Research Performance Progress Report (RPPR) Data Dictionary

Guide for Using the Data Dictonary

August 2012

Introduction

The purpose of this document is to provide grant-making agencies with an overview of the Research Performance Progress Report (RPPR) initiative and the RPPR Data Dictionary.

The document is organized into three sections:

- Background information about the RPPR initiative and the Data Dictionary development process
- A technical overview of the RPPR Data Dictionary, including its organization, and design and technical considerations.
- An appendix that contains the RPPR data dictionary.

Background on the RPPR

The RPPR is an initiative of the Research Business Models (RBM) Subcommittee of the Committee on Science (CoS), a committee of the National Science and Technology Council (NSTC). The Working Group began this effort in May of 2004, with the objective of creating an alternative to the Performance Progress Report (PPR) that would service the needs of the research community. Leveraging the National Science Foundation progress report format as its model, the Working Group developed a standard format for reporting performance on federally funded research projects.

The RPPR policy letter was signed on April 21, 2010, and established a standard RPPR format for use by agencies and awarding offices that support research-related activities. Per the policy letter, each agency was required to post an implementation plan by January 21, 2011, that addressed the agency's decision to use either the RPPR paper or electronic format, as well as the agency's anticipated implementation date. Detailed information and background on the RPPR policy and agency implementation plans can be found on the RPPR Website.

Following the issuance of the policy letter, NSF hosted a meeting with research agencies and the Office of Management and Budget (OMB) to discuss the RPPR implementation. The meeting provided participants with historical context and background of the RPPR, information around the process for clearance of agency implementation packages, answers to specific questions. At that meeting, the research agencies agreed to collaborate on developing a standard government-wide RPPR data dictionary. This will help ensure consistent government-wide implementation at the data level, resulting in improved data integrity and consistency, and fostering inter-agency collaboration and information sharing.

RPPR Data Dictionary Development Process

Representatives from research agencies interested in using the RPPR met to develop a government-wide RPPR data dictionary. The National Science Foundation's (NSF's) developed a draft RPPR data dictionary as a starting point and sent it agency representatives for initial comment. Once comments were collected, organized, and analyzed, four cross-agency meetings were held to discuss and resolve comments and gain consensus on a draft RPPR data dictionary.

RPPR Data Dictionary Overview

As grants management becomes more automated, improved process efficiencies that are built into systems will continue to subsume current labor-intensive manual and paper-based processes. This conversion to systems-based processing relies on data in electronic formats. Once stored on databases, the data expands an agency's abilities to transform it into crucial, insightful, meaningful, and useful

information. It allows agencies to efficiently enter, display, query, report, consolidate, and share information.

To store data in a structured manner where it can easily be accessed and used efficiently, a data dictionary is required. Analogous to any word found in an English dictionary, there are similar components found in a data dictionary, including: the data's name, its definition, its field length (i.e. how many characters), and how it is used.

In addition to providing structured data for an agency's system, a data dictionary can also be used to share information between systems, whether these systems are within an agency or external to it. However, this requires that both the data creating and receiving systems use the same data dictionary so that they can "speak the same language" to each other. For example, verbal and written communication between two people would be impossible if one used an English while the other used a Russian dictionary as the basis for communication. Similarly, a mismatch in data dictionaries between two systems also results in an inability to share information. Thus, the solution, whether it is two people or two systems that want to share information, is an agreement to use a common dictionary.

The RPPR data dictionary provides a common data platform for systems and provides agencies an efficient means to share and receive information. In addition, the dictionary is useful for those agencies that build interface capabilities with grantee systems. This would provide grantees with the ability to create the RPPRs in their system, and submit them to the agencies via an interface. While originally conceived for research agency usage, the data dictionary could also be used by other agencies that have a need for collecting similar performance data.

The RPPR data dictionary defines all the data elements that may be collected per the RPPR policy. It does not include any additional agency- or program-specific reporting components that an agency may require. As specified by the RPPR policy letter, agencies may develop additional reporting components; "however, to maintain maximum uniformity, agencies should minimize the degree to which they supplement the standard categories." Any agency-specific requirements will require additional OMB review and clearance under the Paperwork Reduction Act (PRA). Each agency must work with OMB to clear its implementation of the RPPR whether it be through an online form, downloadable fillable form, paper form, or some other mechanism. The data dictionary helps agencies with the technical implementation of its RPPR reporting mechanism by providing a detailed data model.

Data Dictionary Organization

The data dictionary, in an MS Excel spreadsheet, is organized into sections and subsections that align with the RPPR policy guidance. The spreadsheet contains a tab for each RPPR reporting component.

- Cover Page
    i. Cover Page
- Mandatory Reporting Category
    ii. Accomplishments
- Optional Reporting Category
    iii. Products
    iv. Participants
    v. Impact
    vi. Changes-Problems
    vii. Special Reporting Requirement
    viii. Budgetary Information
    ix. Demographics

Each tab has seven columns that provide descriptions and attributes for each data element. The columns are:

1. Field # - For organizational ease, each data element has a field number, starting with number one for the first data element and following a numeric succession for all subsequent data elements.

2. Data Element Name - Each data element has a name that logically defines its meaning, and attempts were made to ensure that the names are self-explanatory.

3. Data Type - Data types define the kind of data attributed to each element. Additionally, each data element contains its unique field length.

4. Validation - Some data elements contain a set of valid values that users can select. Agencies will need to define these values in their databases and implement a mechanism for users to select the appropriate values. For example, a drop-down list showing valid values could be used.

5. Definition - The definition provides a description of each data element.

6. Standards Sources - Every effort was made to leverage current accepted standards for each data element, pertaining to an element's field length and definition. When a current standard was applied, the source was annotated. Some data elements fell under the "New Standard" designation. This designation was created for two possible scenarios: a data element deviated from a known standard; a data element that has no known current standard.

7. Notes - This field may contain suggestions about additional instructions that may be needed. It also contains the rationale of the work group for deviating from existing government-wide standards.

Design and Technical Considerations

This section provides a high-level description of design and technical considerations to aid agencies in developing and implementing the RPPR. However, it is assumed that agencies will have complete discretion over detailed RPPR designs that satisfy their unique functional and business needs. This includes, but is not limited to:

- Graphical user interface
- Data editing and processing logic
- Databases and data schema
- Interface designs (e.g. system to system interfaces)
- Systems access and security

The data dictionary does not prescribe how an agency should design, develop, and implement the RPPR. This allows agencies complete freedom to adopt an approach that meet their unique business needs. When developing and integrating the RPPR in their grants systems, agencies can benefit from the following design and technical considerations:

1. Graphical User Interface and System Functionality

The data dictionary is intended to only provide an inventory of all data elements (and their attributes) for the RPPR. While agencies are required to follow the RPPR organizational structure, they are free to design the graphical user interface and system functionality to support the RPPR data collection process. Agencies are not required to collect all the data included in the data dictionary. Because agencies have different systems, it is important to allow this independence so that agencies can build

their RPPR functionality that logically dovetails with the look, feel, and workflow of their particular systems.

2. Data Editing and Processing

An important component of the system design is how the RPPR will be collected, edited, and processed. The primary consideration should be user access and system security to the RPPR workflow. Because the system collects RPPR data from external users (i.e. grantees and possibly other agencies via system to system interfaces), it must have a public facing component. This requires a higher level of security than an internal system encapsulated by agency's strong network security.

With a strong system security mechanism in place, agencies will need to consider what RPPR data will be entered by the user, pre-populated, and interfaced with other systems. It is assumed that at this point that agencies have already designed the database to accommodate all the fields. An important consideration will be minimizing grantee administrative burden. Some fields in the RPPR can probably be pre-populated with data that agencies already have in their grants management systems such as grantee information found in the RPPR cover page. In addition, agencies may pre-populate data by interfacing with other systems such as invention information from iEdison. The RPPR represents a complete set of data elements that can be collected by the agencies. However, agencies can choose to collect fewer elements to conform to business requirements. This is most relevant in the Products section (e.g. citations) where agencies may choose not to collect all the data elements.

In addition to identifying the data that can be pre-populated, agencies will need to define user-entered data. This data will probably include both mandatory and optional fields; as an example "Middle Name". Agency systems should have data editing and processing logic to ensure that all required data is collected, edited for completeness and errors, and validated. The data dictionary itself does not determine what data are required or optional.

3. System-to-System Interface Development

Agencies may want to publish the data dictionary and interface specifications/services should they decide to make system-to-system interface possible. Grantees with a large number of government grants that have the system and infrastructure to build an interface to an agency's grants system will benefit from increased reporting efficiency. This interface makes it possible for grantees to create reports from their own system and submit reports electronically via an interface. Government systems can be built with interface editing capabilities to allow grantees to make the appropriate edits before final file acceptance. This ensures that information captured by the government systems have passed the required edits, resulting in consistent data integrity.

4. Database Development/Specific Data Dictionary Considerations

Data Types

Because different agencies use various databases that result in unique naming conventions for data types, the data dictionary uses generic data type labels for each element. During the design process, agencies can translate these generic data types into specific data types unique to their databases. For example, CHARACTER (generic) can be translated to VARCHAR for SQL databases. The data dictionary contains the following generic data types:

- CHARACTER – includes both letters and numbers
- NUMERIC – numbers only and used when mathematical calculations are possible
- DATE – universal data format
- Field Lengths

To arrive at the maximum field length for each data element, the "largest common denominator" approach was adopted when there was not an accepted government-wide data standard that could be adopted. This approach took every research agency's field length requirements for each data element and used the longest required field length as the standard. Agencies can, of course, collect data with shorter field lengths, but the schema should accommodate the agreed upon largest field length.

The "largest common denominator" approach benefits both the grantee and agency. Grantees with system-to-system interface with agency systems can use one specification to satisfy each agency's field length requirements without the possibility of truncating data. Likewise, agencies choosing to share information with each other can create and receive complete data streams without truncation.

Pre-defined Field Values
Some data elements have pre-defined values that grantees must pick without the option of typing in the answer. For example, the data element "Report Term or Frequency (annual, semi-annual, quarterly, other)" has four pre-defined values: Annual, Semi-Annual, Quarterly, and Other. Agencies will have the freedom to design how this information is displayed and entered by the grantee; as examples: pull-down menus or radio buttons.

To ensure that pre-defined values are consistent across agencies, each value has a database value; for example: Annual = 1, Semi-Annual = 2, Quarterly = 3, and Other = 0 (Zero). Simplified database values ensure a more accurate methodology for inter-agency data sharing. In the example above, hard coding spelling/wording errors could occur if the database used the named values (e.g. Annual). The result of this would be mismatched and inaccurate data. However, if simple numerics were used to correspond to the named values, coding errors are less likely. Grantees would not see the numeric values; instead they would see the more understandable named values.

Multiple Database Rows
The data dictionary contains several sub-categories where it is possible to have more than one valid answer. For example, the "PD/PI: Name" sub-category contains eight data elements relating to the PD/PI information such as "Last Name, First Name, E-mail Address, etc.". If the grantee needs to report multiple PD/PIs, agency systems will need to make it possible for grantees to enter multiple PD/PI information. In this case, the database will be required to accommodate multiple rows for each data element. Agency policy and processes will dictate how many entries a grantee can enter for those sub-categories that can potentially have multiple answers.

Attachments
Most narrative fields have a character limit of 8000 characters. Agencies may want to give grantees the option of uploading attachments, and agency systems would need to implement uploading capabilities. This is a design question that is not dictated by the data dictionary.

"Nothing to Report" and "No Change" Checkboxes
Should grantees have nothing to report on some fields (mostly narrative fields), agencies may want to implement "Nothing to Report" checkboxes or similar mechanisms. Mandating an answer for every field, even if it is a "Nothing to Report", ensures reporting completeness and avoids the possibility that grantees may skip a field.

Appendix: Data Dictionary

See document titled" RPPR Data Dictionary_Aug 2012