received no response
communication lost	At least one reply was received by the archiving team from the target researcher but communication was lost before any data were sent
data lost	PI or other data manager confirmed that the data no longer exist
non digital	Data exist in a non-digital format (excludes digital PDFs, includes hand-written or type-written data)
unwilling to share	PI or other data manager refused to share data
requested funding	PI or other data manager agreed to share data only if additional funding was provided

1. Project Status Reporting

> Twenty-five years after the EVOS, for how many projects funded by EVOSTC can we collect data?

Fig 1: Recovery process

Final status of all projects from which data were requested. Black bars are projects for which data were successfully acquired, grey bars represent projects for which no data were acquired.

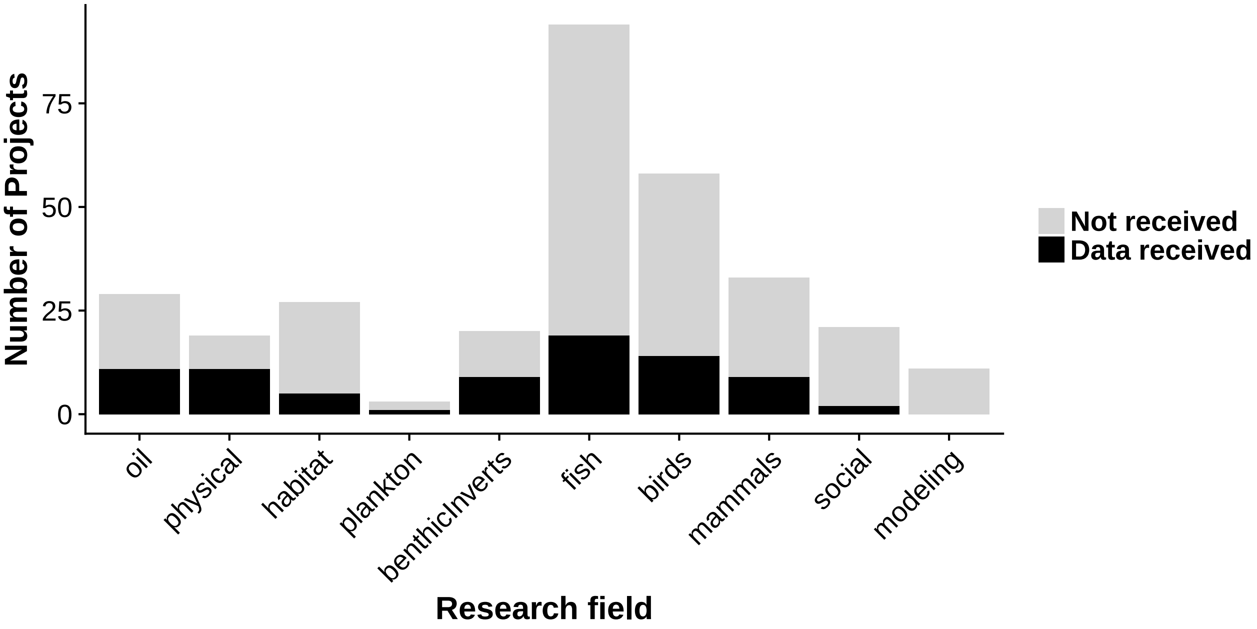


2. Data Characteristics

Are there differences in data reporting based on characteristics of the data?

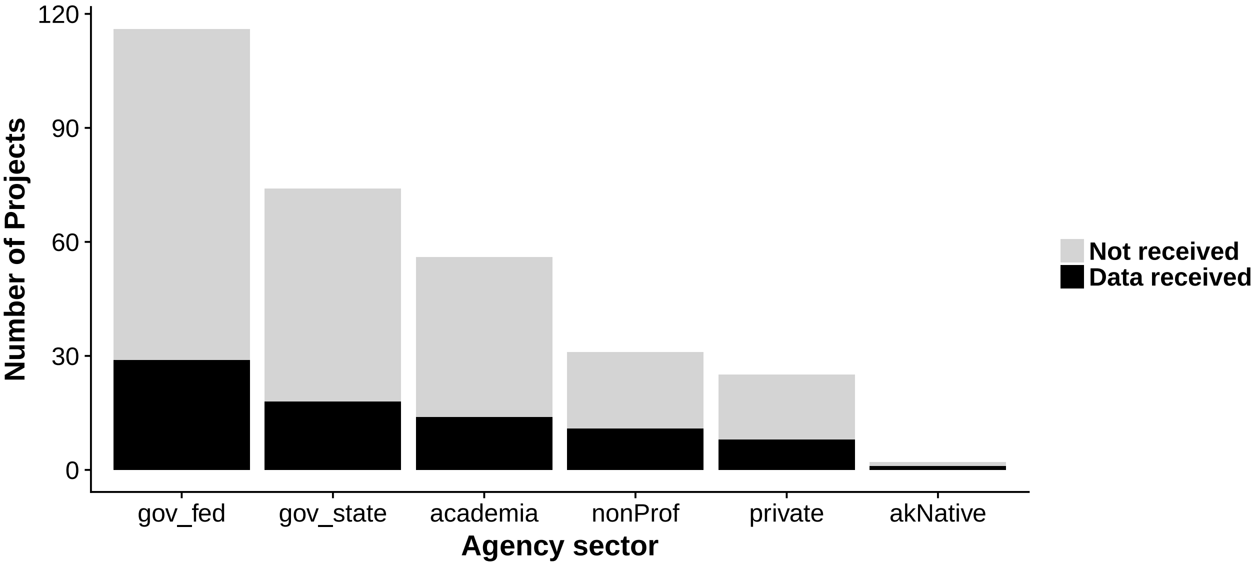
Fig 2: Research Field

Total recovered and non-recovered projects by research field.



S1 Fig: Grantee agency sector

Total recovered and non-recovered projects by agency or institution sector.

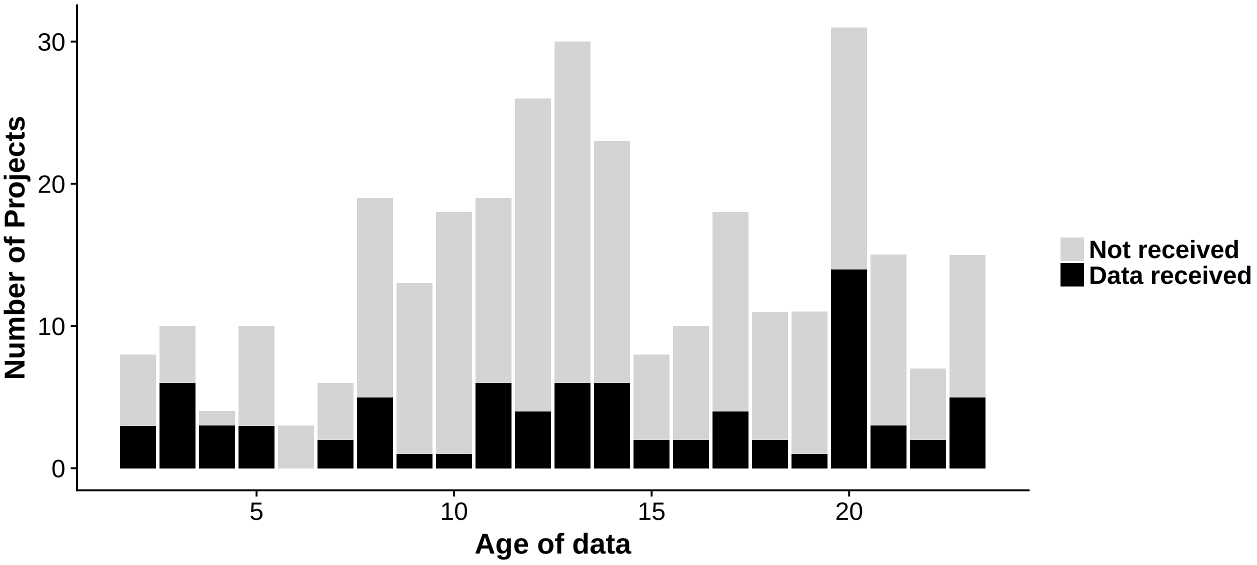


S2 Fig: Age of data

Is the availability of data correlated to how old the data are?

Total recovered and non-recovered projects by project age. Age is calculated based on number of years between the last year of EVOSTC funding and start of the archiving project (2012).

The x-axis might not be initially intuitive but we are representing years since a project ended in opposite-chronological order to show any effect of increasing age of a dataset. We chose this design to be able to compare to Michener et al. 1997 - fig 1.

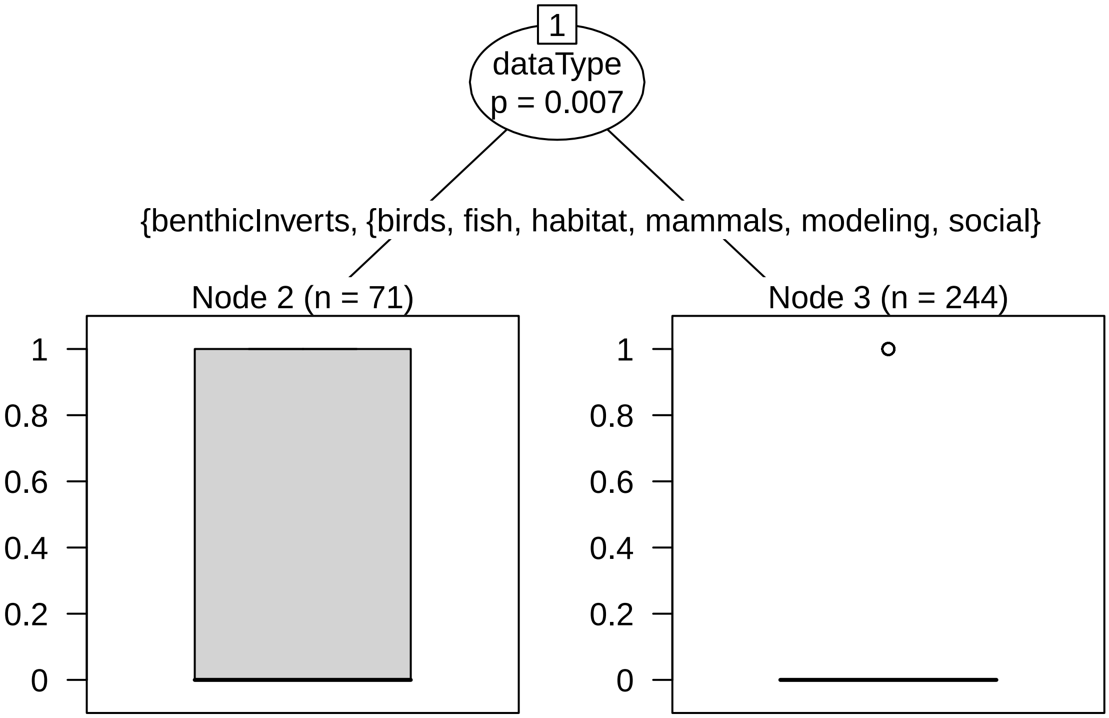


3. Which characteristics are most important?

How do the important characteristics influence the output (success)?

Classification tree

Classification tree of which data variables and characterizations predicting successful and non-successful data recovery. Data field was the only variable that could be used to predict data availability.



4. Reasons for not sharing

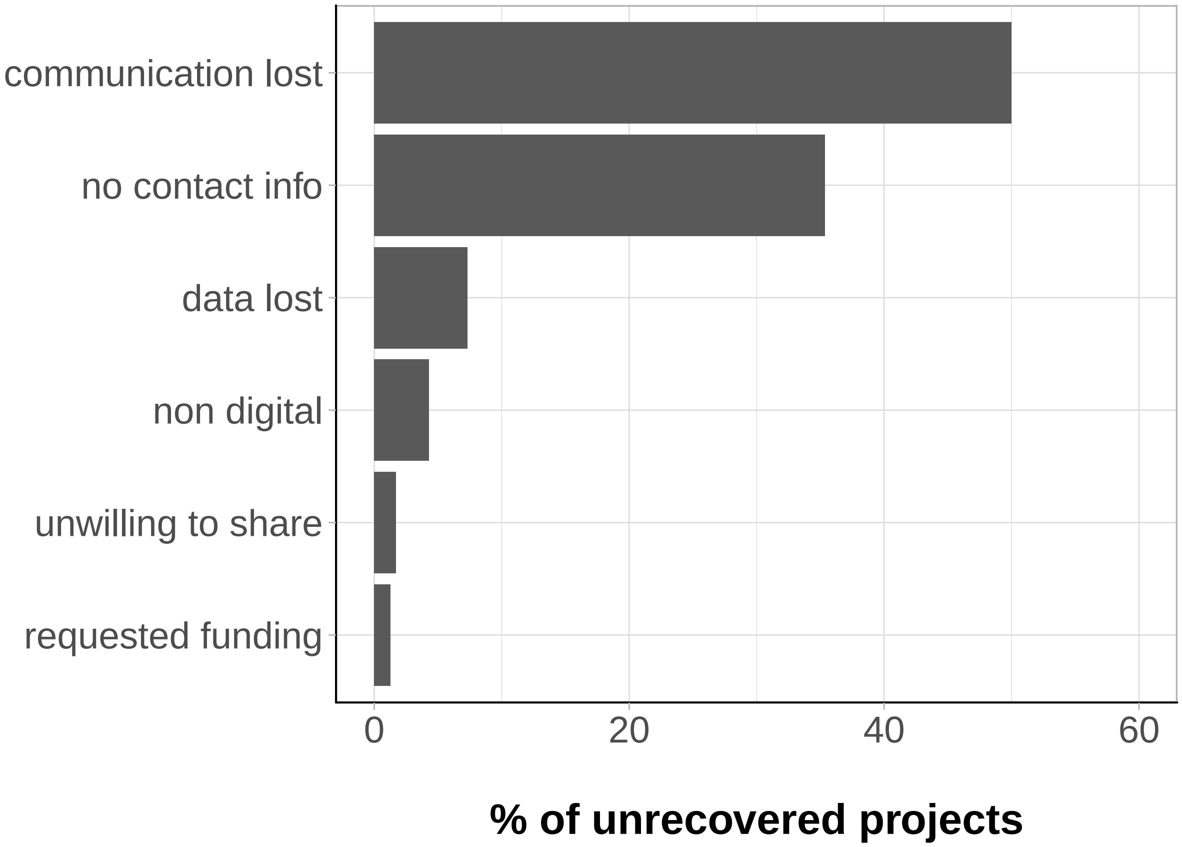
Why don't people share data?

Reasons given for not providing data. Communication loss and lack of contact information were the main reasons data were not obtained.

Data details: Here I used all of the data that were not collected and sorted them into categories based on recovery notes:

- All datasets labeled "emailed" were grouped with a other datasets that were labeled "unrecoverable" because no contact information could be found.
- All datasets labeled "replied" we put into their own category called "communication lost".
- The "data lost" category includes all datasets in which someone confirmed that the data no longer existed either due to damage, non-persistent formats (not including printed/non-digital), etc.
- The other are self-explanatory, but represent confirmed responses as to one of these reasons.
- I tried to separate out the datasets that were not recovered but "should be on the CD" to sort these into their own category, but we acked sufficient notes to confidently add this grouping.

Fig 3: Why not share?


communication lost
no contact info
data lost
non digital
unwilling to share
requested funding
% of unrecovered projects