NOAA Technical Memorandum OAR WPO-001

Research and Development Portfolio Analysis: Lessons from Select Federal Agencies

September 2022

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

NOAA/Weather Program Office
Silver Spring, Maryland
September 2022

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

NOAA Technical Memorandum OAR WPO-001

Research and Development Portfolio Analysis: Lessons from Federal Agencies

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

NOAA/Weather Program Office
Silver Spring, Maryland

September 2022

NOTICE

Acknowledgements

Thank you to the current and former staff and contractors who participated in interviews for this project, including the Department of Energy (DoE); Dr. Yanxhi (Ann) Xu, *Co-Founder and CEO at ElectroTempo, Inc and former Senior Technical Advisor for Impact and Assessment, ARPA-E, DoE,* and Patrick Finch, *Lead Associate, Booz Allen Hamilton supporting DoE*; the National Institute of Health (NIH), and National Science Foundation (NSF). Thank you to the reviewers at the National Oceanic and Atmospheric Administration, Gina Digiantonio, John Ten Hoeve, Dorothy Koch, David Turner, and Michael Smith, who also provided valuable contributions to this work.

List of Acronyms

Abbreviation	Description
NOAA	National Oceanic and Atmospheric Administration
ARPA-E	Advanced Research Projects Agency - Energy
DOE	Department of Energy
FAIR	Findable, Accessible, Interoperable, and Reusable
HIPAA	Health Insurance Portability and Accountability Act
NIH	National Institute of Health
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
NLP	Natural Language Processing
PACOI	Portfolio Analysis Community of Interest
PDF	Portable Document Format

Lexicon

Term	Definition
Data	Recorded information, regardless of form or the media on which the data are recorded
Data Asset	a collection of data elements or data sets that may be grouped together
Data Science	Interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from large and/or complex data sets.[1]
Social Science	The process of describing, explaining and predicting human behavior and institutional structures in interaction with their environments
Textual Data	Data in the form of language expressions such as keywords, phrases, clauses, sentences, and paragraphs.

[1] NIH Strategic Plan for Data Science, 2019. <https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf>

1 Introduction

With the passing of the Foundations for Evidence-based Policymaking Act of 2018 (Public Law 115-435, "Evidence Act"), there is increasing demand for Federal Agencies to characterize and evaluate their research and development portfolios. While planning, monitoring, and evaluating R&D is a continuous process, the Evidence Act highlights an even higher level need - to understand how an entire research portfolio contributes to societal outcomes. As such, Agencies are in need of methods and tools to conduct portfolio analysis.

Portfolio analysis examines an agency's R&D portfolio from the program to the enterprise level in support of planning and decision making. Portfolio analysis supports both strategic and tactical decision making. At the strategic level, portfolio analysis can give an overall picture of how the currently supported research is advancing the agency's priorities and goals and identify gaps in the mission. At a tactical level, portfolio analysis provides an additional data-informed tool to assist program managers in setting priorities and funding decisions.

Portfolio analysis utilizes a suite of textual data assets and computational tools to digest large quantities of information and summarize it in a digestible form. A recent increase in the availability of textual data assets, including proposal, project, and publication databases that systematically catalog scientific work and outputs, as well as advances in the fields of qualitative coding, machine learning, and natural language processing has opened up an increasing arsenal of portfolio analysis capabilities.

As an applied science agency, the National Oceanic and Atmospheric Administration (NOAA) conducts science in service to society. Analytical portfolio analysis capabilities present an opportunity for the agency to better understand the linkages and connections between its research and development (R&D) portfolio and the intended societal benefits. As NOAA looks to build out and expand its existing capabilities, the agency can learn from how other U.S. government agencies have implemented textual management and portfolio analysis systems. These examples highlight best practices and challenges to anticipate in system development as well as for potential uses and outcomes of such a system.

This paper gives an overview of the inception, development, and maintenance of portfolio analysis management systems at three other Federal agencies: the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Department of Energy (DoE). The information in this document was gathered largely from interviews with agency staff, as well as from formal documentation. It also discusses how these systems have been used to inform and enhance agency operations. The best practices and lessons learned help devise a process for the development of a similar system at NOAA.

2 Agency Examples

NOAA staff researched comparable portfolio analysis management systems at three other Federal agencies to identify best practices and lessons learned. While every agency functions uniquely, a broad understanding of system development, maintenance, and uses for textual data has universal implications.

The team interviewed agency staff from NSF, NIH, and DoE who have developed and utilized methods and tools to perform portfolio analysis. Loosely structured video interviews included discussions around (1) strategy or process development to determine requirements for textual data management and analysis; (2) actors involved in system development and use; (3) barriers encountered in system development; (4) data production methods; (5) system use over time; and (6) types of data included in the system and types of analysis conducted with those data. These topics helped develop an understanding of the underlying purpose and use of portfolio analysis management systems at each agency, how data systems in support of portfolio analysis were developed and maintained, and how staff utilized these systems to manage their R&D portfolios.

3 Purpose and use of Portfolio Analysis Management Systems

The interviewed federal agencies use their portfolio analysis management systems to support overarching agency goals, monitor objectives, and measure performance. Potential uses include developing a better understanding of and facilitating workforce development, and adhering to legislative requirements for open data and information. Example analyses and their uses are listed out in Table 1.

Table 1: Example Use Cases of Portfolio Analysis

Analysis	**Uses of analysis**
What accounts for a lower rate of funding of NIH applications submitted by African-American/black scientists relative to white scientists? (Hoppe et al, 2019)	Assist with direction of future funding priorities in the agency
Where are the research gaps in area x?	Help to prioritize internal research in the area and/or drive external funding decisions
Who works in the field of x and is qualified to evaluate proposals in this field?	Identify new reviewers to invite for specific topic areas to support increasing diversity in reviewer pools
What scientific advances are most likely to translate into clinical research? (Hutchins et al., 2019)	Predicting translational progress in biomedicine

What is the impact of high risk high reward research?	Reporting out on metrics of program impact such as number of projects partnered with other government agencies, and number of patents issued.

Underpinning all of these examples are two core requirements:
(1) The data is accurate.
(2) The data is in a machine readable form.

With these requirements met, the data can then be analyzed utilizing a system for searching and clustering data, including textual data.

4 Developing and Maintaining Portfolio Analysis Management Systems

Developing a System

All three agencies use in their portfolio analyses a significant amount of textual data, often scraped from research proposals and project reports. Using third-party software tools and internal development teams, these agencies created a database of these textual data assets in the form of machine-readable documents, often encoded in the portable document format (PDFs). They then utilized software for data analytics including R and Solr to analyze the information contained in those documents through textual mining and natural language processing (NLP).

Building Social Capital

Each agency began with a single project champion with a goal to develop a textual data management framework and, in the cases of DoE and NIH, hire staff to maintain and support the data infrastructure. Beyond the project teams, support among agency leadership and potential system users was critical for success. In order to increase support among agency staff for textual data systems, the basic principles of social and economic exchange theory (to minimize burden and cost and maximize benefit and reward) apply. Existing systems at NIH, NSF, and DoE have worked to minimize the burden of data entry by standardizing formats and creating a centralized location for documents to allow for auto-population of metadata when staff or external collaborators enter an article into the system.

The process for developing systems to input, manage, access, and analyze textual data across other agencies varied, spanning both top-down and bottom-up approaches. NIH recognized the need for analytical resources to support data-driven decision making and built up a robust team, starting with five employees and building up to thirty staff including software developers to build a tool around their requirements. NSF's search tool began with just one staff member, using an individual server to house data and tools. The system's user base at NSF has expanded over many years, though it is still centralized in one office. Gradually, other staff are embracing the

capabilities of the repository. Over time, the system structure evolved to answer new questions, and new standards and capabilities were developed on an ad hoc basis as needs become apparent. Usability and speed of the systems at NIH and NSF have developed more or less in parallel, catering to specific needs at each organization.

Good communication and engagement helped rally support and interest from staff at various levels. The bottom-up approach amongst select staff to develop a system at NSF contrasts to the coordinated approach featuring monthly demos and diverse focus groups at DoE. This more top-down approach at DoE helped secure support at the agency and gradually improved their internal user-base. At NIH, the top-down attention led to surveys, on-site briefings, office hours, and training curricula that aided the agency in promoting and expanding the user base of the analytical tools.

Managing Data

Safety and security of the data is paramount to maximize the usefulness of the systems. NIH provides a valuable model for protecting data security and privacy due to the personal and sensitive nature of biomedical and associated data sources in concordance with The Privacy Act of 1974, and The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Since different textual data sources have different privacy and security needs, these three agencies developed separate systems for internal- and external-facing document databases. Agency staff involved in the development and maintenance of these systems discussed the need to follow legislative requirements for open data, protect privacy and security, and adhere to principles for digital data such as findability, accessibility, interoperability, and reuse (FAIR) data principles (Wilkinson, 2016). Staff noted the enhanced ability to learn from and collaborate more internally and with national labs and experts in other fields with centralized systems for textual data.

5 Challenges and Solutions

Standing up a new data system at a Federal agency is a large and complex task, and while each agency has unique hurdles, some common challenges can be addressed with transferable solutions (Table 2). Other federal agencies identified many similar challenges as those facing NOAA. Barriers to overcome include: monetary and temporal costs of generating and managing data for the agency and for associated scientists; siloed data ecosystems; lack of interoperability between datasets; difficulties analyzing value and efficiency of data science tools; and minimal existing third-party tools and services that meet the specific data science and NLP needs of the agencies (NIH, 2018). Personnel associated with ARPA-E specifically identified the cost and time associated with gathering and formatting PDFs as a major reason that their system hasn’t expanded to other programs at DoE despite significant interest.

As part of the effort to overcome these barriers, agencies identified a need to extend data plans to consider textual data. The inclusion of both quantitative and qualitative data in the 2020 Strategic Plan For Data Science at NIH follows the agency's acknowledgement of the growing amount and importance of various types of big data and associated metadata in biomedical fields. The NIH Plan also mentions the timely relevance of cloud storage and machine learning among other artificial intelligence techniques that can be leveraged within an agency or through public-private partnerships to increase the value of well organized and clean data through "cost-effective ways to capture, access, sustain, and reuse" data (NIH, 2018).

Cloud services also emerged as a way to house large textual databases on a common architecture, with both digital and physical infrastructure for data storage, connectivity, management, and sharing. Some agencies had better collaboration with their information technology offices leading to quicker implementation, suggesting that a cross-office approach with a clear strategy can be beneficial to align goals and actions in a collaborative manner prior to implementation.

Table 2. Data management, workforce, and big picture lessons learned

Consideration	Best Practice	Challenge to Anticipate
Initial System Development	Find a champion in leadership and build staff capacity to build up system for broad acceptance and integration; anticipate and communicate slow ramp-up	Resistance to change, especially in cases where existing systems need to be incorporated and given demands for time and money
Workforce implications	Build staff capacity in textual data science	Cost of new personnel, buy-in for new duties and tools amongst existing staff
Privacy and Security	Develop both internal and external systems to maintain internal security and privacy while providing transparency to the public	Additional capacity needed to develop and maintain two systems
Data and AI ethics	Follow FAIR data principles; Understand how automated elements of the system works	Potential for "black box" processing with AI limiting human checks and balances; perpetuation of existing inequities when learning from past data
Public and political scrutiny	Data limitations and biases should be mitigated, and clearly addressed up front	Increased transparency both internally and externally can bring additional scrutiny.

6 Conclusion

Identifying lessons learned and best practices from other Federal agencies in the planning, development, and implementation of textual data asset management and analysis will inform NOAA's own strategy development. The interviews helped identify best practices for adopting a textual data management system including, identifying where textual data assets lie in the agency and organizing them in machine readable formats. Other elements of a new system, such as where to house the system, who to involve, and what success looks like are varied.

With all of this information, NOAA will be able to learn from these agencies and subsequently develop a right-sized strategy to uniquely fit the agency's needs and resources. More formal systems for textual data asset management will lead to the ability to perform project, portfolio, and enterprise level analyses to measure success towards NOAA's overarching goals and objectives.

7 References

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

Hoppe TA, Litovitz A, Willis KA, Meseroll RA, Perkins MJ, Hutchins BI, Davis AF, Lauer ML, Valentine HA, Anderson JM, Santangelo GM. (2019) Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science Advances* 5 (10). https://doi.org/10.1126/sciadv.aaw7238

Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. (2019) Predicting translational progress in biomedical research. *PLOS Biology* 17(10): e3000416. https://doi.org/10.1371/journal.pbio.3000416

National Institute of Health. NIH Strategic Plan for Data Science. June 2018 <https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf>

Wilkinson M, Dumontier M, Aalbersberg I. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18.