NOAA Technical Memorandum OAR WPO-002

Hidden data: Using text data to answer questions of value, societal impact, and portfolio analysis

September 2022

Michele (Micki) Olson, OAR/WPO
Alice Grossman, AAAS Science and Technology Policy Fellow
Gina Eosco, OAR/WPO*
Laura Newcomb, OAR/OSS*
Joseph Conran, OCFO/PRSSO*

* These authors contributed to study design and research question development, but did not contribute to analyses or the writing of results and interpretations for this report.

NOAA/Weather Program Office
Silver Spring, Maryland
September 2022

noaa NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION / Office of Oceanic and Atmospheric Research

OAA Technical Memorandum OAR WPO-002

Hidden data: Using text data to answer questions of value, societal impact, and portfolio analysis

Michele (Micki) Olson, OAR/WPO
Alice Grossman, AAAS Science and Technology Policy Fellow
Gina Eosco, OAR/WPO
Laura Newcomb, OAR/OSS
Joseph Conran, OCFO/PRSSO

NOAA/Weather Program Office
Silver Spring, Maryland

September 2022

NOTICE

Acknowledgements

Thank you to our interviewees and survey respondents for sharing your thoughts with us. Thank you to the reviewers at the National Oceanic and Atmospheric Administration, Stephen Elliott, John Ten Hoeve, Dorothy Koch, and Jeffrey Wielgus, who also provided valuable contributions to this work.

List of Acronyms

Abbreviation	Description
CaRDS	Capabilities and Requirements Decision Support
FCA	Formulation and Congressional Affairs
ICR	Information Collection Request
NESDIS	National Environmental Satellite, Data, and Information Service
NOAA	National Oceanic and Atmospheric Administration
NOSIA	NOAA Observing System Integrated Analysis
NRDD	NOAA Research and Development Database
NWS	National Weather Service
OAR	Oceanic and Atmospheric Research
OMB	Office of Management and Budget
PRA	Paperwork Reduction Act
R&D	Research and Development
TPIO	Technology, Planning and Integration for Observation
WPO	Weather Program Office

1 Introduction

The 2018 Foundations in Evidence-based Policymaking Act of 2018 (Public Law 115-435, "Evidence Act") emphasizes the need for incorporating data into decision-making, including program evaluation, such as evaluating NOAA's research and development portfolio. To that end, NOAA has institutionalized several resources for collecting and archiving information (i.e. data) about its research and development (R&D) portfolio at both the agency (e.g., the NOAA Research and Development Database or "NRDD" and the NOAA Institutional Repository), line office (e.g. the National Weather Service Customer Satisfaction Database), and program (e.g Sea Grant's Planning, Implementation, and Evaluation Resources System) levels. NOAA staff use these resources to conduct organizational analyses at the request of both external (e.g., Congress) and internal (e.g., NOAA staff offices) entities. Often called portfolio analysis, these systems are valuable tools in analyzing and characterizing an agency's R&D portfolio (Grossman et al, 2022).

For example, one such resource, the NRDD was established in 2017 to provide "a single repository for consistent, comprehensive information on NOAA's R&D portfolio" (2019 NRDD Annual Report). Data housed in the NRDD can inform answers to questions regarding the composition of NOAA's R&D portfolio. For example, the NOAA Science Council reviews transitions quarterly to track progress and identify potential barriers to transitioning R&D. To inform this review, the NOAA analysts use a combination of NRDD data and supplemental information from NOAA Line Offices, to assess the NOAA projects transitioned into operations, commercialization, application, and/or other use (i.e., transitioned from Readiness Level 8 to 9, Gal-Edd et al., 2020).

As illustrated in the example above, the value of these institutional resources comes from their use. In order to evaluate how these resources can best serve the agency in planning, monitoring, and evaluation of R&D through a portfolio analysis management system, user needs are paramount. This project pursues the first step of this process by identifying user needs of institutional management systems.

We conducted a series of exploratory interviews in Spring 2021 with analysts conducting organizational analyses across OAR, as well as a few key OAR stakeholders in other NOAA line offices. This report presents these interview results, which were conducted with the purpose of answering three primary questions:

1. What inquiries do NOAA/OAR staff receive related to planning, managing, and evaluating NOAA R&D?
2. What barriers do NOAA/OAR staff perceive exist in answering these inquiries?
3. What information does NOAA/OAR currently not collect but needs to in order answer these inquiries?

2 Method

Participants were selected using a purposive sample technique, whereby the authors first created a list of a potential participants based on their ability to provide: (a) in depth knowledge about various NOAA datasets, databases, performance measures, analysis types, science portfolios, etc., and (b) diverse insight based on their role and experiences within the agency.

A combination of individual and group interviews were used, which are discussed in more detail below. Twenty-four participants completed individual interviews, whereas 31 participants completed group interviews, for a total of 55 participants. Note-takers for both individual and group interviews were trained by the second author to ensure that the notes adequately captured and provided a complete overview of what was discussed.

Finally, as this is an OAR case study, participants were primarily from OAR (~76%) and included both federal and contractor staff.

Table 1: Number of participants from each NOAA Line Office for individual and group interviews

Line Office	**Individual Interviews**	**Group Interviews**
National Environmental Satellite, Data, and Information Service (NESDIS)	5	0
NOAA Headquarters	1	5 (1 group interview at 5 participants)
National Marine Fisheries Service	0	0
National Ocean Service	0	0
National Weather Service (NWS)	2	0
OAR	16	26 (2 group interviews at 13 participants each)

Office of Marine and Aviation Operations	0	0
Total (n=55)	**24**	**31**

Individual Interviews

Prior to the individual interviews, interviewees were instructed to fill out a short survey and indicate which type(s) of data they typically use (e.g., peer reviewed publications, transition plans) to make decisions and if these types of data originate inside their office, in another NOAA office, or external to NOAA.

Interviews followed a semi-structured interview protocol (see Appendix), whereby all interviewees were asked what data and/or databases they commonly use, if they developed any tools, resources, or databases that utilize this data, their office performance measures, and what type(s) of questions they typically try to answer on a regular basis.

Group Interviews

Select offices were invited to participate in group interviews. This method was used as a way to gather thoughts from many relevant participants simultaneously, as opposed to having to conduct individual interviews with many participants separately. In the group interviews, participants were asked how their office defines success, what type(s) of questions they try to answer, and the specific types of data they commonly use.

Data Analysis

Using NVivo qualitative analysis software, individual and group interview notes were analyzed inductively using a combination of automated[1] and manual coding by the first author. The knowledge gained through the interviews informs the following sections of this document.

All data analysis was completed by an author who was *not* interviewed for this project. The authors of this report who were also interview participants did not conduct or influence data analysis nor the subsequent results presented below. Instead, these authors helped contextualize results within the NOAA organizational framework, as well as aided in exploring applications and implications of results.

[1] NVivo Automated Coding Strategy

4 Results

The following section provides an overview of the type(s) of questions participants receive and/or would like to know, as well as the type of data needed to answer these questions, and perceived or emergent barriers.

Value

By far the most prevalent questions participants asked and/or recieved centered on *what is the value of NOAA information, data, research, products, and services, and to whom is it valuable?*

Participants noted that current individual data sources cannot fully answer these types of questions. For example, the NRDD provides data on past and present R&D projects, but the questions people are asking require additional data sets to answer in full. As such, NOAA data sources, systems, and/or databases must be linked.

For example, one participant asked, “Is there some way to visualize groups of how [research] connects to the economy, or [line office] missions so we can visualize our place in the chain of the process? We fund lots of research that funnels into different things...I’m thinking of ways to visualize our impact.” Another participant similarly noted, “[what] If we had a system of systems to connect to operations and then back to R&D and show the interconnections? Is there a way to measure the success of research investments, like [linking] journal articles in the NRDD?”

Participants mentioned specific types of connections between systems they would like to see established to better answer their questions, including connecting the

- NRDD to the NOAA Observing System Integrated Analysis (NOSIA). NOSIA is a NOAA-wide portfolio analysis and data suite hosted by NESDIS that connects observations to NOAA-internal downstream databases, models, and products alongside prioritization, organizational, and data satisfaction information,
- NRDD to the NWS Capabilities and Requirements Decision Support (CaRDS) database, and
- NOSIA to the NOAA Institutional Repository.

All the specific data sources participants mentioned are included in Table 2.

Table 2: Data sources mentioned during interviews

Data Source	**Abbreviation**	**Public within NOAA?**	**Led By**
Capabilities and Requirements Decision Support	CaRDS	Yes	NWS
Earth Resources Observation Systems	EROS	Yes	NESDIS
Fiscal Year 2020 Weather Act Gap Analysis database	n/a	No	OAR & WPO
Grant, Contract, Cooperative Agreement, Proposals/Applications within funding offices	n/a	No	Program Offices
Grants Online	n/a	Yes (need to register for separate login)	External to NOAA
NESDIS User Database	n/a	No	NESDIS
NOAA Institutional Repository	NOAA IR	Yes	OAR
NOAA Observing System Integrated Analysis	NOSIA	?	NESDIS
NOAA Research and Development Database	NRDD	Yes (need to register)	OAR
NWS Virtual Lab	VLab	Yes	NWS
NWS Customer Service Database	n/a	?	NWS
Paperwork Reduction Act (PRA) Information Collection Request (ICR) Packages	PRA ICR packages	Accessible to General Public via OMB site	RegInfo.gov Information Collection Review database for Paperwork Reduction Act submissions https://www.reginfo.gov/public/do/PRAMain
Planning, Implementation, and Evaluation Resources System	PIER	No (can access via approval from Sea Grant)	OAR Sea Grant
Lab/Program Specific Databases	n/a	No	NCEI, NWS, AOML, WPO all noted they developed internal

			databases to track their work through commercial platforms.

Finally, participants also noted the challenges or barriers in establishing connections in these data sources across line offices, since such connections would largely rely on cross line office collaboration. The challenges mentioned by participants include:

- Stovepipes between Line Offices
- Lack of relationships between Line Offices
- Different missions, and therefore foci, between Line Offices

Societal Impact

Related to questions of value were questions that centered on "*what is NOAA's societal impact, relevance, or outcomes?"*

Several participants noted that, in their opinion, NOAA lacks information and/or data about societal outcomes. For example, one participant stated, "We have a mission and measure the technical aspects (e.g. tornado lead time, accuracy, deviation from track of hurricanes). We have the accuracy measures, timeliness, etc. but we don't have outcomes. We don't know how to think about [societal] outcomes."

Similarly, other participants stated that NOAA needs more information on its end users. For example, one participant asked, "Who are we reaching? That is what is most important for us to track and how can we improve the delivery of this information?" Specifically, these participants indicated that NOAA needs end user information to assess if service delivery or racial disparity gaps exist.

One participant noted that similar to answering questions of value, NOAA's societal impact can be informed through analyzing documents containing qualitative or text data. For example, documents such as social science research proposals, contracts, Office of Management and Budget (OMB) Paperwork Reduction Act Information Collection Request (PRA ICR) packages, and less formal end user databases contain useful information about what societal outcome data NOAA might have. These documents are a form of text data. For example, research proposals have useful information related to what method researchers are using, who they are sampling (e.g., emergency managers, forecasters), and the geographic areas they are studying. By extracting key pieces of information from these documents, NOAA could build a database of meta-data, which would provide the agency with an accurate picture of what (a) societal impact

information has been collected, and (b) what types of societal impact data NOAA does not currently have, but should collect via future R&D investments. When end user information is embedded in documents, NOAA can also examine connections between types of funded research and impacts on certain societal groups, which can help prioritize future investment for underrepresented or marginalized populations.

Portfolio, Legislative, and Policy

OAR participants specifically noted they often have to answer overarching portfolio, legislative, and policy questions that emerge from NOAA leaders.

Although the NRDD is the data source typically used to answer these types of questions about an entire research or science portfolio, participants noted several limitations with NRDD data. For example, as one participant stated, "[a] Limiting factor is the accuracy and completeness of data. If complete and accurate, then the NRDD would always be faster." Another participant noted that in querying topics, different words are often used to describe the same phenomena. For example, in a request to categorize NOAA's satellite investments, the word "satellite" may not appear in any of the NRDD fields. Instead, information may be under "geo stationary," "GOES," or other associated terms. The NRDD, then, may benefit from an advanced indexing analysis - evaluating the process of assigning a suite of keywords to R&D projects.

Furthermore, two data calls from fiscal year 2020 highlight limitations with relying solely on the NRDD. The first request asked for a complete list of all NOAA funded social science projects and/or activities. Those responsible for assembling the information started with the NRDD but due to missing and/or incomplete entries performed a manual data call. Specifically, these individuals determined, "[we] Were asking more specific questions than the NRDD could provide. Enough gaps existed where we had to reach out [to the line offices]."

Another request asked OAR programs and labs to map their current activities and investments onto sections of the Weather Research and Forecasting Innovation Act of 2017 (Public Law No: 115-25, aka the "Weather Act,"). The purpose of this request was to help NOAA to better understand which areas of the legislation require further R&D investments. However, budget information was needed to answer part of this request, and the budget information in the NRDD was deemed incomplete by participants. Furthermore, this information is not tracked by Formulation and Congressional Affairs ("FCA") or NOAA Legislative Affairs; therefore a manual data call was deemed to be required for all programs and labs.

Since the information for these tasks was not located in any central location or in standard formats, both of these tasks took multiple weeks, as new data needed to be collected and

analyzed. This finding indicates that NOAA could diminish the time and effort needed to complete these tasks, and optimize answering other questions from NOAA leadership and Congress, if the data needed for analysis was either consistently collected or easier to access and use.

5 Discussion

This report outlines the results of an interview study whose purpose was to determine:

1. What inquiries do NOAA/OAR staff receive related to planning, managing, and evaluating NOAA R&D?
2. What barriers do NOAA/OAR staff perceive exist in answering these inquiries?
3. What information does NOAA/OAR currently not collect but needs to in order answer these inquiries?

NOAA/OAR staff primarily receive questions and/or would like to answer questions related to NOAA's value. One key finding revealed that participants expressed a desire for more access or connections amongst NOAA data sources (including databases). Although this was not explicitly mentioned by participants, connecting these data sources with each other could establish a value chain data system. A value chain is a process map that shows the progression from how an R&D output links to other outputs and to societal benefits, such as from observations to models to forecasts to warnings to empowering citizens to take protective action. Linking data and databases across these areas, such as linking projects in the NRDD to societal impact data, could then create the first foundational step toward economic valuation.

Another set of inquiries centered around NOAA's societal impact and answering overarching portfolio, legislative, and policy questions . Interestingly, participants noted that the data needed to answer these questions could be and/or is primarily textual data. Specifically, participants noted that if NOAA had the capability to analyze and connect its textual data assets, the agency would be better poised to assess the value and impact of various programs and portfolios, as well as its impact on society. Furthermore, the ability to link these data sources, which are composed primarily of textual or qualitative data, would allow for more sophisticated types of textual data analysis, including text mining and other natural language processing techniques, as these types of analyses require a larger corpus of data than an individual data source can provide.

However, barriers exist for this type of data. Specifically, how NOAA's textual data is produced and organized, as well as its completeness, correctness, and consistency, plays a role in the

ability to answer the questions raised by participants. Existing NOAA strategies and visions for data (e.g., the NOAA Data Strategy) and R&D should inform future systems to ensure unified effort towards reaching agency-wide goals and objectives. Federal legislation around data (e.g., Foundations for Evidence-based Policymaking Act, Public Law No: 115-435) can also be considered, such that any new data system supports and increases efficiency in adhering to requirements.

6 Limitations & Future Directions

Certain methodological limitations should be noted. Specifically, only one author analyzed the interview data. Ideally for the purpose of triangulation, more than one individual researcher would analyze and corroborate study findings.

In addition, future work should prioritize collecting data from a more representative sample, including from a wider array of Line Offices, staff offices, programs, as well as from individuals with varying job duties. Future work should also consider studying if there is an opportunity to develop a unified system of NOAA wide textual data assets.

Finally, given the results of these interviews, future work may also consider developing a strategy that centers on textual data. Such a strategy could be modeled on the existing NOAA Data Strategy.

7 Conclusions

NOAA/OAR staff identified a need for resources centered around answering questions related to NOAA's value, societal impact, and portfolio and legislative analysis. These interviews and focus groups highlight that these resources are primarily in the form of textual data and the potential for uses of textual data at NOAA.

Participants identified a latent demand for access to organized, machine readable textual data both to improve accuracy and efficiency of current analyses and to answer new questions. Some of the analysis of these data would be valuable exploratory research, and respondents indicated a great need to increase efficiency and ability in answering questions coming down from NOAA leadership as well as external stakeholders. Examples of current data gaps show immediate areas for improving existing textual data management systems, and widespread potential uses suggest a widespread opportunity to increase data availability and system access for NOAA users.

8 References

Foundations for Evidence-based Policymaking Act of 2018. Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/4174

Gal-Edd S, Laster M, & Newcomb L, (2020). The NRDD Annual Report: NOAA Research and Development Database. Retrieved from:NRDD Annual Report 2020_2_AUGUST-2021.pdf

Grossman A, Eosco G, Newcomb L, Conran J, Olson M. (2022). Research and Development Portfolio Analysis: Lessons from Select Federal Agencies. NOAA Technical Memorandum. [doi]

National Oceanic and Atmospheric Administration, (2020). NOAA Data Strategy: Maximizing The Value of NOAA Data. Retrieved from: https://sciencecouncil.noaa.gov/NOAA-Science-Technology-Focus-Areas

National Oceanic and Atmospheric Administration, (2019). NRDD Annual Report 2019. Retrieved from: NRDD Annual Report 2019 - Final Draft - 031920.pdf

Appendix: Interview Protocol

Questions for Everyone

1. What is your role and position?
    a. How long have you had that role?
    b. How long have you been at your agency in that or other roles?
    c. How does your role fit in with your office's mission?
2. What are the processes, data-driven or not, that support your agency and office missions?
3. What data do you have? (reference back to survey response for NOAA staff)
    a. Tell me more about the text data you listed in the survey. Which of these documents were already created by your office versus submitted by others to your office?
4. Has your office developed tools and resources to utilize text data?
5. What was the process to develop the tools and resources you have to utilize text data?
    a. Top down or bottom up?
    b. Ease of institutional buy-in?
    c. How new are they?
    d. Was there an external or internal mandate that spurred the development?
6. What NOAA and non-NOAA text data do you use?
    a. And how?
7. What NOAA databases and resources do you use on a regular basis?
    a. Which do you find easy to use? And why?
    b. Which do you find difficult to use? And why?
    c. Where do you go if you can't answer your question with existing NOAA databases?
8. What NOAA databases and resources do you use infrequently or just once?
    a. Which do you find easy to use? And why?
    b. Which do you find difficult to use? And why?
9. Has your office/Program ever created its own database or central system, and if so, for what purpose?
10. What types of questions do you try to answer with the NOAA data you use?
    a. For example, taskers, congressional inquiries, budgeting, communications, ….
11. What types of analysis do you currently perform on NOAA text data/resources and metadata?
    a. Are there analyses you would like to be able to do but can't?

Question for NOAA staff

1. Follow up on document list from short survey form

    a. How are these documents accessed internally and externally?
    b. What are they used for?
2. What performance measures do you currently measure and report on projects? On programs? Online offices?
    a. Where do you get the data for these metrics?
3. What performance measures do you not currently have, but would like to use to evaluate NOAA programs? Projects? Line offices?
    a. What are the barriers to using these measures?

Questions for NOAA Data Enterers (NRDD, Library, etc.)

1. Into how many different databases/systems do you enter information?
2. Which systems/databases are easiest to input to? Most difficult? Why?
3. How accurately do you think the database usually captures information?
    a. What are the barriers to entering more accurate and precise information (ex time, money, accountability, update frequency...)?
    b. What elements make the data entry easy, quick, etc.?
4. How often are you contacted to revise information that you have entered?
    a. What are the reasons?
5. How often do you contact people for user support?
    a. Can you tell us about your experiences with user support?
6. How much time does it take to enter the information you are charged with entering?
    a. Is this a reasonable amount of time for you?

Questions for NOAA Data Extractors

1. What NOAA resources and databases do you frequently use?
    a. What for?
    b. What, if any barriers have you encountered?
    c. (How) did you overcome them?
    d. Is there a feedback/comments channel? Have you provided input?
2. What NOAA resources and databases have you used infrequently or just once?
    a. What for?
    b. Why do you use these less frequently?
    c. What, if any barriers have you encountered in extracting data?
    d. (How) did you overcome them?
3. Are you aware of any other resources or databases that you don't use but might be useful to you?
4. Which systems/databases are easiest to extract the information you need about social/textual data? Most difficult? Why?

5. Is there anything I didn't ask about that you think I'm missing?
6. Who else should I talk to?