# Process Evaluation of the NIH Research Categorization Program

The influence of the informatics core on categorization quality, effort, and acceptability

**National Institutes of Health**

**Office of Extramural Research, Division of Planning and Evaluation**

**November 18, 2014**

# Evaluation Summary

## Introduction

The research categorization program provides categorization services for the National Institutes of Health (NIH) research portfolio to ensure transparent and consistent categorical spending reports to the NIH and beyond. The research categorization program is supported by the Research Condition and Disease Categorization (RCDC) system, which categorizes the NIH research portfolio using the thesaurus-based natural language processing system Collexis™© in conjunction with a weighted search term list called a fingerprint. The process evaluation described herein was conducted to compare the quality of the categorization results produced by the current fingerprint approach of the RCDC system with two alternative categorization technologies. The evaluation was conducted over the course of the Fiscal Year (FY) 2013 categorization cycle by performing the usual categorization activities using the two alternative categorization technologies in parallel with the RCDC system. The overarching goals of the evaluation were to assess the quality of the categories produced, the effort involved, and the perceptions of stakeholders regarding the three different categorization approaches.

## Summary of Evaluation Findings

**The evaluation was conducted to address three primary questions:**

1. What is the quality of the categories created using different methods?
2. What is the relative effort involved in creating/refining categories for the different methods?
3. What are users’ perceptions of the different methods?

**The evaluation results support the following conclusions:**

* The statistical machine-learning method (Recommind™©) produced results comparable in quality to those of the RCDC fingerprint method. All three categorization methods, RCDC fingerprint, RCDC Best-fit, and Recommind, categorized NIH’s research projects reasonably well (>70% accuracy for most categories).
* The two machine-learning approaches involved the least overall effort for the 15 evaluation categories. However, for categories in fundamental maintenance status, which comprised 72% of the reported spending categories produced by fingerprint in FY 2013, the least effort was expended on categorization using the RCDC fingerprint system. For the general maintenance categories, which comprised 26% of the reported spending categories in FY 2013, the least effort was expended using the Best Fit system, and the RCDC fingerprint system involved less effort than the Recommind system. It is unclear whether efficiency would be gained through experience and process refinement if a statistical and/or machine-learning process were adopted for ongoing categorization.
* Based on survey results, the machine-learning systems were less “transparent” than RCDC in terms of users’ understanding and being able to explain how projects are categorized.
* During the development of new categories, significant time is devoted to research on the scientific topic of the category, and this time would be required regardless of the categorization method.
* The low proportion of projects that were validated by IC Experts was a complicating factor for the analysis of categorization results.
* Although IC Experts devote time to RCDC activities, the overall cost associated with their participation in the NIH categorization program appears to be relatively small for an endeavor of this scale (estimated to be $316,345 for validation and session time in FY 2013). Overall, IC Experts reported spending approximately 4.87 minutes per validity comment.

## Recommendations

The results of the process evaluation indicate that both the thesaurus based and statistical categorization approaches produced acceptable categorization results. Estimates of precision and recall were comparable for the categories produced by the RCDC fingerprints and Recommind’s statistical machine learning system. Furthermore, the standalone application provided by the Recommind system possessed features that, if incorporated into the RCDC user interface, would improve the user experience. However, NIH’s categorization program goals are unique within the field of semantic analysis, i.e., using categorization for budget reporting. Therefore, it is unlikely that any commercial off-the-shelf product would provide a fully functional categorization solution for the NIH research categorization program. Incorporation of a statistical machine-learning system into RCDC would involve significant modifications to the current trans-NIH categorization program that would involve a multi-year period of transition and change management. Because the Recommind system was not unequivocally better than the current Collexis system as a whole, and text mining technologies are still developing and changing rapidly, wholesale adoption of a single categorization method as a replacement for the Collexis thesaurus-based approach is not recommended at this time. Rather, the RCDC governance group recommends the following actions:

1. The machine-learning functionality currently available within RCDC should be further developed to work with mathematical as well as thesaurus-based categorization engines to support ongoing experimentation and testing of new categorization technologies as they emerge.
2. The NIH Research Categorization Program should leverage the expertise that already exists at NIH to establish a collaborative, trans-NIH team of experts in semantic analysis. This team should actively investigate technologies that could be developed to produce a hybrid categorization system that combines the best features of thesaurus-based and statistically-based text mining capabilities. Text mining expertise was identified in Institutes, Centers and Offices across NIH, including but not limited to NLM, CIT, DPCPSI and the ICs.
3. The trans-NIH team of experts should establish criteria against which to compare categorization methods and report their progress and findings to the RCDC Governance Committee.
4. The Process Evaluation should be repeated in three to four years to ensure that these recommendations are implemented to the advantage of the NIH Research Categorization Program and to conduct an *in situ* evaluation of the alternatives nominated by the trans-NIH team of experts. The evaluation and its results should be reported to the RCDC Governance Committee.