National Institute of
Environmental Health Sciences

To tag or not to tag: Learning to achieve information balance

Helena L. Kennedy, Jennifer B. Collins, Kristianna G. Pettibone, Christina H. Drew
Division of Extramural Research and Training; National Institute of Environmental Health Sciences

Abstract

At what point does the cost of data collection outweigh the data it produces?

How and when should federal agencies decide?

At the National Institute of Environmental Health Sciences (NIEHS), we are trying to answer those questions and learn how to streamline our grant tagging process. NIEHS funds approximately 1,200 active grants. In 2012, we developed the Grants Coding Database (GCDB) to help categorize and tag our portfolio of grants with up to 600 attributes. While many tags are instrumental in responding to requests received from leadership, Congress, etc., others have never been used or requested.

This poster:
- Gives an overview of the GCDB;
- Describes results from the analysis; discusses some of the lessons learned, and
- Presents possible future directions, including the use of new tools -- such as natural language processing -- to help streamline the process.

Three NIEHS Strategies for Coding Grants Applications

Program Class Code (PCC): The PCC is an eight-digit code that forms the basis for assigning grant applications to NIH Program Officers, Grants Management Specialists, and to particular areas of science supported by NIEHS. The code must be quickly and accurately assigned to ensure that Staff are aware of their work-load and responsibilities in a timely manner. The current PCC process is manual.

Research, Condition, and Disease Categorization (RCDC): RCDC is a computerized reporting process applied across all the National Institutes of Health (NIH). Each grant application is automatically assigned a "fingerprint" comprised of the topics present in the title, abstract and specific aims. The RCDC codes are used in reporting NIH funding levels for 265 research, condition, and disease categories to Congress.

Grants Coding Database (GCDB): The GCDB contains detailed scientific and programmatic information about NIEHS grants that are manually coded according to a series of rules established over time. The database allows for consistent and efficient data entry, review and reporting. It also allows staff to generate reports and visuals describing particular portfolio characteristics.

NIEHS Funded Research Grants are Systematically coded in the GCDB

NIEHS started coding its funded grants in 2012 as a way to be proactive in answering questions about the portfolio. The GCDB contains over 3,500 competing NIEHS grants active in fiscal year (FY) 2011 and grants that have been funded since then through the present. New grants are added in batches, typically after each Council meeting, three times per year.

The grant coding system is primarily designed for grant information retrieval. Coding information allows users to rapidly generate lists of grants based on specific criteria that go beyond blind word searches and broad categorization.

Many coded items capture ideas not defined by keywords, such as novel experimental approaches, windows of susceptibility, or broader applications of the work. Coding data can also help users characterize pre-defined grant sets, such as grants from a particular RFA, or look across the entire portfolio at one or more key ideas.

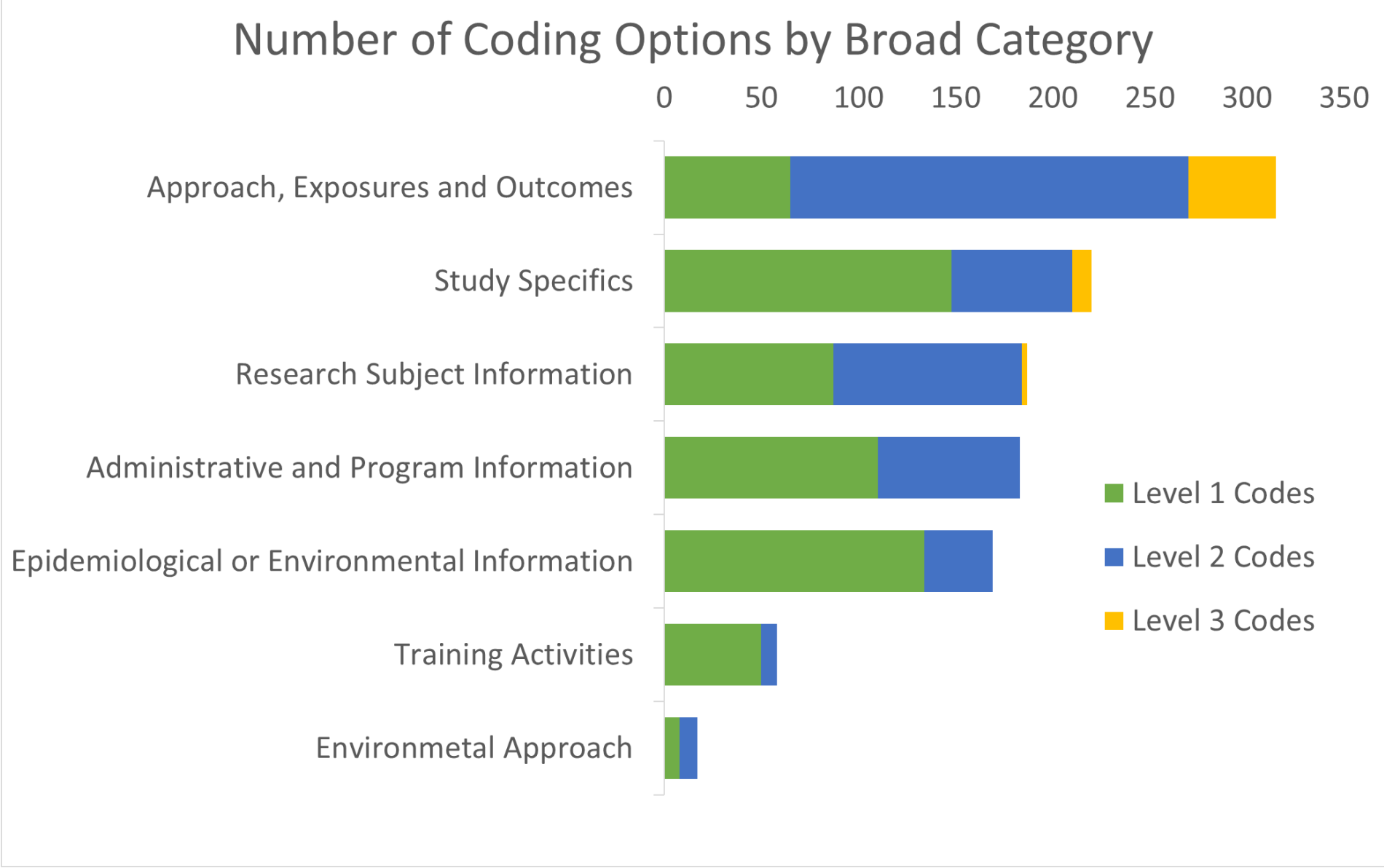
Information Captured in the GCDB Covers a Wide Variety of Topics and Granularity

- Coders can select up to 600 different tags across 7 broad categories of codes
- Each category can have up to three levels of coding; with each level getting more specific:

Approach, Exposures and Outcomes
Air Pollution
Particulates
Organic Farm Dust

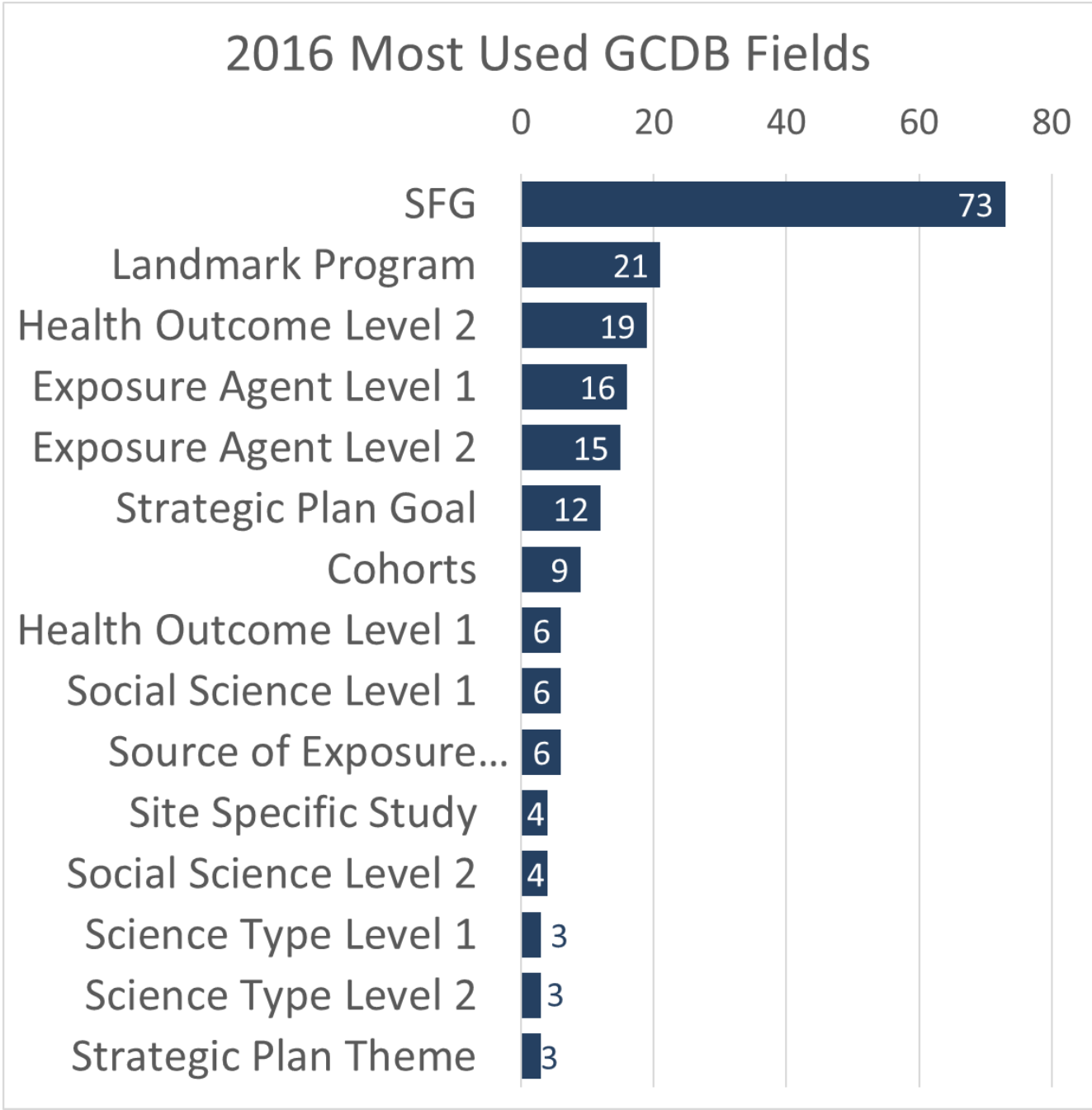
Study Specifics
Mechanistic Studies
Endocrine
Estrogen/Androgen

Research Subject Information
Prenatal/Pre-conception Exposure
Prenatal Exposure
Second Trimester



GCDB Codes are used to Identify Grant Portfolios

- We use GCDB data to answer a wide variety of portfolio questions.
- The GCDB codes complement the standard grant information (RFA, FY, principal investigator, grant number, etc.) when identifying portfolios.

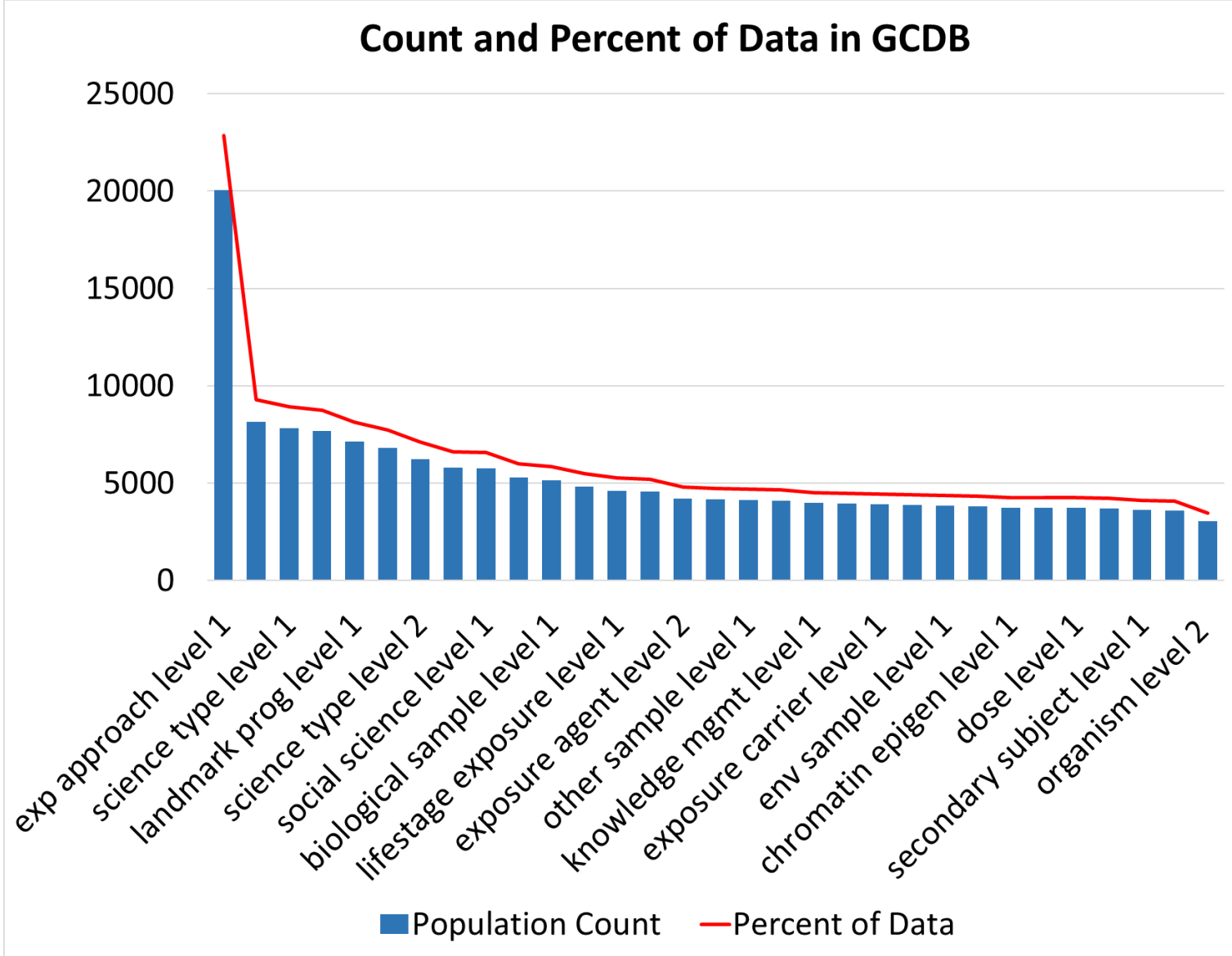


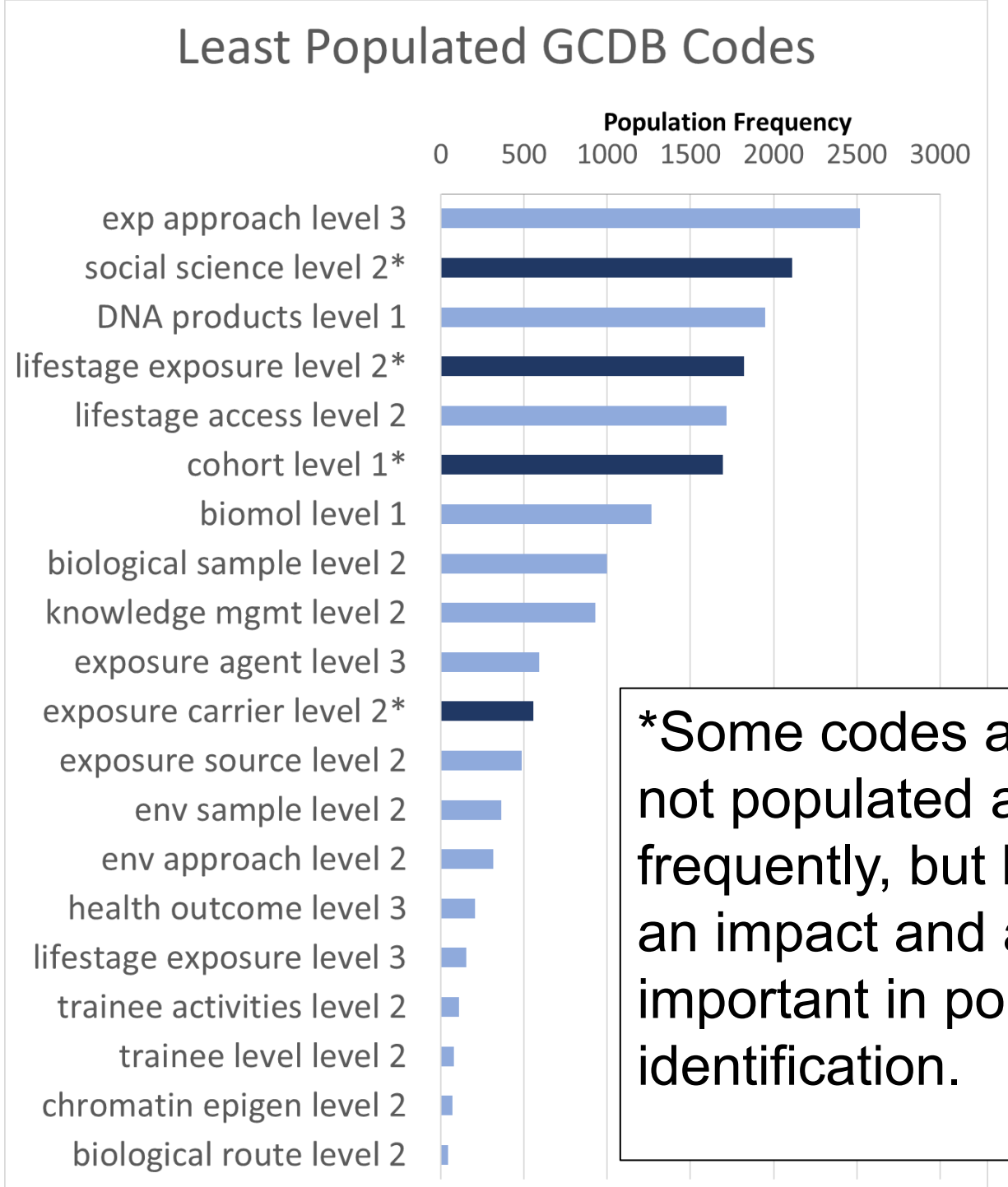
Portfolio Uses
- Presentation slide sets
- Bibliographies
- Data requests from external offices

In 2016, the most used GCDB field was "Science Focus Group (SFG)." These are important topics defined by a group of NIEHS staff that come together to select the topic and determine what kinds of grants should be included. To date there are 43 different SFGs, such as Children's Environmental Health and Air Pollution.

How Could We Streamline the Coding Process?

The 30 most populated fields in the GCDB represent 90% of the data in system.



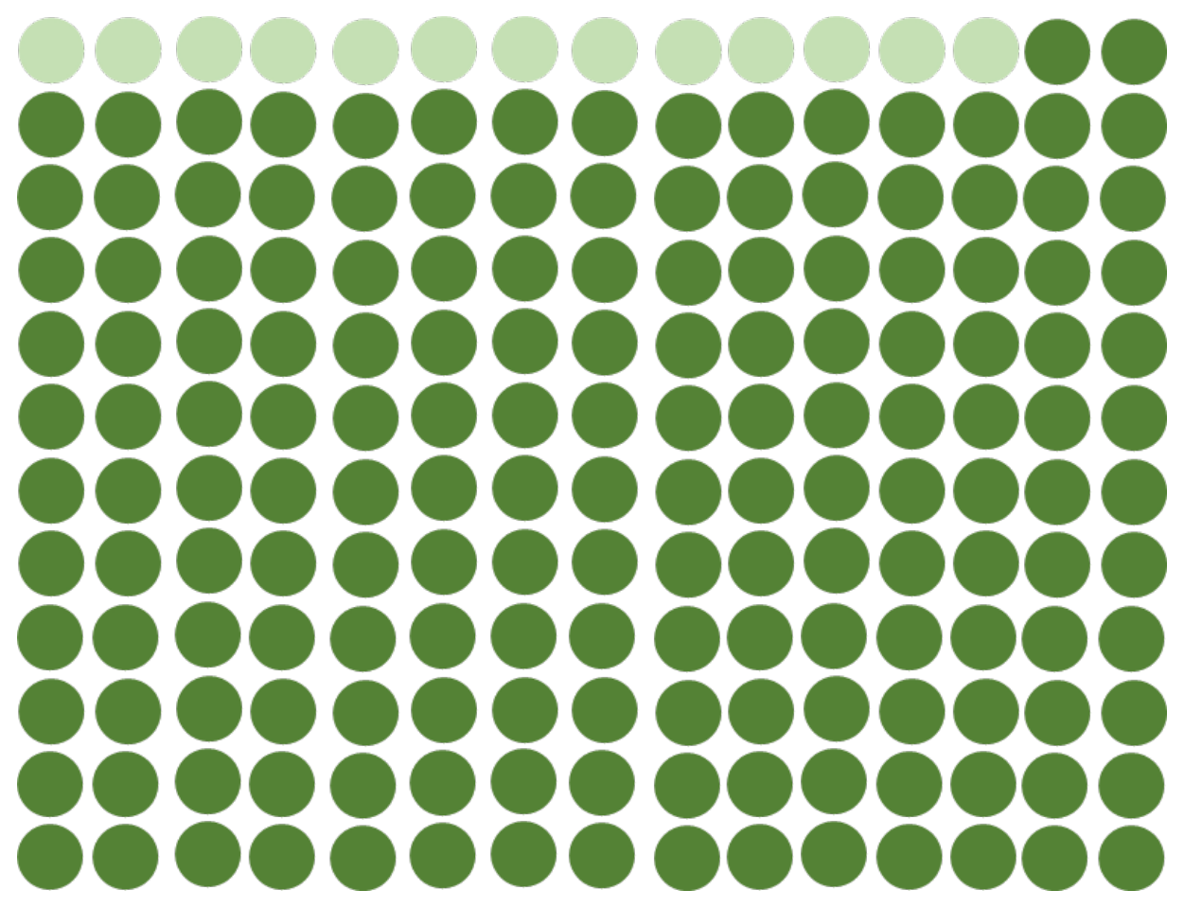


*Some codes are not populated as frequently, but have an impact and are important in portfolio identification.

Leveraging RCDC Codes for the GCDB

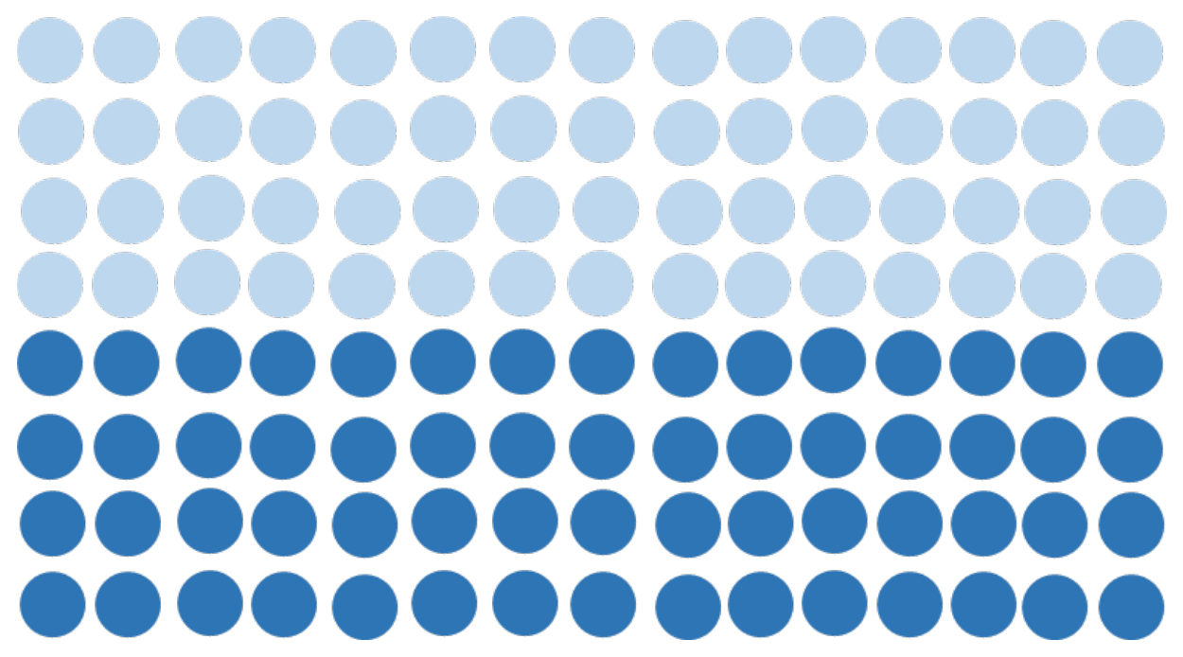
What would we lose if RCDC was used to identify portfolios?

- The majority of the RCDC "fingerprints" match the highest level of GCDB coding ("level 1"). If we used only RCDC codes, we could lose specificity required for portfolio identification, typically found in "level 2" codes.



13 of the 180 Exposures

and



60 of the 120 Health Outcomes

have RCDC fingerprints

Findings

- We only use a handful of codes to identify portfolios.
- We need to use a combination of impact and frequency of use to decide what to stop coding.
- We would potentially lose ~75% of the key exposures and/or health outcomes if RCDC was used exclusively for portfolio identification.

Next Steps

- Work with RCDC team to develop more environmental health-relevant exposure and health outcome fingerprints to the automated system.
- Continue discussions with staff regarding usefulness of codes and what can be streamlined based on data.
- Research new tools such as natural language processing to streamline coding.
- Present findings and recommendations to leadership.

Acknowledgements

The authors acknowledge and thank DERT staff for their critical feedback in both the evaluation and suggestions for future directions.

We would also like to acknowledge and thank the SaRPS team at MDB, Inc. for their help in extracting data from the GCDB.