

EVOSTC data collection paper - Data Visualizations

Jessica Couture

10/31/2017

Contents

Background	1
MS Plots:	1
1. Project Status Reporting	1
2. Research field	2
3. Classification tree	4
4. Reasons for not sharing	4
Supplementary figures	6
S1 Figure. Awardee agency sector	6
S2 Figure. Temporal trends?	8
S3 Figure. Temporal trends in reasons?	10

Background

The data and plots produced here contribute to a publication (“Funder imposed data publication requirements seldom inspire data sharing”) to report results of a two year data archiving effort by a small 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 Exxon Valdez oil spill in Alaska in 1989. Since then, the EVOSTC has funded hundreds of projects and in 2012 we began a project to recover and archive the data collected in these EVOSTC funded projects.

For this paper we ask 5 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 certain **research fields** that are more likely to make data available than others?
3. Are there certain **sectors** that are more likely to make data available than others?
4. Is the availability of data correlated to how old the data are? (temporal relationships?)
5. Why did people refuse to share their data?

Warning: package 'dplyr' was built under R version 3.4.2

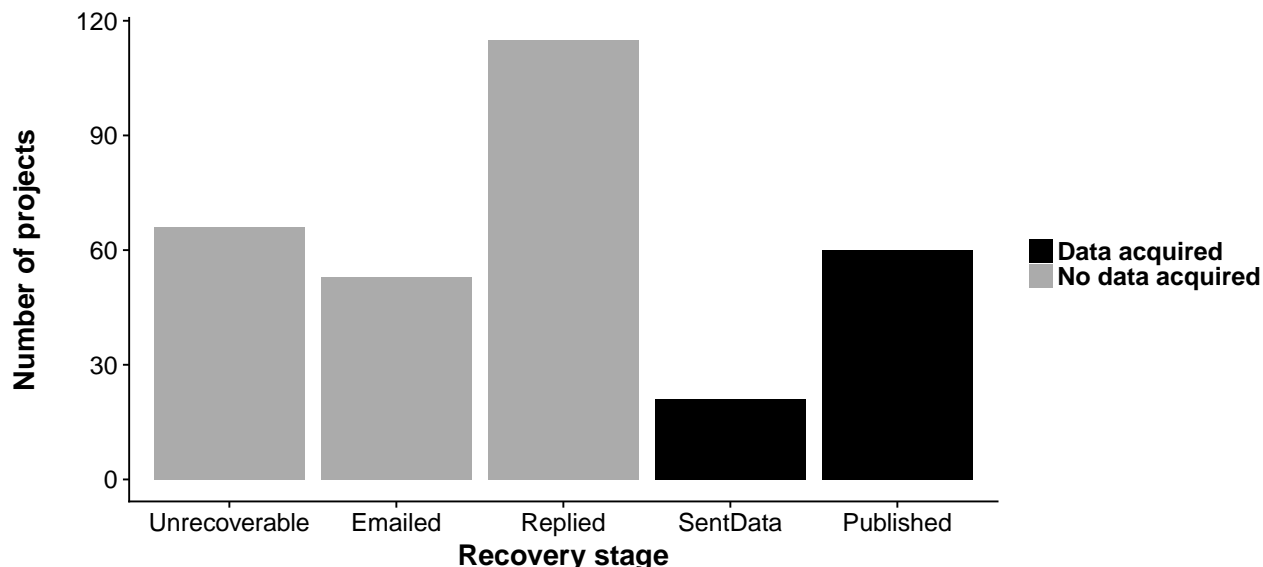
MS Plots:

1. Project Status Reporting

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

Figure 1: Final status of all projects requested. Black are bars are projects for which data were successfully acquired, grey bars represent projects for which no data were acquired.

NOTES: For this first plot, we decided to use the original plot generated from redmine that simply displays the number of datasets in each staus at the end to the 2 year data collection effort. Still having issues getting the arrow in the x-axis title without the period or other character after it.

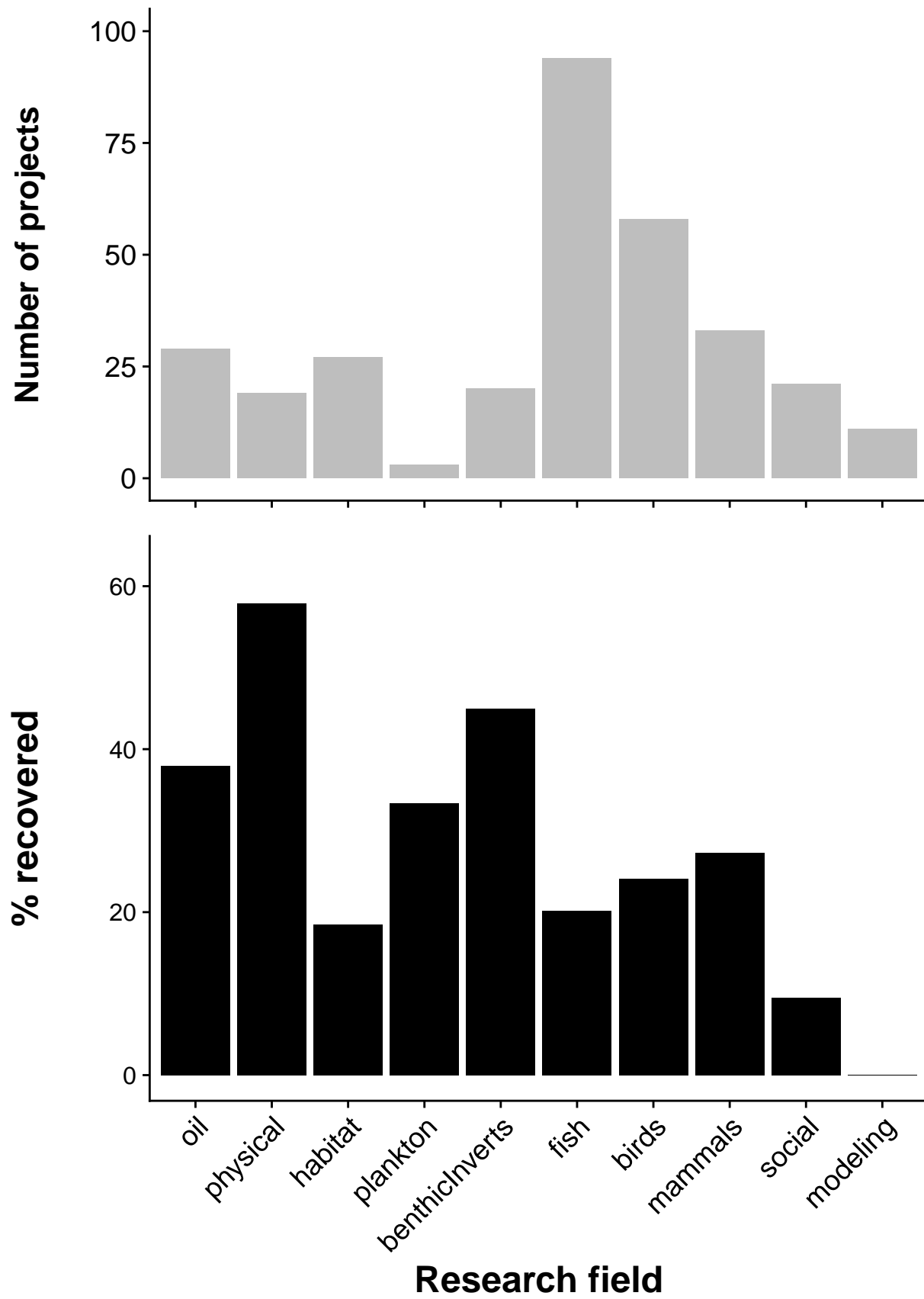


2. Research field

Are there certain **research fields** that are more likely to make data available than others?

Figure 2: Percent success of projects by research field. The top plot is total number of projects funded for each field, the bottom plot shows the percent success for the given field.

NOTES: We stack $nProjects$ /percent graphs to display variation in $nProjects$ between groups as well as variation in relative responses for each group.

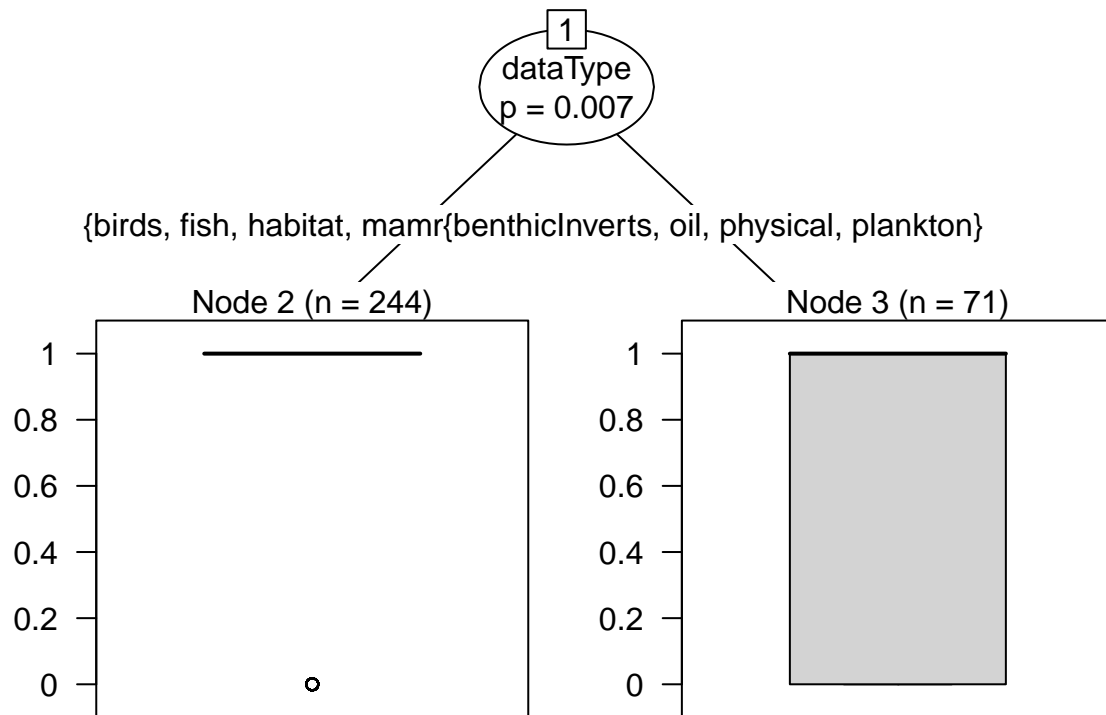


3. Classification tree

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

Figure 3: Classification tree of which data variables and characterizations predict successful and non-successful data acquisition. Data field was the only variable that could be used to predict data availability.

```
rslt2$dataType<-factor(rslt2$dataType)
partreeBio<-ctree(succ~agSubGrp+end+dataType,data=rslt2)
plot(partreeBio)
```



4. Reasons for not sharing

Why don't people share data?

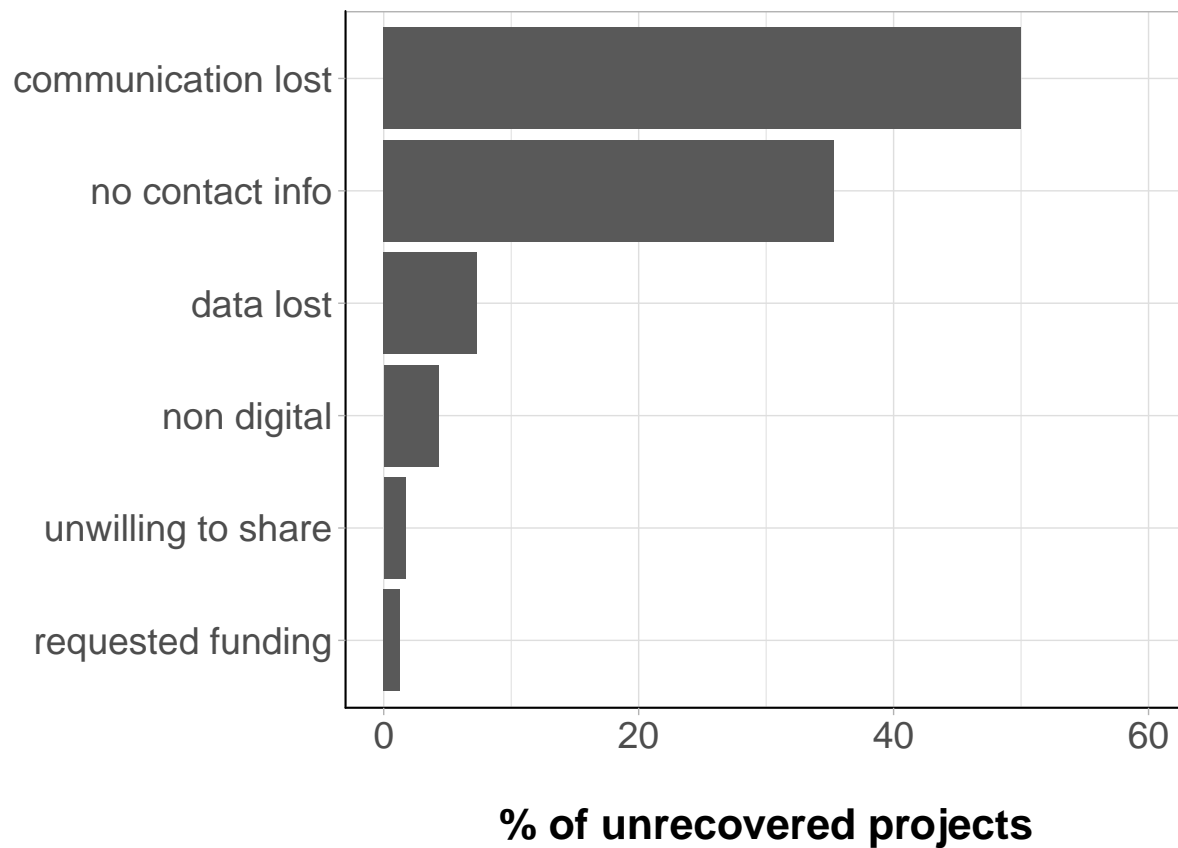
Figure 4: Reasons given for not providing data. Communication loss and lack of contact information were the main reasons data were not obtained.

NOTES: Here we flip the axes to separate this analysis from the others. These data are not a sub-grouping of our results like the others, but rather a deeper analysis of only the unrecovered datasets. Also, unlike the other data above, we did not run stats here.

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

- 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 asked sufficient notes to confidently add this grouping.



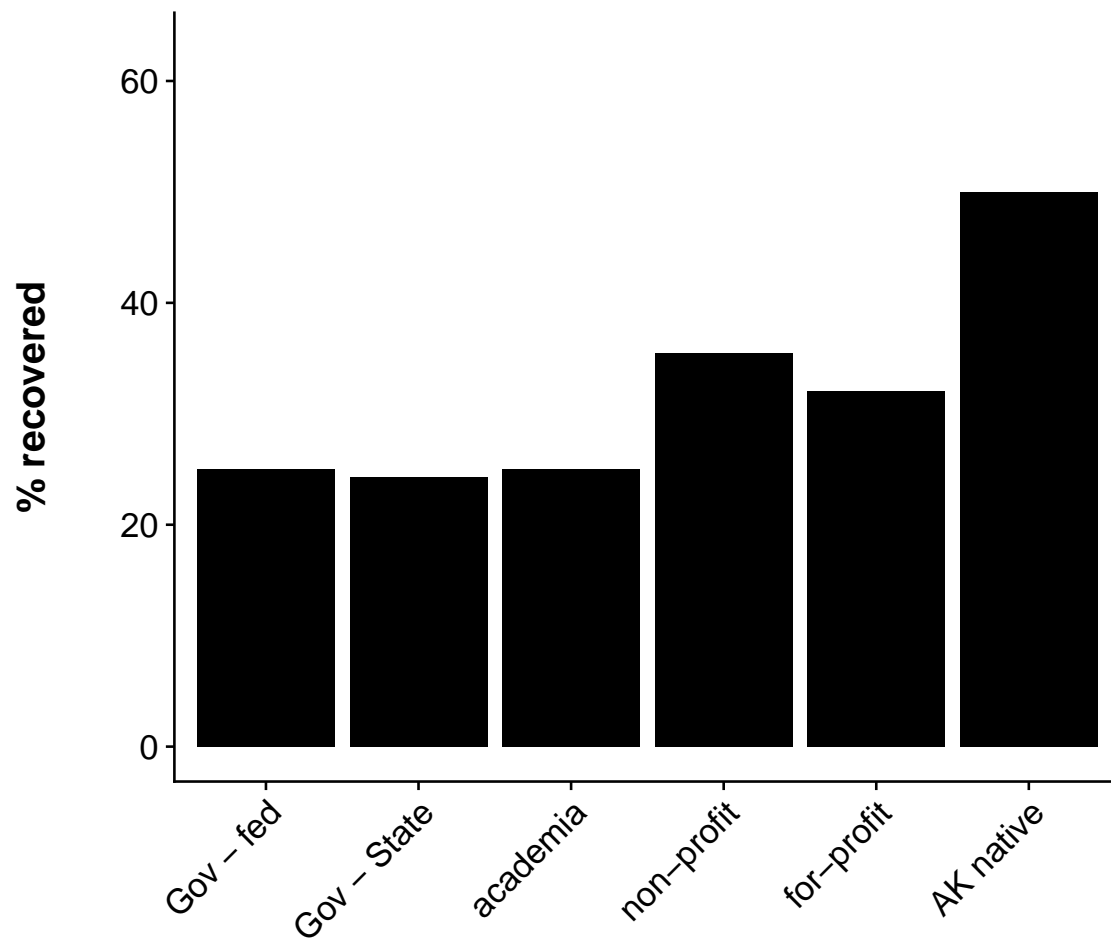
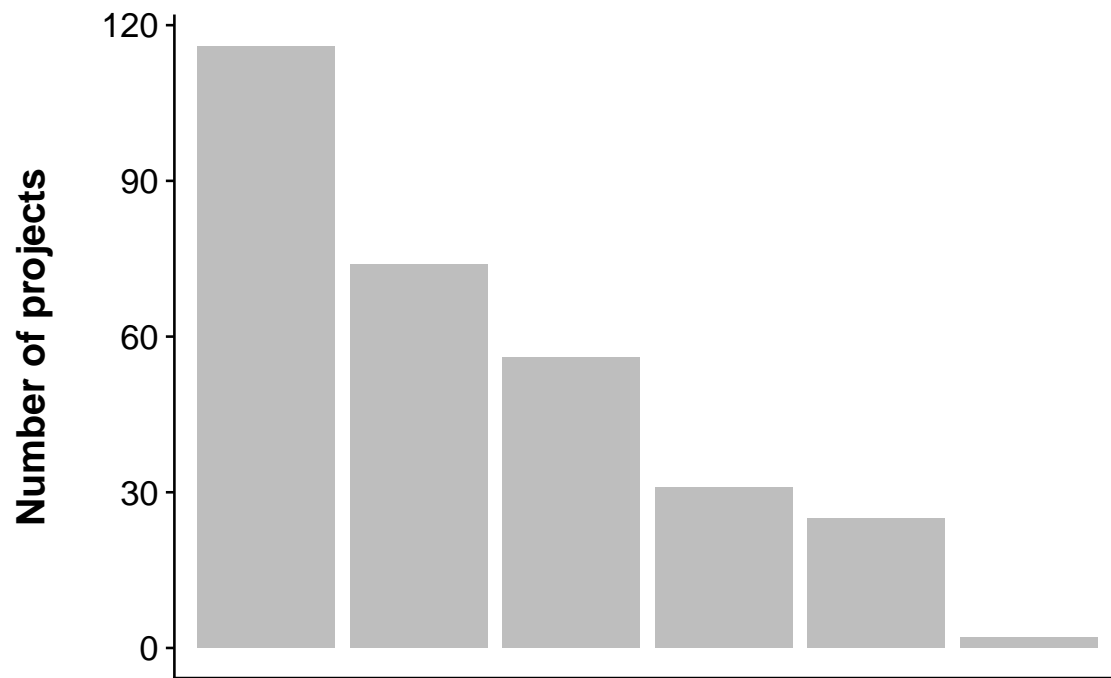
Supplementary figures

S1 Figure. Awardee agency sector

Are there certain **sectors** that are more likely to make data available than others? (based on PI's affiliation)

SuppFigure 1: Percent success of projects by agency sector. The top plot is total number of projects funded for each sector, the bottom plot shows the percent success for the given sector.

NOTES: We stack nProjects/percent graphs to display variation in nProjects between groups as well as variation in relative responses for each group. These data show interesting opposing trends in nProjects and %recovery



Sector

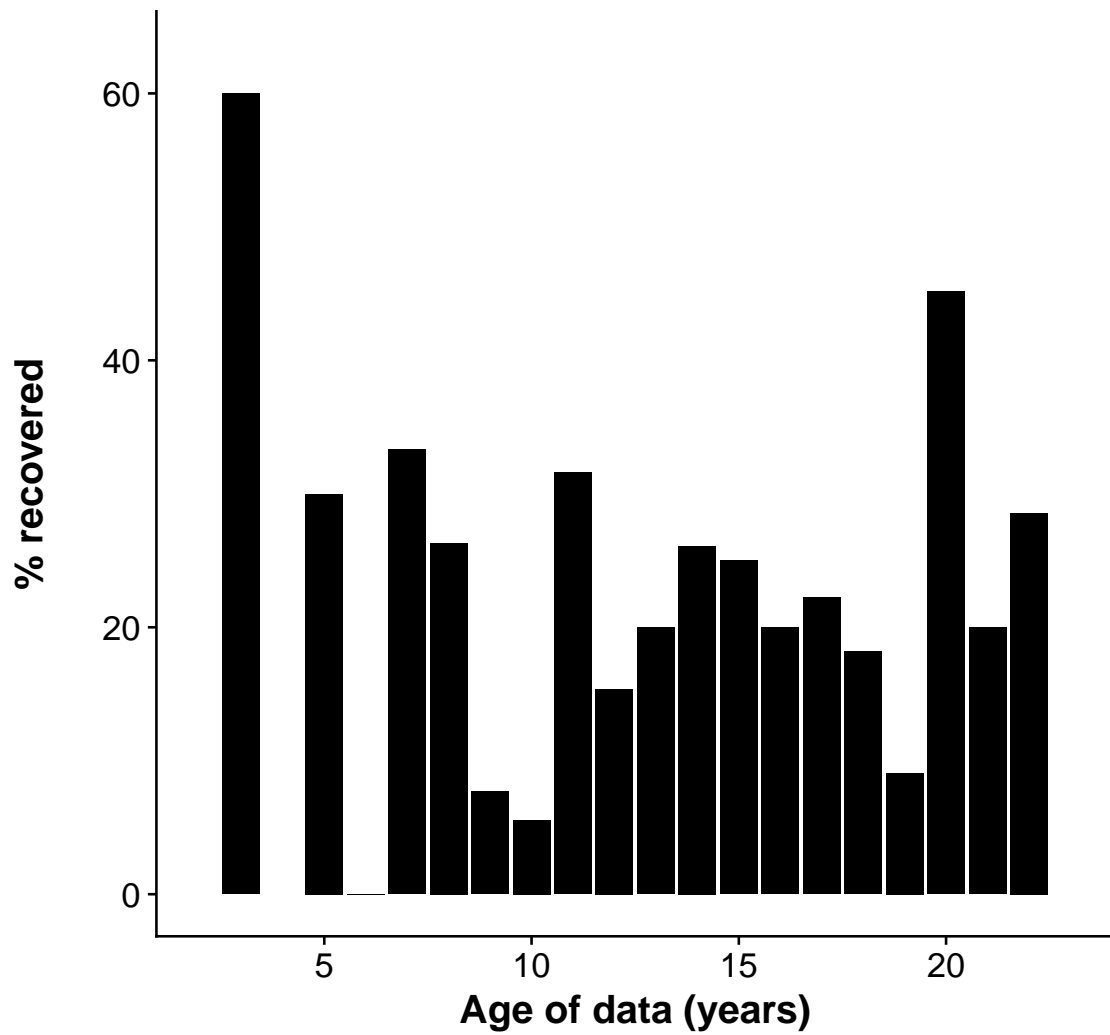
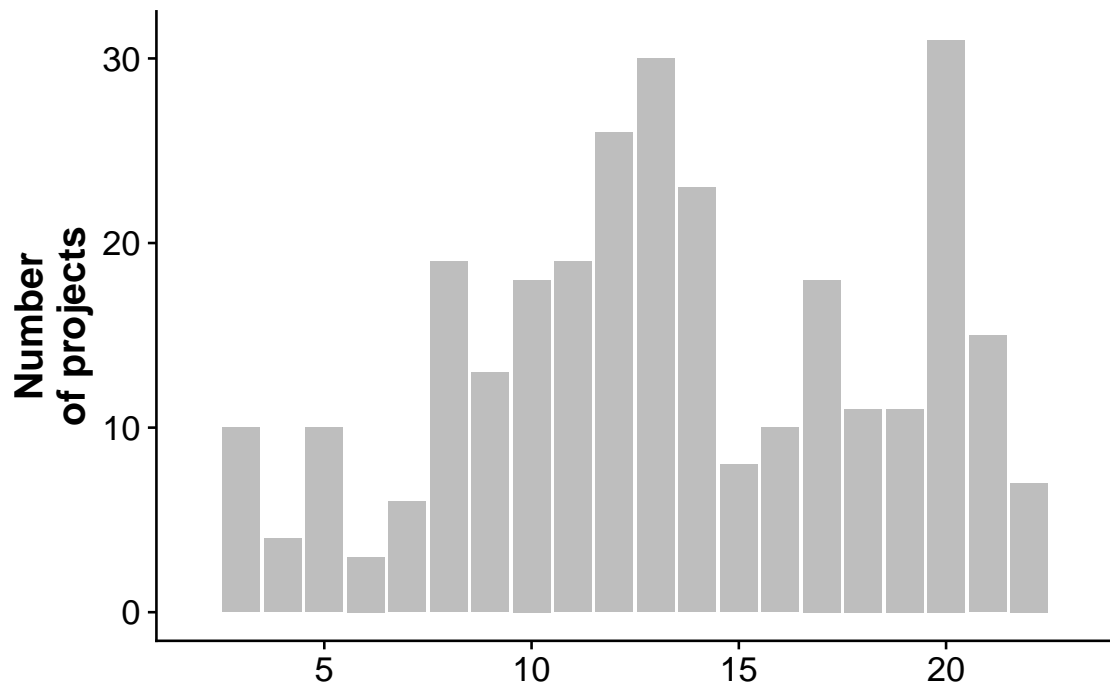
S2 Figure. Temporal trends?

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

SuppFigure 2: Percent success of projects by age of data. Age is calculated based on number of years between the last year of EVOSTC funding and start of the archiving project (2012). The top plot is total number of projects that ended each year, the bottom plot shows the percent success for the given year.

NOTES: We stack nProjects/percent graphs to display variation in nProjects between groups as well as variation in relative responses for each group.

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.



S3 Figure. Temporal trends in reasons?

Are there temporal trends in the reasons given for not sharing data?

SuppFigure 3: Counts of reasons for not sharing data based on the age of the data.

To test if there were temporal trends in reasons given for not sharing data, we plotted the frequency of each reason for not sharing based on the age of the data.

