Putting the Cart Before a Lame Horse:
A Case Study for Future Initiatives to Automate the Use of Administrative Records for Reporting Government R&D

Christopher V. Pece[1]
National Center for Science and Engineering Statistics
National Science Foundation

Summary: This paper will propose that future initiatives for government R&D reporting standards are necessary in order to make effective use of administrative record data in the measurement of R&D and to ensure the accuracy, consistency, and quality of data on R&D funding and performance. Despite the theoretical promises posed by administrative record data for measuring government funding and performance of R&D, without standard methods for how R&D is identified for reporting within government financial systems the advantages of using administrative records data cannot be met.

The United States' national portfolio of research and development (R&D) totals more than $456 billion annually as of 2013.[2] Of this amount the Federal government contributes the second largest share at $122 billion[3] annually through the work of over 15 federal departments (e.g., Department of Defense and Department of Health and Human Services) and 70 sub-agencies (e.g., the Defense Advanced Research Projects Agency and National Institutes of Health), as well as 15 independent agencies such as the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF). Measuring the federal contribution to R&D is part of the mission of NSF' National Center for Science and Engineering Statistics (NCSES). All of the data collected from federal agencies about their funding for R&D are derived from traditional surveys. Advances in information technology during the past decade suggest that administrative data may be a viable option for collecting information about federal funding of R&D over traditional survey methods. However, most of the information about federal funding for R&D that are of the greatest interest, such as place of performance and type of R&D, are simply not aligned with agencies' own financial data systems. In other words, financial information is not directly transferable from the accounting systems and financial reports for statistical purposes, and instead require intervening calculations, summarizations, and "guess-work". Thus, with regard to government funding of R&D, administrative data are unreliable without developing appropriate linkages between financial and management systems. This paper will recommend that future initiatives in government-wide standards for identifying and reporting federal R&D spending within government financial and management data systems should be developed before administrative data can be used to supplement, much less replace, tradition funder (or performer) surveys of R&D. The observations and recommendations are likely to be applicable to data compilation efforts across a number of science and technology domains.

[1] The author wishes to acknowledge the contributions of Jeffrey Alexander, Christina Freyman, and John Benskin of SRI International for their technical assistance and corresponding reports in support of the Administration Records Project upon which this paper is based; and John Jankowski, R&D Statistics Program Director, National Center for Science and Engineering Statistics, National Science Foundation for strategic, planning, and technical guidance on both the Administrative Records Project and the development of this paper.

[2] National Science Foundation, National Center for Science and Engineering Statistics. 2015 *U.S. R&D Increased in 2013, Well Ahead of the Pace of Gross Domestic Product*. Info Brief NSF 15-330, Arlington, VA.

[3] Ibid.

Following recommendations from the National Academies of Science, Committee on National Statistics (CNSTAT) to improve the efficiency and quality of the *Survey of Federal Funds of Research and Development* and the *Survey of Federal Support for Science and Engineering to Universities, Colleges, and Nonprofit Institutions* – colloquially referred to as the Federal Funds Survey (FFS) and the Federal Support Survey (FSS) respectively – NCSES established the Administrative Records Project (ARP). The premise of ARP, suggested by CNSTAT, was "the survey data are of insufficient quality and timeliness to support many of the demands put on them"[4]; furthermore, "the surveys are increasingly difficult to conduct in times of constrained resources, and their technological, procedural, and conceptual infrastructure has not been modernized for procedure or content in contrast with other surveys in the portfolio [of NCSES]."[5] One option offered for modernizing the survey is to more intensively use microdata in administrative records that are part of the standardized automated reporting systems in the key federal agencies that provide the bulk of federal support to academic and nonprofit institutions."[6] The problem is this assumes is that federal agency financial systems already hold detailed information about the nature of their R&D funding. This suggests that all of the information about an agency's R&D funding activities are already captured within their administrative records, furthermore it assumes that all agencies maintain the same level of attentiveness to this information. CNSTAT correctly noted that "administrative record-based data collection system will be of use to [NCSES] only if those records include information in the categories [NCSES] needs to collect."[7]

Recognizing the need to modernize and streamline the data collection processes while maintaining data quality, and in response to the recommendations from CNSTAT that NCSES "should initiate work with other federal agencies to develop several demonstration projects to test for the best methods to move to a system based at least partly on administrative records"[8] the ARP was established to look at ways to move towards a system of using administrative records in lieu of traditional surveys of federal agencies' R&D spending. This paper will present findings from ARP and show that despite lofty goals, the promise of administrative records for capturing federal R&D, cannot meet the detailed statistical needs of NCSES to reliably and consistently capture federal expenditures for the Conduct of R&D, Type of R&D, Field of R&D (FORD), Place-of-Performance, Performer Type, or R&D Plant. Furthermore, without future initiatives to develop government-wide standards for R&D tied directly to agency financial and management systems, raising expectations on the value of using administrative records data is imprudent.

In order to present a case study for how to improve information about R&D funding in the future this paper is composed of four sections. We will *(I)* describe the challenges with obtaining administrative records data files from the federal agencies. A number of legal and administrative hurdles presented difficulty in obtaining federal agency administrative record files for the ARP study. Secondly, *(II)* we will describe the process used to review and compile information from the data files that we were able to obtain. This section will also provide results of the study to replicate agency R&D reports solely from administrative records. Without internal controls and standards for the identification of R&D we cannot use administrative record data to recreate the level of detail provided from the traditional survey

[4] National Research Council, *Data on Federal Research and Development Investments: A Pathway to Modernization*, 2010, pg. 1.
[5] Ibid., pgs. 1-2.
[6] Ibid., pg. 15.
[7] Ibid., pg. 51.
[8] Ibid., pg. 60.

methods; as a result we are forced to rely on individuals' program-specific knowledge in order to obtain information through surveys. Then *(III)* we will review the results of using supplemental source data from publicly existing information on federal R&D and show that they are often incomplete and inconsistent to be used for detailed statistical information about R&D in lieu of traditional survey collections on Federal R&D spending. Finally *(IV)* we will conclude with a look forward at the initiatives that should be considered in order to make administrative records actually feasible in the future through the development of government-wide standards for identifying and reporting federal R&D spending and the need to develop internal controls for managing information on R&D. Standardize administrative data is necessary to ensure data quality on federal R&D investments and any practical use of federal administrative records.

Part I: Obtaining and Assessing Federal Agency Administrative Records for Measuring R&D

NCSES asked the CNSTAT of the National Research Council to convene a panel of experts to review and offer recommendations to improve and modernize the FFS and FSS. In late 2010 CNSTAT issued their final report, *Data on Federal Research and Development Investments: A Pathway to Modernization*. In this report, the panel made several recommendations on how NCSES could change the processes for collecting R&D data for the annual FFS statistics. Upon implementation these recommendations should improve the timeliness and accuracy of the survey results, reduce the burden on the responding agencies, and improve the coordination between NCSES and data providers in those agencies. The panel made three long-term recommendations for NCSES to explore the possibility of extracting data relevant to R&D activities directly from agencies' administrative data systems, and using those data to produce statistics on federal funding for R&D in lieu of traditional surveys, specifically:[9]

- NCSES, in cooperation with OMB and OSTP, should seek to have all federal agencies that fund or conduct R&D incorporate R&D descriptors (tags) into administrative databases. Ideally, in order to enable identification of the R&D components of agency or program budgets, tags should identify: a specific field of science and engineering; whether a record applies to R&D or R&D plant; and whether the recorded activity is basic research, applied research, or development.

- NCSES should seek endorsement from OMB to work with other R&D funding agencies to incorporate intramural data into existing and future databases or to directly access intramural spending information from performer databases.

- NCSES should initiate work with other federal agencies to develop several demonstration projects to test for the best methods to move to a system based at least partly on administrative records.

In early 2012, NCSES initiated the ARP Pilot to determine the design, process, and feasibility of using administrative data as the basis for producing FFS statistics. In lieu of specific approaches for actually using administrative data from CNSTAT NCSES developed three methods for using agency administrative data, these included:

[9] Ibid., pg. 5.

- **"Clone" accounting records.** This approach involves obtaining copies of relevant agency accounting records and constructing a crosswalk between these "clone records" and the data elements necessary to create Detailed Statistical Tables (DSTs) that are currently generated from data collected by the FFS.

- **Data Tagging.** This approach involves using eXtensible Business Reporting Language (XBRL) or similar forms of reporting to embed predetermined data "tags" into the accounting systems of federal agency respondents and collect/extract the data to produce DSTs.

- **"Third party" records.** This approach involves using "third party" administrative records, such as those obtained from the Federal Procurement Data System (FPDS), Federal Awards Assistance Data System (FAADS), USAspending.gov, and recovery.gov, to create DSTs currently generated from data collected by the FFS.

With technical support from SRI International NCSES hosted a knowledge-building workshop with federal officials to review the recommendations from the CNSTAT panel and the three options for testing administrative records data for R&D. NCSES was able to obtain reaction and alternative considerations from representatives of federal agencies regarding the proposed approaches, solicit input on the actual benefits or challenges of these approaches, and invite new approaches to be proposed. The workshop was held in November 2012 with over 40 attendees from the following agencies:

- Defense Advanced Research Projects Agency (DARPA)
- Defense Logistics Agency (DLA)
- Department of Energy (DOE)
- Department of the Air Force
- Department of the Army
- Department of the Navy, Office of Naval Research (ONR)
- Environmental Protection Agency (EPA)
- National Aeronautics & Space Administration (NASA)
- National Institutes of Health (NIH)
- National Institute of Standards & Technology (NIST)
- National Oceanic & Atmospheric Administration (NOAA)
- National Science Foundation (NSF)
- Office of Management and Budget (OMB)
- Office of Science and Technology Policy (OSTP)

Both NCSES and agency participants acknowledged that using administrative records for survey purposes presented some specific challenges, including:

- The records themselves are designed for agency-related purposes rather than for national statistics and therefore may not be detailed enough.
- Definitions of data items with the same name may be inconsistently applied across agencies.
- There could be confidentiality concerns with data accessibility.

Most agencies represented dismissed the feasibility of using third party administrative data sources as lacking the detail and comprehensive coverage that NCSES needed for generating statistics on agency R&D funding. A handful of agency representatives noted they may be open to experimenting with the clone-file approach while a few recognized the potential value of data tagging but expressed concerns over implementation, costs, and security concerns. NCSES followed up with representatives from six agencies whose processes and data systems offered the best opportunities to test the feasibility of using administrative records to generate R&D statistics. NCSES developed a protocol to review the quality of potential administrative records for measuring R&D.[10]

DoD Dependent Agency Administrative Records

Although DARPA and ONR had some of the more sophisticated data systems and processes for reporting R&D and tracking R&D projects within their own systems, and would have been good test cases for both the clone file and data tagging approach, NCSES was unable to obtain any administrative data files from these agencies. The clone file approach would not work for these agencies because the files would have presented specific information on classified R&D projects and programs. All information from classified R&D projects, other than the contract award ID, are highly restricted even within the agencies themselves and they could not provide these data to NCSES. Similarly these agencies would not be able to provide descriptions or other data needed to develop data tags specific to R&D as it may reveal classified information.

NIST Administrative Records

NIST also had one of the more sophisticated processes for identifying R&D in their accounting systems and even had a number of NCSES survey-specific codes imbedded into their systems of record to identify projects that would be captured in order to complete the NCSES data calls. Unfortunately, NIST was under no obligation to provide detailed administrative records to NCSES and requested NCSES compensate NIST $20,000 for costs incurred to provide the additional detailed records from their data systems that went beyond the in-kind work associated with the standard survey response. NCSES did not want to create a precedent of paying agencies to provide data regarding their R&D funding to the survey and declined to pursue NIST's administrative records any further.

NSF Administrative Records

This left NASA, NIH, and NSF as the only viable agencies remaining to pilot the use of administrative records for R&D. Fortunately data from NSF was the easiest to obtain as no Memorandum of Agreement (MOA) was needed since NSF is the parent agency for NCSES; staff in the Office of Budget, Finance, and Award Management (BFA) were able to transfer a series of files directly to NCSES. NSF's raw transaction records were extracted from the NSF Financial Accounting System (FAS) using a SAS query executed by NSF budget office staff. These files constitute the same data records that BFA staff used to respond to the FFS directly. Many of the necessary variables for reporting to the FFS were either contained in the transaction record or could be easily derived from existing variables in the transaction record. To assign each record to a FORD category for the FFS the NSF budget office applied an algorithm that analyzes the funding organization, funding program, and appropriation. NSF budget staff were also able to provide this algorithm. Initial discussions with NSF's BFA staff indicated that it

[10] See Attachment A.

would not be possible to reproduce all data required by the FFS purely from administrative records, as certain variables are still calculated with substantial human involvement.

NIH Administrative Records

Obtaining data files from the NIH was more complex than from NSF for several reasons. NCSES drafted an initial MOA based on templates developed by the Federal Committee on Statistical Methodology specifically for the acquisition of administrative records files between federal agencies. Nonetheless this did not make a material difference in the timing or process for obtaining approval of the MOA by NIH. Negotiations between NCSES, NIH program staff, IT staffs, and NIH Office of General Council took nearly a year before NIH would sign-off on the MOA. It was only after approval that NCSES was able to obtain the requested data files.

The NIH administrative records data came from multiple sources spread across different offices each with purview over their respective data. For example, data on R&D contracts and some intramural research records were imported from the NIH Business System (NBS) and the Department of Health and Human Services Departmental Contracts Information System (DCIS) Database into a central database called the Information for Management, Planning, Analysis, and Coordination (IMPAC II), the internal NIH database of extramural and other spending records maintained by the Office of Extramural Research (OER). NIH uses the IMPAC II system to generate the specific values reported to NCSES for the FFS combined with values from the NIH Budget Office for applied versus basic research allocations and intramural research values. At NIH, a large portion of the data elements can be derived from data available in the IMPAC II data system, which tracks extramural grant funding. IMPAC II also pulls contract data from the NBS. The major complication in reporting data at NIH to the FFS is the tracking of intramural research spending (that is, research performance). As there is no project-level system for intramural research across the entire NIH, the NIH Budget Office relies on submissions from the individual Institutes and Centers (ICs) to compile and classify that research spending. As a result, any data elements that contain some or all of the NIH intramural research spending cannot be extracted directly from administrative systems. Various amounts for categories of intramural research spending are derived from the figures published in the NIH Budget Mechanism table in its annual submission for the President's Budget Request. The NIH Budget Office generates the figures directly and so they do not constitute administrative records.

NIH does have a sophisticated system for allocating its R&D funding to specific fields of science and engineering. This system employs a text analysis process leveraging both expert knowledge and machine learning, and classifies individual projects based on the project title (if available). While this is also an estimate, as some titles may not be easily associated with a single discipline, it is systematic and consistent over time. Unfortunately NCSES was not provided access to, or classifications from, this system and would need to classify the Field of Research categories based on information provided in the existing systems of record.

NASA Administrative Records

Initially NASA staff had expressed an interest in the use of administrative records data to help improve their reporting of R&D expenditures to both the FFS and the OMB data calls for the President's Budget. Initial interviews with NASA staff in the Office of the Chief Financial Officer (OCFO) Budget Division – the area charged with responding to the FFS – and the Systems Division, revealed that similar to NIH the

records currently generated to respond to the FFS are held in a many systems; therefore, responding to the FFS requires a complex process of extracting records from these systems and collating them to provide the totals requested in the survey. In addition, and unlike NIH or NSF, several key data elements are not captured at all by NASA information systems, including:

- Unique project identifier (helpful for linking records across systems)
- Type of Obligation (Non-R&D, R&D, or R&D Plant)
- Type of R&D (basic or applied research, or experimental development)
- Performer Type
- Place-of-Performance
- Field of Research and Development

These data elements are critical to producing the responses required in the FFS. To cope with this issue, NASA has adopted a number of workarounds of varying degrees of rigor to generate rough approximations of those data elements. Therefore, while the basic transaction information is extracted from existing systems, there is a substantial amount of manual processing that must be performed to complete the FFS. To highlight the complexity of the R&D data call, at no point in this process is any narrative information about R&D activities (e.g., program or project descriptions) linked to any of the transaction records. As a result, there is no way to describe the specific R&D activities associated with any transaction record. This represents an obvious impediment to using either clone records or XBRL tagging to replicate the responses that NASA has submitted to the FFS. Nonetheless NASA provided a file with nearly 20,000 transaction records from their Budget Data Warehouse in order for NCSES to pilot how one might use these to compile statistics on the agency's R&D funding.

Summary

Identifying agencies with data systems that are robust enough to offer details about their R&D initially turned out to be far fewer than CNSTAT or NCSES would have assumed. Complicating the issue is that a substantial amount of R&D is related to national defense activities and these data are often, as a matter of agency policy, unavailable due to national security concerns and the potential for unauthorized disclosure of classified project details and other information. For the non-defense agencies it is important to first have a protocol to identify the structure and availability of data about R&D within agency systems and how those data systems are structured.

Even if the agencies' data systems look promising and may have some of the necessary information about R&D funding and performance, there are often a number of policy and legal hurdles to convince agencies to share this information beyond its own organizations employees. This issue is not unique to those interested in administrative records data for R&D only but are a broad issue affecting the statistical community in general. For example, in 2014 the Office of Management and Budget released an official Memorandum for the Heads of Executive Departments and Agencies under M-14-06 *Guidance for Providing and Using Administrative Data for Statistical Purposes* to assure agency heads that they should seek to accommodate the statistical needs of the federal statistical system's use of administrative records. Further acceptance and recognition of statistical uses of administrative data still require a cultural shift in agencies regarding the willingness to share these data (accepting that they may not be perfect) beyond their own programmatic needs.

Surveys benefit from the human-based interpretation of what should or should not be reported from extracted administrative records. Differences in individual nuances of those definitions and interpretations can be readily vetted with the survey. However, such opportunities for review and interpretation do not exist when simply extracting raw data from a system of administrative records. Standardization of R&D definitions and classifications is necessary to ensure some level of accuracy that comes from the human elements of survey response are echoed in the use of administrative records.

Part II: Application and Findings from Pilot Effort to use Administrative Records for Measuring R&D

This section provides a brief overview of the process used to compile information from the administrative record files obtained from NSF, NIH, and NASA and results of the pilot efforts to use these data to replicate results from the FFS. NCSES initially proposed two methods for direct use of agency administrative data, this included data tagging and a clone-file approach.

Data Tagging Approach

Under the data tagging approach agencies would augment their existing data systems with NCSES R&D taxonomies for the FFS, using eXtensible Markup Language (XML)/XBRL. Agencies would not have to surrender control of the process, which was a concern under the clone file approach. Data tagging involves an increased focus on data standardization. It involves looking at ways to add tags to agencies' existing systems to facilitate use of the data by public policy makers.

In order to extract an accurate list of R&D projects from administrative data, NCSES needs to understand and be able to answer the following questions:

1. How is R&D defined?
2. Given the definition, how is it interpreted by agency staff?
3. Given the interpretation, how is each project coded in the agency's administrative database(s)?

These questions are especially important when trying to consistently classify R&D projects from different federal agencies, which may have different definitions, interpretations, and coding practices. Many projects fit more than one category, presenting a classification challenge. If individual project level data are available, however, tagging to multiple categories can provide flexibility in reporting.

The CNSTAT report noted that one impediment to the quality and timeliness of the FFS is the lack of alignment between each agency's internal reporting taxonomies and the taxonomies contained in the NCSES surveys. The misalignment required agencies to review their records manually and determine how to apportion their R&D transactions based on the NCSES taxonomies. The data tagging approach, would make use of an XBRL software module to take the R&D transaction records extracted from an agency's systems and map that agency's existing internal reporting taxonomies to the relevant NCSES survey taxonomy. In this way, the agency could automate the "translation" of the classification systems for that agency into the appropriate NCSES classifications.

The proposal to develop a module based on XBRL was based on a few factors:

1. XBRL is used primarily to enable organizations to map data records between two analogous taxonomies, and to tag records with the appropriate reporting taxonomy. This function is exactly the role envisioned for the ARP software module.

2. XBRL is the dominant standard for "tagging" financial reports from public corporations, due to its implementation by the Securities and Exchange Commission for regulatory filings. The dominant vendors of financial management and transaction reporting software, including Oracle and SAP, offered XBRL modules as part of their software systems. Therefore, it seemed safe to assume that government agencies, in their push for greater data and software standardization, would also begin adopting XBRL for financial reporting.

3. It was also assumed that each agency used internal classification systems for many of the concepts described in the NCSES survey taxonomies, such as type of recipient (performer), place of performance, funding purpose, etc. Agencies already included similar fields in their reports to those in the Federal Procurement Data System (FPDS). Therefore, with the possible exception of the Field of Research classification, there should be within each agency an internal classification that could be mapped to each of the NCSES survey taxonomies.

In designing the implementation phase of the project NCSES explored the development of a software module that would ingest transaction reports generated by the agency's Oracle or similar system directly into the module, crosswalk the classifications in each transaction record to the comparable NCSES classification, and produce a machine-readable file with records categorized in accordance to the NCSES survey taxonomies. The agency could then send that file directly to NCSES and its survey contractor for processing, calculations, and reporting. This conceptual/proposed process is shown in **Figure 1**.

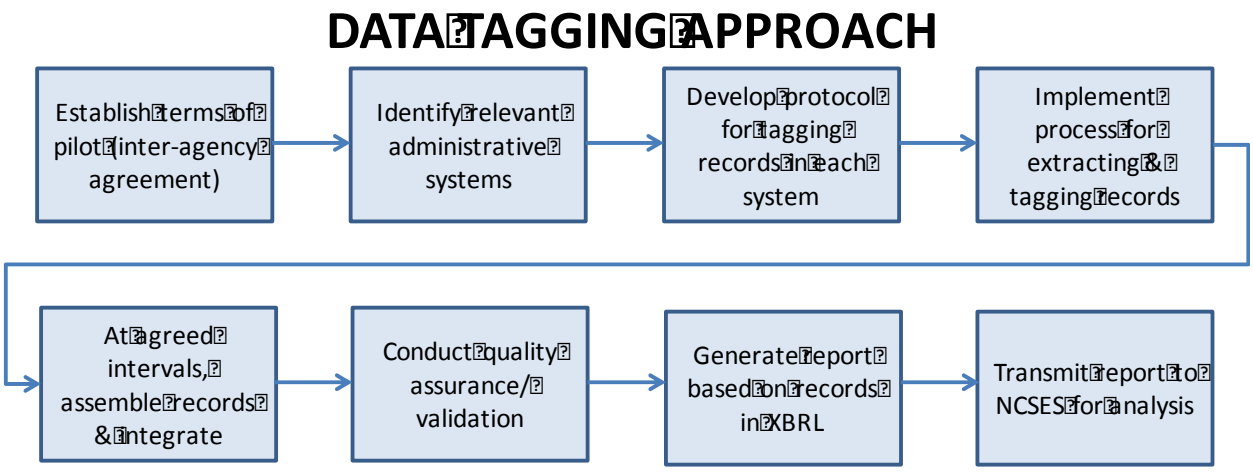

Figure 1. Imagined work flow of the data tagging approach.

During the interviews with agency representatives to obtain data files for the pilot NCSES discovered the following:

First, almost all agencies interviewed for the pilot had not yet begun to implement XBRL as part of their financial management software. Even for agencies using a comprehensive financial management system from a vendor such as Oracle, implementing XBRL compliance would entail modifications to that

system and additional costs. The ARP could not budget for the modification for other agency's core data systems – doing so could also be a violation of federal appropriations law.

Second, many agencies (including NIH, NASA, and components of the Department of Defense) assembled their responses to the NCSES surveys from multiple incompatible systems. There was rarely a single reporting system that could output a complete set of transaction records that could be processed into the survey response file. Therefore, any software module provided by NCSES would have to interface with multiple systems within each agency, greatly increasing the cost and complexity of an XBRL module.

Third, there was much less consistency than expected in the reporting classification schemes used by different agencies, and even inconsistencies in the classifications used with a particular agency for different transactions. Furthermore, many agencies' budget systems lacked a classification for the concepts contained in the NCSES survey. For example, few agencies recorded the place of performance for an R&D contract, relying instead on the mailing address of the recipient. Various agencies lacked classifications for the type of performer, the type of expenditure, and the purpose of funding. Responses to the NCSES surveys were generated using processes that were almost entirely manual, using heuristics and decisions on classification that were not easily replicable or necessarily consistent across years. In short, the agencies lacked internal taxonomies or database attributes that could be mapped to the NCSES survey taxonomies.

Taken together, these findings indicated that there was very little opportunity to leverage an agency's internal transaction record structures to generate the "tags" that would be affixed by the XBRL software module. As a result, even developing a pilot module for a single agency would be much costlier and time-consuming than initially envisioned by NCSES or presumed by the earlier CNSTAT panel. Also, there would be very little opportunity to achieve economies of scale, as the module used by each agency would have to be highly customized to that agency's systems and data structures. The review of agency records by the project team showed that the "translation" and mapping of taxonomies could not be easily automated, making the development of a new XBRL software module infeasible, leaving the clone-file approach as the only other alternative to make use of administrative data.

Clone File Approach

Under the clone file approach, information relevant to R&D funding activities would be extracted from agencies' central accounting systems and copies would be delivered to NCSES. NCSES would be responsible for mapping the supplied data to the survey variables and then used to create the statistics typically generated from the FFS responses. This approach could speed up the dissemination of data and provide consistency in the treatment and classification of R&D activities across agencies since NCSES would crosswalk all the data. NCSES would need data dictionaries as part of the administrative record deliverables in order to facilitate the crosswalk operation.

The analysis of agency transaction data sought to answer two questions:

1. Using the data provided by an agency, can all the DSTs for the FFS be reproduced? In other words, do the transaction records contain all the fields required to produce the DSTs?

2. What is the difference between the published DSTs for the 2012 FFS and the tables created from the agency-provided administrative data for Fiscal Year 2012? In other words, using

the data elements identified in #1 for FY 2012 data, how do the resulting DSTs compare to the published DSTs?

Preparation of “cloned" accounting records analysis involved examining transaction data from selected federal agencies as a general basis for either approach. This imagined process is shown in **Figure 2**.

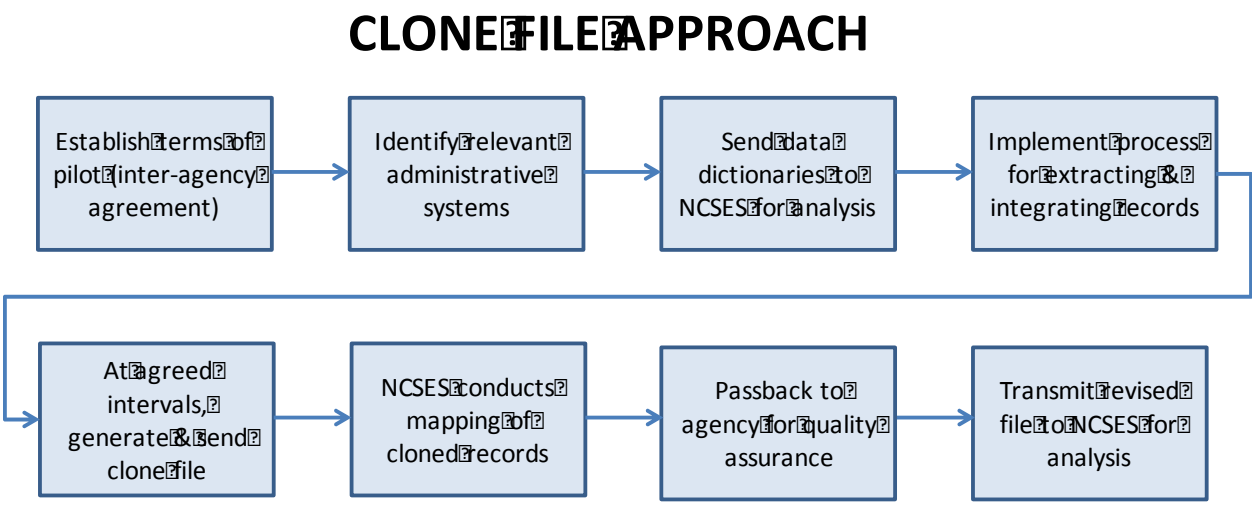

Figure 2. Imagined work flow of the clone file approach.

To support the development of the ARP pilot, but also to respond to concerns about the lack of documentation in the current agency responses, NCSES produced mappings of the taxonomies and data elements in specific R&D agencies and the classifications and variables collected in the FFS for NSF, NIH, and NASA. The agencies varied in the extent to which they could produce survey responses from administrative data systems.

NSF Proof-of-Concept

For NSF, nearly all variables in the survey were derived from data files extracted from the agency’s Financial Accounting System (FAS). A notable exception was DST Table 1: Outlays for R&D and R&D Plant. The NSF system is set up to track obligations. Outlays (for NSF this includes only transfers to extramural performers) are summed across projects and issued periodically, without any reference back to the projects associated with each transfer. Also, NSF uses a heuristic mechanism for determining the Field of Research for any given set of projects based on the funding Directorate and Program. This method, while systematic, produces an estimate of the actual allocation of R&D funds to fields, as certain programs may fund projects that cross multiple disciplines.

NCSES received two main files from the NSF budget office – a raw transaction records file and a processed records file in Microsoft Excel format. The raw transaction records were extracted from the NSF FAS using a SAS query executed by NSF budget office staff. The processed file was derived from the raw records file by mapping one or more FAS variables to the relevant variables in the FFS DSTs. For example, to assign each record to a Field of Research to report to the FFS, the NSF budget office applies an algorithm that analyzes the funding organization, funding program, and appropriation to assign a field.

The raw transaction records were processed to extract the data collected by the FFS, and then analyzed to produce facsimiles of the DSTs derived from the FFS. NCSES found that the data provided by NSF could be used to reproduce all the DSTs for the FFS. In addition, most of the DSTs that were produced matched the published FY 2012 DSTs generated from the 2012 survey process. Generally, the values generated from the transaction records were reasonably close to the figures contained in the FY 2012 DSTs. Discrepancies between the figures were in general much less than 1% of the DST amount and can most likely be attributed to rounding or an update in records between the original calculations and when the records were transfer to NCSES for the pilot.

Though the results of this experiment show that the DSTs can be reproduced using the NSF transaction records, updating the performer crosswalk and tagging of fields may improve accuracy. Though the error analysis revealed small random errors that probably occur due to miscoding, it does not appear that there are any insurmountable obstacles in using transaction records to reproduce the FFS DSTs from the NSF. However, NSF is not a very complex organization and all of its R&D is extramural in nature. NIH on the other hand is more complex in terms of its R&D activities and organizational structure so that the promise of using administrative records for measuring R&D from NSF proved short lived when trying the same approach with data from the NIH.

NIH Proof-of-Concept

Under the MOA NIH provided the following files from NIH for the ARP clone-file approach for FY 2012:

- An Excel file containing 58,017 transaction level records of grants, contracts, and intramural projects obtained by the Division of Statistical Analysis and Reporting (DSAR) in Office of Extramural Research (OER) from the IMPAC II database.
- An Excel file containing records of R&D plant expenditures extracted from the IMPAC II database.
- An Excel file containing Personnel Pay Cost and dollar values for basic research and applied research.

For this project, the raw records in the first two files listed above were used to extract the data collected by the FFS, and then analyzed to produce facsimiles of the DSTs where possible. NIH reports the same numbers for obligations and outlays, so only obligations were calculated. Since basic research, applied research, and development information was not included in the records, it was not possible to reproduce the FFS statistics with basic and applied research totals from the transaction records. However, the NIH Office of Budget (OB) calculates a high-level aggregate number for basic and applied research,[11] plant, and personnel costs. DSAR then uses the relative proportion of the IMPAC II records to calculate the money spent on the various FORD categories of R&D and R&D plant. Using human-provided analytical data in addition to the administrative records data, all of the DSTs can be reproduced using the NIH transaction records, but we could not produce them from the administrative data alone. Many figures reported to the FFS are generated through manual manipulation or compilation by NIH staff, and cannot be generated solely from administrative records. Relying solely on transaction records will generate values that are significantly different than reported values.

[11] NIH states that they do not do experimental development and so only classified their expenditures into basic research and applied research.

Although NCSES was able to calculate the same value for intramural research using NIH's methodology, this methodology relies on the assumption that all research not accounted for in extramural performer records in IMPAC II is, by default, intramural. NIH was not able to provide any information on administrative systems that might provide comprehensive records of intramural research, although IMPAC II provides some information for a limited number of intramural projects.

NCSES could only produce values for NIH-specific data in 65 of the 102 statistical tables produced for the FFS report. Some of these tables (14, 98, and 105) only contained Department of Health and Human Services (HHS) data in aggregate. We produced values for these tables since in theory if all agencies in HHS participated in this project then it is important to know that the NIH values could be obtained from the records, even though the numbers did not always match due to the missing data from other HHS agencies such as the Centers for Disease Control and Protection. NCSES was able to reproduce statistics in 15 of the 65 tables. Unfortunately NCSES was unable to reproduce exactly the numbers for Fields of R&D for all HHS R&D obligations. (We were, however, able to reproduce the numbers for fields of science and engineering for universities and colleges because those values came directly from the transaction records). For 54% of the records included, the FORD classification assigned by the NIH text-analytics process did not match the classification assigned manually during the clone file compilation process. Without access to the NIH proprietary system of assigning FORD, the only way to assign FORD classification is by a review of the project abstracts or descriptions. Many FORD classifications are not included in the NIH classification scheme. For the purposes of this project, however, this analysis suggests that assigning FORD tags is not straightforward. It is not possible to judge if an alternate semi-automated tagging process would assign projects more precisely than the current NIH process or how consistent this might be with the standard FORD categories. There is no formal crosswalk between the NIH proprietary disease classification algorithm and the FORD used by NCSES.

While NIH administrative systems for tracking extramural R&D spending appear to be complete, accurate, and comprehensive, a comparable level of data is not available for intramural R&D. Until that situation changes, it is impossible to recreate in full a substantial majority of DSTs currently generated from the NIH data submission to NSF without the additional inputs. However, it appears that a slightly different set of percentages were used for the FY 2012 FFS submission than what was originally provided to NCSES under the pilot. We were not able to exactly reproduce any of the breakdowns of basic research (all performers and for colleges and universities performers) despite applying the provided 53.2% value. Although most of the data were very close, which leads us to conclude that while only a few tables can be reproduced using only the transaction records, the inputs of the supplemental data may enable the production of all the DSTs that include NIH values.

NASA Proof-of-Concept

NASA was the third agency that NCSES worked with to pilot the use of administrative records under the clone file approach. NASA's data records were particularly problematic, as its main administrative data system lacks data elements indicating critical information such as the identity of R&D performers, or whether a given transaction is or is not related to R&D. Most records are associated with programs and organizations, not a functional area such as research. In general, NASA uses an internally developed heuristic to determine if a funded program is related to R&D, and if so, how it should be classified using the FORD taxonomy. It also uses that heuristic to identify the type of R&D (basic, applied, or experimental development). Most of the data reported by NASA to the FFS allocates the spending by

performer, type of R&D, or field of R&D. Therefore, none of the tables in the FFS can be completed using NASA administrative records alone.

Based on interviews with agency officials and through inspection of available data elements, NCSES developed a mapping to establish the relationships between NCSES survey taxonomies, the classifications in the agency transaction records corresponding to those taxonomies, and the specific data elements and systems used by the agencies to complete the FFS.

NASA provided a file with 18,648 raw records from their Budget Data Warehouse. Since program information is included in the records, the basic research, applied research, and development tables could be reproduced using the NASA-supplied heuristic. As noted above, the files were missing key variables such as Field of R&D and performer type. Very few DSTs could be reproduced using the NASA transaction records extracted from the Budget Data Warehouse. NCSES was able to calculate values for only nine DSTs of the 65 DSTs relevant to NASA's activities, and, of those nine, only six tables had data that matched the data reported to the FFS.[12]

We hypothesize that NASA projects could be classified by field of R&D using machine-learning and text analytics. To do so, they would need project abstracts. They could not locate a public source of this information for us to reference. Thus, the NASA administrative data, in their current state, are untenable for use by the FFS and the data submission still relies on extensive staff interpretation and data manipulation, which could not be re-created by NCSES to produce similar results.

Summary

Although the use of administrative records may appear to be theoretically advantageous and in many cases some of the data needed to complete survey requests for R&D funding even the most advanced agencies do not poses comprehensive administrative data regarding their own R&D activities. Administrative records data can only be used for statistical purposes if the levels of detail in the records themselves are more detailed than the information needed for the statistical variables that the records detail can be rolled up. In the case of R&D the level of detail that policy makers and data users actually need and use are more detailed than any information that might exists in agency records. Therefore any future use of administrative records for statistical purposes require efforts to standardize the information about the nature of R&D projects to make these data (or anything related to R&D funding or performance) of any comprehensive utility.

Part III – Use of Supplemental Source Data in Producing Federal R&D Funding Statistics

One approach noted by the CNSTAT panel was to take advantage of information already available in public data systems. The panel's report listed several of these public databases (e.g., Grants.gov, USAspending.gov, and Federal Procurement Data System) as potential sources for NCSES. The potential viability of these databases is supported by a variety of legislation, from the E-Government Act of 2002 to the Federal Funding Accountability and Transparency Act (FFATA) of 2006, which require improvements to the availability and accessibility of data about all organizations receiving federal funds.

[12] The topics of the six tables covered *Federal obligations and outlays for research and development and R&D Plant, by agency*; *Federal obligations and outlays for research and development, by agency*; *Federal obligations for research and development, by type of work, by agency*; *Federal obligations for research and development to foreign performers, by region; Federal obligations for basic research to foreign performers, by region*; and *Federal obligations for research and development, by state or location*.

As a result, the CNSTAT report makes a compelling case for NCSES to utilize these systems as part of its data collection effort on federal R&D:

> These improvements have the potential for reporting on federal R&D spending at the project level and associating fiscal year obligations with such attributes as performer, performing institution, and geographic location. These and other attributes could serve as the foundation for [NCSES] data collection efforts, which are also tied, in part, to fiscal year obligations and implicitly require the aggregation of project-level data into agency-wide data.
>
> Under these new laws, a supporting infrastructure is being developed across the government under the leadership of the Office of Management and Budget (OMB) that has the potential of improving the quality and timeliness of administrative records on government expenditures and, by doing so, to provide at least part of the data on R&D expenditures that are now extracted by means of surveys. Before these new initiatives, administrative records did not have the capacity to provide current or reliable information on federal R&D expenditures.[13]

The CNSTAT report also notes that NCSES should avoid any systematic changes to the basic survey methodology of imputation or estimation, instead, the path to modernization lay within making greater use of administrative records.

> [NCSES] could consider alternatives…[which] include reporting incomplete information earlier, providing a preliminary report with imputation for late respondents, or designing a simple schedule to the form that could be completed more easily and quickly as the basis for publishing a preliminary report.
>
> The panel reluctantly concludes that [NCSES] should stick with current practice. Each of these alternatives has pros and cons. Incomplete data might be misinterpreted, and preliminary totals would necessarily be on the low side because of missing data. Imputation for late respondents would introduce a new source of error and could make the data less accurate. Some data users hold that the potential for increased error due to omission or imputation is less desirable than more timely publication. [14]

Although CNSTAT suggested a number of advantages to using existing publicly available datasets to help move the collection of federal R&D spending from a survey-centric model, the panel's report noted a number of potential hurdles with these systems. "For example, the FFATA-enhanced administrative databases on contracts and grants are still maturing, and little work has yet been done with the major reporting agencies to set the basis for direct [NCSES] exploitation of their administrative records."[15] With regard to USAspending.gov the report noted the portal "allows users to generate detailed reports on external federal spending by performing institution, performer type, and geographic location. However, the website does not enable users to generate reports of federal spending by character of work or field of S&E, and it lacks information on intramural R&D."[16] Nonetheless, the reported noted

[13] National Research Council, Committee on National Statistics, *Data on Federal Research and Development Investments: A Pathway to Modernization*, National Academies Press: 2010. Pg.50.
[14] Ibid., pg.46.
[15] Ibid., pg.60.
[16] Ibid., pg.59.

that "NSF should develop the capacity for mining the standard and newly enriched government-wide contracts and awards databases to extract comprehensive information on R&D spending."[17]

In an effort to develop several demonstrated projects to test for the best method to move to a system based on administrative records[18] we sought to review the strengths and weaknesses of several federal-wide data systems identified in the report. The objective of this effort was to determine if the data from these federal-wide data systems could be used to supplement or supplant the FFS in its current form by providing similarly national uniform quality statistics of federal R&D funding and performance. The following sources, identified in the CNSTAT report, were examined for data on federal grants and contracts related to R&D in the summer of 2012:

- Federal Assistance Awards Data System (FAADS)
- Grants.gov
- Federal Audit Clearinghouse (Single Audit Program)
- Recovery.gov
- MAX Information System
- USAspending.gov
- Federal Procurement Data System (FPDS)

Although a number of these data systems had some of the variables that could be used by NCSES to collect and compile information for the FFS, there are a number of systematic short comings with these databases which render their use for official national statistical purposes problematic.

At best, some systems are more comprehensive and have more information than others. For instance, while the FPDS has information that would allow us to classify contracts by type of R&D through the FPDS codes for Basic Research, Applied Research, and Development, similar information needed to classify the type of R&D for grants in USAspending.gov is not available. As a result, we cannot compile a complete data set with the same variables applied to both grants and contracts consistently or uniformly.

At worst, information needed to properly compile the data is missing entirely. For example, data on agency intramural activities are not present in any of the systems reviewed. USAspending.gov, for instance, does not record data on federal employee salaries that would be needed to compile intramural R&D expenditures.

In many instances where the systems do have identifiers for type of external performer or type of R&D, the quality of these identifiers are inconsistent so that we cannot rely on them at face value to provide comprehensive nationally reliable statistics. For example, within FPDS universities are not always classified as Higher Education; in other instances for-profit corporation are classified as non-profits. Thus we cannot systematically rely on these identifiers as the sole source and would have to re-compile the information to reduce the presence of this type of non-sampling error.

[17] Ibid., pg.70.
[18] Ibid., pg.60, see Recommendation 4-3.

Project description information needed to identify the Fields of R&D is either, weak, inconsistent, or not available and does not allow for direct classification of projects. For example, while the FPDS has a Product and Services Code (PSC) that provides some information about the nature of the contracted services, they are more analogous to project outcomes (e.g., Defense Aircraft, or Coal) rather than general fields of science. Although the PSC may identify the project as coal related, for example, we have no information about what kind of research is being done with coal to assign the appropriate FOSE (e.g., Geological science or Atmospheric science). There is no easy solution to this problem, and is evidence of why researchers in the field of Science Policy are turning to alternative means, such as Topic Modeling and other automated language algorithms, to classify projects into various sciences and sub-disciplines, but these are not fully developed or necessarily capture the true intent of the project through word instances alone.

Based on comments from federal agency representatives at the NCSES sponsored inter-agency meeting to discuss *Options for Leveraging Administrative Records Data Collection on Federal R&D Activities*, there was broad consensus from federal agency representatives that these data systems would not be useful because they do not have the details NCSES is looking for. Specifically, USAspending.gov does not provide a mechanism to parse basic vs. applied research, development, or training. In addition, the system does not have enough granularity regarding the nature of the project funded that would meet the level of detail needed by NCSES.

Although we recognize the conceptual potential within these systems from a theoretic point of view, we also see an inherent contradiction in the CNSTAT panel's suggestion that these existing federal-wide data systems could provide enough information for NCSES to supplement or even supplant the FFS. The panel's report notes that "Reporting agencies are challenged by the fact that their internal records do not use the categories [of information] requested by [NCSES]."[19] It is this lack of detail specific to the statistical needs of NCSES that supports the CNSTAT panel's recommendations 4-1 and 4-2, which call for NCSES to develop data tags specific to R&D. However, since the data contained in these federal-wide databases originate with the agencies themselves, the same data limitations within agencies' existing systems limit the usefulness of the data once it is incorporated into other federal-wide systems beyond the agency of origin.

Federal Procurement Data System (FPDS)

Of all of these systems the Federal Procurement Data System (FPDS) showed the most potential for promise. NCSES obtained files from the FPDS public website; these files contain raw data from the submitting agencies. They are available to the public without special permission. For each agency, FPDS provides a compressed ZIP folder with a number of XML files detailing records from a given fiscal year. Each ZIP file contained two files relevant to this analysis: AWARDS (awarded contracts) and IDV (indefinite delivery vehicle contracts).[20] The AWARDS file contained 111,205 records from 59 agencies. The IDV file contained 6,475 records from 38 agencies. From these files, NCSES extracted only those

[19] Ibid., pg.17.

[20] The Indefinite Delivery Vehicle is a contract that has been awarded to one or more vendors to facilitate the delivery of supply and service orders. IDVs permit the issuance of orders for the performance of tasks or the delivery of supplies against prepositioned contracts and agreements.

records within PSC (Product Service Code)[21] category A (R&D) from the FY 2012 agency submissions. Therefore, the operating assumption was that all records in our AWARDS and IDV master files represented R&D activity.

In order to determine approximately how many of the downloaded records contained one or more detectable errors, NCSES created a sampling strategy that balanced the need to inspect records in every stratum and the necessity of being efficient. Overall, the AWARDS sample showed errors and inconsistencies pertinent to this analysis in 105 out of 298 records, and the IDV sample in 84 out of 220 records. If this had been a statistically significant sample the error rates would be 35% and 38%, respectively.

Notable issues included:

- Use of "FALSE" as a default value in data columns
- Use of generic titles by agencies for groups of foreign performers
- FPDS categories are not necessarily mutually exclusive; FPDS organizational Type response choices overlap TRUE/FALSE fields
- Neither FPDS nor the System for Award Management (SAM) has a unique tag for "Institution of Higher Education"
- No reliable tag to identify foreign performers
- All Federal Funded Research and Development Centers (FFRDC) records identified either in the AWARDS or IDV files were marked "FALSE" for "is Federally Funded Research And Development Corp"
- Errors in geographical coding
- Missing data

We do not know how much funding in the FFS represents grants and how much represents contracts. FPDS on the other hand covers only procurements (that is, only contracts and not grants). Additionally, intramural R&D spending from FFS will not be captured by FPDS.

Table 1 shows the FFS R&D extramural total obligations compared to "Obligated Amount" totals derived from FPDS data.

Table 1. Federal extramural obligations for research and development, by type of R&D, FY2011 (dollars in thousands)

[21] FPDS categorizes contracts by product or service codes. According to FPDS, "These product/service codes are used to record the products and services being purchased by the Federal Government. In many cases, a given contract/task order/purchase order will include more than one product and/or service. In such cases, the product or service code data element code should be selected based on the predominant product or service that is being purchased. For example, a contract for $1000 of lumber and $500 of pipe would be coded under 5510, Lumber & Related Wood Materials." From the FPDS Product and Service Codes Manual August 2015 Edition. https://www.acquisition.gov/sites/default/files/page_file_uploads/PSC%20Manual%20-%20Final%20-%209%20August%202015_0.pdf.

Data	Total R&D	Basic research	Applied research	Development
NCSES FFS Survey TOTAL	$100,346,100.0	$24,443,793.3	$20,916,512.8	$54,985,825.6
FPDS data analysis AWARDS	$51,835,532.0	$10,507,681.0	$12,769,585.2	$28,558,265.8
FPDS data analysis IDV	$207,680.0	$80,426.6	$28,422.9	$98,830.6
FPDS data analysis TOTAL	$52,043,212.0	$10,588,107.5	$12,798,008.1	$28,657,096.4

The overall totals for FPDS show only 52% of the total value of FFS data. At least some of this difference may be explained by the FPDS sums containing only procurements, as noted above. The proportional distribution of total funding under Basic Research, Applied Research, and Development is roughly similar, as shown in **Table 2**.

Table 2. Comparison of proportional distribution of FFS and FPDS totals for research and development, by type of R&D, FY2011

Data	Total R&D	Basic research	Applied research	Development
NCSES FFS Survey TOTAL	100%	24.4%	20.8%	54.8%
FPDS data analysis TOTAL	100%	20.3%	24.6%	55.1%

Nevertheless, when the totals are broken out by type of R&D for selected agencies for FFS and FPDS, neither the absolute totals nor the proportional distribution of activity show the same degree of agreement. In many cases, both are markedly different between FFS and FPDS. Although there were a few examples where the totals are close, for instance for NASA and the Department of Energy, these may be no more than chance matches, as the overall totals showed approximately 50% more activity in FFS than in FPDS.

The analysis showed that it is possible, by supplementing FPDS data with System for Award Management (SAM)[22] data, to reclassify performers according to the NCSES taxonomy used for FFS. However, it also showed that the totals derived from FPDS data by Type of R&D by Agency are often markedly different from the survey data. Without being able to examine the original contracts one cannot be sure if this is the result of PSC miscoding, missing data, or some other error. Another issue is that whereas FPDS covers only procurements, FFS includes both grants and contracts, and one cannot determine from FFS data how much funding represents grants and how much represents contracts. Additionally, intramural R&D spending from FFS will not be captured by FPDS. These factors suggest that FPDS data, at least currently, cannot reliably reproduce agency submissions to the FFS. In addition,

[22] The System for Award Management (SAM) took effect on July 30, 2012 and combined a series of Federal legacy databases and websites for the management of Federal procurements. These previous systems included the Central Contractor Registry (CCR), Federal Agency Registration (Fedreg), Online Representations and Certifications Application (ORCA), and the Excluded Parties List System (EPLS).

there are no corresponding data sets for grants data that might be used in conjunction with the FPDS to estimate total R&D funding by agency similar to that captured by the FFS.[23]

Summary

We can conclude that the use of existing federal-wide databases for data on R&D performance and funding are not a viable option for the Administrative Records Project to pursue prior to the implementation of some additional, and necessary, data standardization initiatives. This conclusion is based on the following observations:

- Systematic weaknesses in these data systems (notably for queries of intramural performers) and the inconsistent reliability of the identifiers available for queries of external performers,
- Lack of availability of data for identifying Type of R&D for grants,
- Inconsistent classification of Type of R&D contracts,
- Lack of consistent uniform data elements for identifying the FORD; and
- Recommendation from the CNSTAT panel that the FFS should remain a census and not be subject to imputation, estimation, or any form of sampling methodology.[24]

Therefore, any potential value in the Administrative Records Project remains with an examination of the feasibility of (i) compiling information directly from the agencies systems based on a review of each agency's data availability; and (ii) developing and using data tags, that would specifically capture relevant information and classifications of R&D needed (not only by NCSES, but also by OMB), and which are consistent with the Frascati Manual's guidelines.

Part IV – Developing Government-wide Standards

The ability to successfully compile agency data (from any source) may depend more on the development and application of data tags, which in turn, are dependent on ensuring data standardization across disparate agency systems in the first place. The CNSTAT report even makes this point in its discussion of reforms to the taxonomy of Fields of R&D, where it is noted: "[F]or purposes of collecting data on research and development statistics in a consistent manner across federal government agencies, it is necessary to establish a common taxonomy that will be useful to the largest number of data providers and users."[25]

The vital role of developing data standards for R&D was also highlighted in a report from the Organisation for Economic Cooperation and Development (OECD) titled: *Measuring Public Procurement of R&D and Innovation: Review of existing data sources, evidence and potential measurement methodologies*. The authors conducted a review of several public procurement data systems in different countries, the U.S. FPDS among them, to determine the viability of them for measuring innovation as well as their quality of capturing R&D data. Although the report demonstrated some potential analytical applications of using data from these systems, the authors noted that these systems will not be a reliable source of comparable statistics in the short or medium term.[26] The authors also noted that

[23] FPDS analysis done on FY 2012 data. The FPDS has since been undergoing upgrades for improved coverage and greater quality control.
[24] Ibid., pg. 46.
[25] Ibid., pg. 32.
[26] Appelt, Silvia, and Galindo-Rueda, Fernando; Organization for Economic Co-operation and Development, Directorate for Science, Technology, and Industry; Committee for Scientific and Technological Policy; Working Party of National Experts on

"comparable statistics produced in such [data] bases in the future are unlikely to be successful in the absence of common reporting standards for procurement actions."[27]

As such it would be prudent for any future interest in the application of administrative data and even in the interest of improving overall data quality on R&D data to invest future initiatives in the development of data reporting and auditing standards specific to R&D that are applicable to all federal agencies that fund or perform R&D; and to extend these standards to private industry application as well as the higher-education sector as well. Creating standards that can be applied by all federal agencies is itself a monumental undertaking, expanding this to other sectors critical to measuring R&D is even more so, and worthy of Blue Sky consideration for how this might be accomplished in the future.

In the U.S. both NCSES and OMB, are responsible for collecting comprehensive data on federal agency R&D funding. Both agencies use the same international definitions and guidelines for measuring R&D but need to rely on agencies to provide them with the R&D data, and must therefore bridge the gap between the formal Frascati classification standards and the actual output of agency systems, a process that ultimately creates inconsistencies in both the interpretation and application of definitions affecting the quality and consistency of the data. This challenge is compounded by a lack of continuity between individual agencies' financial accounting systems and other information systems containing project-level data about specific R&D projects' funding and performance. For example, some agencies must manually link data from their financial system on a specific contract or grant with a specific statement of work (which could vary from a short paragraph description to a five page abstract) to determine if the project in question should be reported as R&D to OMB and NCSES; and then also determine the appropriate field of R&D for the R&D project. In other words, for many agencies there are few if any automatic identifiers within their accounting systems specifically for R&D, and even fewer for further levels of disaggregation for items such as field of science. Those agencies that do have some inherent identifiers for R&D may not have identifiers for all components of R&D, and the ones they do have are unique to the agency and not necessarily comparable with those of another agency. These kinds of issues create opportunities for errors and inconsistency in R&D data reporting between and among federal agencies, and make the process of aggregating R&D data very time-consuming for both the agencies and the R&D data collectors. An example of the effect of these issues can be seen in **Table 3**, which shows how the federal government ends up reporting three different sets of figures describing R&D spending without certainty about which is more accurate. The gap of about $6 billion between the President's Budget and NCSES data, primarily a difference in Department of Defense figures, is narrower than it used to be, in part because of NCSES's efforts to identify sources of data discrepancies and resolve them.[28]

Table 3. Comparison of Federal government data collects of R&D in aggregate, by source, FY 2014

Science and Technology Indicators: "Measuring Public Procurement of R&D and Innovation: Review of existing data sources, evidence and potential measurement Methodologies" April 8, 2013 §101, Pg. 42.

[27] Ibid., §94, pg. 40.

[28] Known sources of discrepancies are: NCSES obligations vs. OMB budget authority, NCSES summer data collection cycle vs. OMB winter data collection cycle, and certain agencies' peculiar practices in reporting these data (DOD, NASA).

R&D Classification	Financial Report of the U.S. Government[29]	FY 2016 President's Budget[30]	NCSES's Federal Funds for Research and Development Report[31]
Total R&D	$123.9	$136.3	$130.3
Basic Research	$34.0	$32.2	$31.6
Applied Research	$28.1	$32.5	$31.3
Development	$61.8	$69.0	$67.4

In attempting to make use of administrative records of R&D for statistical purposes has led NCSES to consider the question of not only how administrative records might be improved but how we might use standardization practices in order to ensure more consistency and accuracy to the international standards for defining and measuring R&D. The question then becomes whether or not NCSES and OMB can create methods and processes to really improve the quality and consistency of the R&D data provided to them by federal agencies without a government-wide standard? While there are obvious capacity gaps at both the OMB and NCSES, there are a number of practical steps that can be taken and if successfully implemented over the long-term would also have implications to make administrative data truly useful. However, without a process to develop standards for identifying and tracking R&D activities, any use of administrative records for measuring R&D are inconsistent and incomplete. The absence of standardization not only impacts the utility of administrative records but of survey responses overall. Any standard setting initiatives should be designed not only to improve the consistency and quality of administrative records, but should also seek to imbed the human interactions necessary to interpret and classify data according to what is being requested. With any survey collection there is an inherent human activity to interpret and decide how the definition is to be applied and as a result how those data may be reported; there is no similar endeavor with administrative records data to ensure the same vetting of definitions and interpretation. Standardization needs to ensure the human-based interpretations of the definitions of R&D can be internalized as part of an agencies own internal controls and conveyed to their administrative records data.

Initially, a project to create new Object Class codes in the Federal budget and accounting standards to identify R&D specifically would make any future standardization relatively easy. Not only would improvements in federal R&D accounting and reporting benefit the R&D data collectors; this effort would also benefit the Nation in compiling statistics on the Gross Domestic Product (GDP). Bureau of Economic Analysis (BEA) recently revised the National Income and Product Accounts (NIPAs) so that R&D is now treated as an investment rather than an intermediate input in the cost of production. On the new basis of reporting, "[g]overnment R&D expenditures will be treated as investments regardless of whether the R&D is protected or made freely available to the public, because the provision of public

[29] The Financial Report of the U.S. Government, FY 201, pg. 227.

[30] Budget of the United States Governments, Fiscal year 2016, Analytical Perspectives, Chapter 19 Research and Development, Table 19-1.

[31] National Science Foundation, National Center for Science and Engineering Statistics. 2016 *Federal Funds for Research and Development: Fiscal Years 2014-16*. Detailed Statistical Tables.

services is part of the economic benefit generated by government R&D."[32] As a result of this change "R&D spending will be reclassified from consumption expenditures to gross investment…As a result, GDP will be boosted."[33] Thus, having consistent, high quality data on R&D is also important to the proper measurement of our national economic activity as well as international comparability based on provisions for the treatment of R&D in the System of National Accounts (SNA).[34] The challenge is that the Object Classes are created by Congress to serve their specific needs. Any change to these would require Congressional action and given existing priorities and concerns this is not likely to be something that can be practically pursued.

At this point a collaborative effort currently underway by NCSES and OMB to establish a Federal R&D Community-of-Practice (R&D COP) is the best opportunity available. This effort is looking at ways to address agency questions on the interpretation of how to apply Frascati principles to meet the constraints of existing systems and management practices. This will be done through the development of specific guidance to resolve agency specific questions and concerns. But it will also seek to work with agencies collaboratively to develop a set of R&D heuristics as a form of internal management controls for identifying and classifying R&D. The development of these standards should also benefit the surveys in addition to administrative records by helping to organize and categorize which activities, projects, and programs, specific to the agency, are consistent with the definition of R&D as per the Frascati Manual. In the summer of 2013 NCSES, OMB, and NASA staff worked collaboratively to develop a heuristic to improve the agency's ability to classify R&D activities from non-R&D activities. The results have reduced some of the manual manipulation of R&D administrative data by NASA staff when reporting to the FFS. But they have also lead to greater continuity in the R&D data reported to both NCSES and OMB. For example, for FY 20092 the total R&D NASA reported to OMB and NCSES showed a difference of over $5 billion. After successful development and implementation of this heuristic the difference was reduced to $50 million for FY 2013. So developing heuristics and internal controls can make a difference towards standardizing what is reported as R&D. But they need to be carefully crafted to ensure compliance with the Frascati Manual. These efforts need to be expanded and recommended as a best practice. Combining these heuristics as an agency's required internal management controls will help to internalize the need and process for ensuring data quality, but this too will take some volition to ensure it can be done given other priorities and political concerns.

So what are some options that we should consider for future improvements not only for administrative data but also the consistency and accuracy of these data about R&D? One option is to include the formal development of data tags to systematically and uniformly identify R&D and other important variables such as type of performer, type of R&D, and field of R&D. The challenge is how should we seek to ensure data tags for R&D are a required component of all federal reporting? In addition, how can these be expanded to address the needs of other sectors where R&D performance is critical. There are already options for how we might facilitate more electronic reporting of these data. The challenge is

[32] Survey of Current Business, *Preview of the 2013 Comprehensive Revision of the National Income and Products Accounts*. March 2013. pg. 15.
[33] Ibid., pg. 17.
[34] The System of National Accounts is internationally agreed standard to set recommendations on how countries should compile measurements of economic activity, enabling greater international comparisons of economic activity. http://unstats.un.org/unsd/nationalaccount/sna.asp.

how do we ensure the records themselves have sufficient means to capture specific and detailed information about R&D that policy makers and other require given the existing constraints?

Detailed questions for potential participants in the ARP Pilot

Item	Question(s)
	1 Policy on administrative records access/transfers
1.1	Who within your agency would need to give permission for your agency to be involved in the Administrative Records Pilot (ARP)? Can you provide us with those points of contact?
1.2	Please describe any legal, regulatory, or administrative restrictions on access to your agency's administrative records. If legal please cite legislation.
1.3	Are your data classified (requiring clearance) at the level of specific grants/contracts? If yes, is there a level of aggregation that can be transmitted in a data file as unclassified information? Also, If your data records are classified, would NCSES/SRI be able to access the data if they were transmitted and stored securely? Would any special permissions/credentials need to be issued?
1.4	Who would need to provide official approval for data files to be transmitted to NCSES/SRI? Do you have any information security and protection requirements that NCSES would need to implement when storing your records on NSF systems?
1.5	Do you believe that your agency is suitable for testing the data tagging approach or clone file approach as they were presented at the November 28th workshop, or both? Would a particular component of your agency be better for testing each approach?

1.6	What processes or routines are conducted by your agency to ensure the quality, coverage, and consistency of administrative records for R&D activity?
1.7	After NCSES/SRI applies a crosswalk and creates the survey data for the reference period, would your agency want to conduct a quality assurance review? If yes, who would do that review? (Follow-up is 2.1)
1.8	Is there a position or office that would be responsible for providing these records on an ongoing basis to NCSES for future survey cycles? (Follow-up is 3.2 and 3.3)
1.9	Are there any current plans to refresh or replace your current financial management or administrative records platform?
2 Identifying relevant records	
2.1	How does your agency's definition of "R&D" differ from the OMB A-11/NCSES definition? As such are there specific items that are excluded under one definition or the other, or are reported in different places?
2.2	How do you distinguish data records related to R&D activities from records for non R&D activities?
2.3	What object codes reference R&D? Do you use any other identifiers to specifically reference R&D activities from non-R&D activities?
2.4	Is information on the content of R&D activities stored in structured data formats, or as unstructured data (such as PDF files)?

2.5	Do you track funding of R&D activities at the project level? If so, do you assign a single, unique identifier to every project?
2.6	At what other levels of aggregation (program, sub-agency component, etc.) do you report data on R&D activities?
2.7	Do you store information on the content of R&D activities (project summaries, grant abstracts, project proposals) separately from financial or accounting data about those activities? If so, in what format(s)? Are the two linked? If so how? If not, what is done in order to associate the two for reporting to NCSES?
2.8	Do you store information on extramural R&D performers separately from records about their funded activities?
2.9	Do you collect and store reports submitted by external performers in electronic form? If so, in what format(s) do you store those reports? Are those reports correlated or cross-referenced with administrative data on the project funding
2.10	To what extent do you store information about sub-awardee performers? If not how do you currently address these in current reporting to NCSES?
2.11	Do data for prior fiscal-year final expenditures (outlays) have a different structure/IT environment from current fiscal-year obligations and next fiscal-year projections? Are the data for forecast years stored in a system separate from current or prior fiscal years?

	3 Collection/integration of data records
3.1	How many different systems collect, generate or store administrative and financial records on the following types of R&D activities? If possible, please name each system and describe its purpose. a. Intramural R&D projects b. R&D projects for external sponsors c. R&D funds to other government performers d. R&D funding for activities at FFRDCs e. Extramural R&D grants f. Extramural R&D contracts g. Independent R&D h. Information on extramural R&D performers i. R&D plant/infrastructure
3.2	For each of the above relevant systems, could you please provide the following: • Which organization/office within your agency is responsible for operating each system? • Who is charged with entering and updating data in each system? In what office and location is that person based (see item 1.8)? • Describe the scope of the records included in the administrative records for these systems. Please describe all variables even if they are not current used for reporting.Are there specific known sources of errors in each system? If so, how do you adjust for those errors (if at all)?

3.3	For each system, please provide the following details on its technical architecture: • What is the operating system/platform for the system? • What database software package or standard is used for managing the records? • What software is used to generate reports from that system? • Does the system interact with legacy systems to generate records, or is it a standalone system? If connected to legacy systems, what middleware technology is used to make calls on the legacy system? (For example, Web services, enterprise application integration, object request brokers, etc.) • What is the user interface to the system (e.g. Web browser, proprietary system, terminal access)?
3.4	Do any of your standard reporting processes require the integration of data from multiple systems? If so, to what degree have you been able to automate those processes (see items 2.4 and 2.7)? Also, what system architecture was used to enable data extraction and integration from multiple systems (e.g. Web services, EAI, other)?
3.5	Are there manual processing steps currently involved in preparing data to respond to the FFS/FSS? If so, where do those steps take place, and who is responsible for performing them (see item 2.7)?
3.6	How do you aggregate data on R&D activities at field offices and agency component offices? Are those data already centralized, or do you need to request data exports from those other offices?
3.7	How frequently do you compile financial reports for reporting or close-out?
3.8	How do you currently compile data to report to OMB for the MAX A-11 system (see item 2.1)?

3.9	Referencing your current reporting methods for the FFS, describe all manual adjustments you make to the administrative records before filing out the survey.
3.10	If you do not compile the data directly (e.g., field offices report to you), do you further edit the data or report exactly what you receive?
3.11	Describe any other processes that occur outside of any system that are an integral part of responding to the FFS. For example, if the split between Basic and Applied research is a percentage applied outside the records system, please discuss that here.
4 Ability to annotate/tag records (data tagging approach)	
4.1	What specific variables from each system are referenced for current reporting?
4.2	How do you determine and store the place of performance for R&D activities (both intramural and extramural)?
4.3	How do you integrate accounting and fields of science? Are these tagged in the same system or do hand calculations need to be made for reporting to the FFS?
4.4	Has your agency looked at the implementation of XBRL for any of its financial management systems (such as grants management, financial reporting, etc.)? If so has R&D been subject to such reporting?

4.5	What metadata exist to assist in cross walking agency records/objects to the FFS/FSS?
4.6	What existing classification systems are used to categorize or aggregate data records related to R&D activities?
4.7	Does your agency use a different or more specific chart-of-accounts from the US Government Standard General Ledger Chart of Accounts? If so, does it address or provide for specific classification of R&D activities? If so can you provide us a with copy?
5 Ability to interpret records and generate crosswalks	
5.1	How do you determine the appropriate classification of R&D activities using FFS and FSS taxonomies? In particular, how do you handle classification by Field of Science and Engineering and by Character of Work (basic, applied research, development) when not self evident?
5.2	Do you have taxonomies in your systems which are related to those NCSES taxonomies? Do you have a data dictionary? If so, can you provide us with a copy?
5.3	To what extent do your variables, classification systems, or record structures change each year? Describe such changes from one year to the next. How are multi-year project addressed when there are changes each year?
5.4	Has your agency conducted an evaluation on the use of the National Information Exchange Model (NIEM)? • If so for what programmatic reporting/use was it evaluated? • Were any R&D datasets or potential R&D variables impacted or included in the evaluation?

	• What did the agency conclude with regard to NIEM adoption? • Who/which office is the best point-of-contact regarding the evaluation or use of NIEM?
5.5	If there were formal specifications on what the tags should be and when to apply them would your agency have resources to apply such tags as part of your on-going business process? • For existing projects? • For new projects as they start? • Could they be applied to out-year projections?
5.6	Would a multi-way crosswalk work better for your data (i.e., a crosswalk for each field rather than for each record)?
6 Ability to export/report records	
6.1	How easy is it to execute customized data runs from your system?
6.2	In what formats do you export data for producing reports?
6.3	What organization (IT support, budget, other) is responsible for creating data runs for each of your administrative record systems?
6.4	Do you involve contractors in generating data runs or reports?

6.5	Does your current system have special data runs to call the data needed for the FFS? If not, could this potentially be built in?
6.6	Do you currently publish or otherwise export accounting data (for reporting purposes, etc.) in XML? Is the conversion to XML conducted in batch mode, or is this a continuous process?
6.7	What is the file structure of the administrative record system outputs that would be sent to NCSES as a clone file?
6.8	If multiple systems will be accessed, will the data be merged or in separate output files?