Characterizing the Investments in NLM’s Grant Portfolio

***RoseMary Hedberg, MLIS Associate Fellow, 2012-2013 Project Leader: Valerie Florance, PhD Director for Extramural Programs***

**Final Report January 2012**

# TABLE OF CONTENTS

Acknowledgements 3

Abstract 4

Introduction 5

Methodology 7

Term-Mapping with Word Banks 7

Semantic Natural Language Processing 10

Second Coding Test 12

Results 13

Discussion and Recommendations 16

References 18

Appendices 19

Appendix A: Original Grant Code Definitions 19

Appendix B: Specific Long Range Plan Goals 21

Appendix C: Initial Coding Test Results 22

Appendix D: Bioinformatics Expanded MeSH Tree Structure 23

Appendix E: Cosine Similarity Results 26

Appendix F: Second Coding Test Results 28

Appendix G: Second Coding Test Results (Revised) 30

# ACKNOWLEDGEMENTS

I would like to thank Dr. Valerie Florance for the investment of her time and informatics expertise on this project. I would also like to thank the Extramural Programs staff for their vital feedback: Dr. Hua-Chuan Sim, Dr. Alan VanBiervliet, Dr. Jane Ye, and Ebony Hughes.

Thank you to Dr. Kathel Dunn, Coordinator of the Associate Fellowship Program, and my Preceptor Jen Jentsch for serving as both sounding boards and voices of reason.

I would especially like to thank Dr. Lan Aronson and Dr. Antonio Jimeno for their individual contributions and for volunteering their time and insight.

Finally, I would like to thank my fellow Fellows Diana Almader-Douglas, Karen Gutzman, and Kevin Read, for providing me with an immeasurable amount of support.

# ABSTRACT

## OBJECTIVE:

Using the current informatics research grant codes established by NLM’s Extramural Programs division, explore the expansion and refinement of grant category definitions to improve their usability and allow for automatic coding for analysis of funding trends.

## METHOD:

The primary approach to expanding the code definitions and improving their usability involved approaching the project from a “data mining/term mapping” perspective by creating a “word bank” of sorts that would associate with each of the six grant codes. The word banks would be created by using terms and concepts from several sources, including the MeSH Thesaurus, CRISP Thesaurus, and term definitions created by the American Medical Informatics Association, the Centers for Disease Control, and the National Center for Biomedical Ontology.

The second approach involved utilizing semantic natural language processing, automatic tagging, and cosine similarity algorithms to match grants to their most similar corresponding codes through machine-learned analysis of the terms and concepts within grant titles.

## RESULTS:

The MeSH Thesaurus and the other sources used proved to be ineffective at creating valid grant category definitions because their corresponding entries in each database were either too brief, too vague, or did not exist based on the novelty of the various fields of informatics. Creating word banks would not be a suitable approach.

Based on a combination of incomplete titles within the sampled test grants and a lack of coded grants by which to predict against, the cosine similarity test produced weak, unreliable results.

## DISCUSSION:

The combination of poor coding test results, vague category definitions, and ineffective data mining and natural language processing experiments resulted in in-depth discussion amongst the Extramural Programs team. This discussion led to unanimously agreed upon refinement and redefinition of the grant categories, thus achieving the main objective without necessitating a complete redesign of the categorization process.

# INTRODUCTION

Extramural Programs (EP), one of six major divisions of the National Library of Medicine (NLM), operates a $44 million grant program for informatics research and development and informatics research training. It is the only component of NLM authorized to award federal grants. These grants support research and development in biomedical informatics, specifically in improving storage, retrieval, access, management, and use of biomedical information. There are several types of grant programs available from NLM, including research grants, resources grants, training support, career deployment support, and grants for small businesses that cover various aspects of informatics. For example, NLM research grants focus on research and development in bioinformatics, and resource grants focus on optimizing the management and use of health-related information.

At the suggestion of the NLM Board of Regents, Dr. Valerie Florance and the EP staff developed a method for presenting any grant program expenditures as an investment portfolio. In the past, the Board was told how many grants were awarded in the different grant categories (research, resource, etc.), but the categories were not given any thematic view. To improve upon the grant presentation effectiveness, Dr. Florance developed a six-topic coding scheme so that the awarded grants could be seen in terms of subject area rather than simply grant mechanisms (see Table 1). These six codes were chosen based on NLM’s university-based biomedical informatics research training program focus areas, as they are meant to cover the range of informatics. The original working definitions of the six codes are found in Appendix A. She wanted to create a set of high-level codes that could be used across time that could also be linked to the Library’s Long Range Plan (LRP) goals (see Appendix B), so EP could clearly communicate how its grants awarding process and decisions fit into the scope and direction of NLM’s research future1.

| Investment Category | Code | Link to NLM LRP Goals |
| --- | --- | --- |
| Bioinformatics | BB | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.2, 3.3 |
| Health Care | HC | Goal 3: Rec 3.1, 3.2, 3.3 |
| User Sciences | US | Goal 2: Rec 2.1, 2.4, 2.5 |
| Basic Informatics | BI | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.3 |
| Public Health Informatics | PH | Goal 3: Rec 3.2, 3.3 |
| Translational Informatics | TR | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.2, 3.3 |

Table 1: EP's Six-Topic Coding Scheme with Corresponding Long Range Plan Goals

Previously, Dr. Florance coded each funded grant herself. However last year, in an effort to shift coding responsibilities to the program officers, she and the officers performed a coding test of 25 sample funded grants. Each of the three program officers and herself used the grants’ titles, abstracts, and (if necessary) specific aims to code the items against her created category definitions.

The initial 25 item coding test resulted in poor inter-rater reliability. Only 12 of the 25 items (48%) had full or near agreement across all coders (see Appendix C). The test showed that refinement of the code definitions would be needed. Also, in order for the investment coding to be useful in portfolio assessment, the unfunded grant applications would need to be coded in the same manner as the funded grants. This would assist EP, as well as the Board of Regents, in more fully understanding what was and was not selected for funding, along with possibly extrapolating trend data.

## Project Objectives and Goals

The initial components of the proposed project included: 1) Using the existing investment codes, assigning codes to all non-awarded research grants, 2008-12 (excluding ARRA); 2) Expanding the code definitions to improve their usability; 3) Creating a coding manual for use by EP staff that includes the code definitions and examples of grants that fit each definition; 4) Recommending a graphic or visualization approach for presenting portfolio investment graphically in public talks.

# METHODOLOGY

## TERM-MAPPING WITH WORD BANKS

The primary approach to expanding the code definitions and improving their usability involved manually creating an ontology, here meaning an explicit specification of a conceptualization2. This meant approaching this project from a “data mining/term mapping” perspective. In this approach, a “word bank” of sorts would be established and associated with each of the six grant codes, so that whenever one of the terms in the word bank would be found in the grant proposal‘s title or abstract, the term (and subsequently, the grant) would map to one of the codes. This would guarantee consistency, reflected in the representational vocabulary, with respect to each individual grant.

After initial discussion and approval of this approach as a valid possibility, the next step involved seeking out similar programs and conceptual frameworks from other grant-awarding institutions and building the word banks based on valid terms, similar to the method employed by the extramural research division of the National Institutes of Health (NIH).

### RePORT and RCDC

The NIH maintains the Research Portfolio Online Reporting Tools (RePORT) as a way to provide access to reports, data, and analyses of NIH-funded research activities, as well as information on NIH expenditures and the results of NIH-supported research. The RePORT site also supports Research, Condition, and Disease Categorization (RCDC), which is a computerized reporting process NIH uses at the end of each fiscal year to categorize its medical research funding in 223 research, condition, and disease categories. While the RCDC does not currently record funding for NLM EP informatics research grants, the categorization process employed would possibly be applicable for the purposes of the EP grant project. Benefits of this approach include consistency, reliability, and detail.

Much of the consistency and reliability stem from solid category definitions, which are series of concepts most relevant to a particular category. These concepts are chosen from the RCDC thesaurus, which consists of more than 180,000 biomedical concepts and synonyms derived from several sources, including:

* NLM’s Medical Subject Headings (MeSH) Thesaurus
* Computer Retrieval of Information on Scientific Projects (CRISP) Thesaurus
* National Cancer Institute’s Thesaurus
* Jablonski’s Dictionary
* Other specific types of concepts from NIH Institutes and Centers

These same sources will be used together for the EP grant coding project in order to create comprehensive and universal definitions and word banks for the six existing informatics research codes.

### MeSH Thesaurus

Once the appropriate resources were found, it came time to define the codes. This began with searching for the code terms in the MeSH Thesaurus through the MeSH Browser, which met with results in a broad range of quality. For example, searching for *Bioinformatics* in the MeSH Thesaurus led to the MeSH Heading *Computational Biology*. The resulting tree structure displayed as such:

**Computational Biology**

**MeSH Heading:** Computational Biology

**Scope:** A field of biology concerned with the development of techniques for the collection and manipulation of biological data, and the use of such data to make biological discoveries or prediction. This field encompasses all computational methods and theories applicable to molecular biology and areas of computer-based techniques for solving biological problems including manipulation of models and datasets.

**Entry Terms:** Bio-Informatics; Bioinformatics; Biology, Computational; Computational Molecular Biology; Molecular Biology, Computational

**See Also:** Medical Informatics  
**Tree Structure:**

Natural Science Disciplines

Biological Science Disciplines

Biology

**Computational Biology**

Genomics+

Epigenomics

Glycomics

HapMap Project

Human Genome Project

Nutrigenomics

Proteomics

Metabolomics

Systems Biology

Figure 1: MeSH Tree Structure for ‘Computational Biology’

In this approach, each of the terms under the umbrella of *Computational Biology* was similarly expanded and defined. For example, in the case of *Epigenomics*:

**Epigenomics**

**MeSH Heading:** Epigenomics

**Scope**: The systematic study of the global gene expression changes due to epigenetic processes and not due to DNA base sequence changes.

**Entry Term:** Epigenetics

*Epigenetic Processes*

*MeSH Heading:* Epigenesis, Genetic

*Scope*: A genetic process by which the adult organism is realized via mechanisms that lead to the restriction in the possible fates of cells, eventually leading to their differentiated state. Mechanisms involved cause heritable changes to cells without changes to DNA sequence such as DNA methylation, histone modification, DNA replication timing, nucleosome positioning, and heterochromization which result in selective gene expression or repression.

*See Also*: DNA Methylation; Morphogenesis

Figure 2: Expanded MeSH Tree Structure for ‘Epigenomics’

When appropriate, certain terms or concepts within individual definitions necessitated expansion for clarification, such as with the preceding *Epigenetic Processes*. These expanded tree structures would then be used to fill corresponding terms in the word banks, in such a way that if, for example, the terms *gene expression*, *epigenetic processes or heterochromization* were ever to be found in a grant’s title, abstract, or specific aims, that grant would map to *Bioinformatics*, as per the tree structure seen above. The fully expanded and defined tree structure for *Bioinformatics* can be found in Appendix D.

### Other Resources

The American Medical Informatics Association (AMIA), the premier group of health care providers, informatics researchers, and information professionals in biomedicine and science, created its own definitions of the major domains of informatics. Out of the areas they research, they define translational bioinformatics thusly:

**Translational Bioinformatics** is the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations.3

AMIA also created a working definition for consumer health informatics:

**Consumer Health Informatics** is the field devoted to informatics from multiple consumer or patient views. These include patient-focused informatics, health literacy and consumer education. The focus is on information structures and processes that empower consumers to manage their own health--for example health information literacy, consumer-friendly language, personal health records, and Internet-based strategies and resources. The shift in this view of informatics analyses consumers' needs for information; studies and implements methods for making information accessible to consumers; and models and integrates consumers' preferences into health information systems.4

The Centers for Disease Control and Prevention (CDC) also understand the importance of informatics research, and they have their own definition of the concept. There are two focus areas of particular interest to the CDC and the National Program of Cancer Registries (NPCR): public health informatics and cancer surveillance informatics.

In order to provide an exhaustive search of all available resources, the search for universal definitions expanded to the National Center for Biomedical Ontology (NCBO) and BioPortal, its biomedical ontologies application. BioPortal allows users to browse the NCBO library of ontologies, search for terms across multiple ontologies, and browse mappings between terms in different ontologies.

## SEMANTIC NATURAL LANGUAGE PROCESSING

The RCDC categorization process inspired another approach for finding a way to systematically code the informatics grants. RCDC is a concept identification system, useful in information retrieval and extraction, data mining, and classification and categorization. The system reads text and assigns an appropriate concept. NLM’s Lister Hill National Center for Biomedical Communications (LHNCBC) developed its own concept identification system called MetaMap. MetaMap, released in 1994, “is a program that automatically maps text to concepts in the Metathesaurus, and a Medical Text Indexer System, which produces automate MeSH indexing of meeting abstracts and suggests appropriate MeSH headings to NLM’s MEDLINE indexers”.5

MetaMap is part of the Semantic Knowledge Representation project, which is concerned with providing both reliable and effective management of the information that is encoded in natural language texts by way of natural language processing. This project works to develop programs that provide semantic representations of biomedical texts by using NLM resources, specifically the Unified Medical Language System (UMLS) and its knowledge sources, the Metathesaurus, Semantic Network, and SPECIALIST Lexicon. For reference, the Metathesaurus consists of over 2 million concepts and synonyms, and the SPECIALIST lexicon contains more than 40,000 entries.

Natural language processing systems are increasingly being used in support of other computer programs, especially as a way to gain access to information inherent in large amounts of text. The traditional approach of natural language processing necessitated complete analysis of every sentence. However, there have been some new approaches developed in syntactic analysis of context-free grammars. As Dr. Thomas Rindflesch of the LHNCBC Cognitive Science Branch states: “There is a growing realization that effective natural language processing requires increased amounts of lexical (especially semantic) information”.6

Even so, there is the possibility of using automatic tagging programs, which have been found to typically be around 95% accurate and contribute significantly to efficiency.

### Automatic Tagging and Statistical Machine Learning

An automatic tagging experiment was proposed and run for this project using the titles of a selection of unfunded grants and the established code definitions. The terms appearing in the definitions and the grant titles were collected and their frequencies estimated. The terms were chosen by identifying spaces and other characters that would indicate boundaries between terms. Additionally, other words from a stop words list, such as prepositions and articles, were removed as they have no meaning in context.

A cosine similarity algorithm examined the overlap between a summary built from the labeled grants from the first coding test with the selection of unlabeled, unfunded grants. The algorithm assigned to each title lines to the most similar grant code from the Initial Coding Test Results (Appendix C) by comparing the words from the titles to terms collected for each code.

Cosine similarity is a commonly-used text mining measure that compares vectors representing the text within documents. The dimensions of the vectors are words, so similar documents will have vectors that are close to each other in space. The similarity measures the cosine of theta, the angle between two vectors.

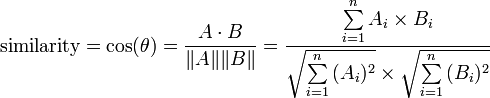


Figure 3: Formula for Determining Cosine Similarity between Vectors

In the above formula, A and B represent respective attribute vectors of the two documents undergoing similarity comparison. Resulting similarity ranges from 0 to 1. If two documents are identical, they will have corresponding identical vectors. Thus, there will be an angle of 0 difference between them, which then gives a cosine of 1. Dissimilar documents will have a larger angle between the two vectors, and as such their resulting cosine measure will be closer to 0. Simply put, the closer the cosine similarity is to 1, the stronger the relationship is between two documents. When applied in the context of this experiment, the grant titles displayed varying strengths of similarity between the codes:

Uncovering and Reducing Health Literacy Barriers to Tobacco Cessation

US|0.2654243731222853

PH|0.22478935866813277

HC|0.1074430618700507

BI|0.03479445003196102

TR|0.02933573244244292

BB|0.0

Figure 4: Example of Cosine Similarity Result Extracted from Grant Title

When the cosine similarity measure is applied, the code with the highest number (and thusly most similar relationship) for each title would then be assigned to that particular grant. Under these parameters, the preceding grant example “Uncovering and Reducing Health Literacy Barriers to Tobacco Cessation” would be labeled *US Consumer Health Informatics*. More results from the cosine similarity experiment are included in Appendix E and discussed further in the Results section.

## SECOND CODING TEST

After the team discussed the initial findings from the term-mapping and natural language processing approaches, it was determined that these would not be appropriate avenues for this course of study. The focus then shifted towards the need to determine whether the initial code definitions were valid and reliable. A second coding test would need to be run. The team kept the established definitions and retested with a different set of grants. This time, the grants would be tested by Dr. Florance and her team of three program officers, with the addition of a naïve but educated user to serve as another analyst. The test results would then be collected, analyzed, and discussed, with the intention of sending the results through the same processing program that had been run previously, in order to see if the correlations were similar, thus providing testing validity.

# RESULTS

## TERM-MAPPING WITH WORD BANKS

### RePORT and RCDC

Each of the resources was evaluated for possible use in the grant coding project, as they seemed the most likely to produce favorable, consistent results based on the similar end goals of both the RCDC and the EP project. Two of the sources were immediately discarded as possible avenues. The CRISP Thesaurus, developed by NIH for use in the CRISP database of research projects funded by the US Public Health Service, contains over 8,000 preferred terms grouped hierarchically into 11 domains. However, the CRISP database was last updated in 2006 and has since been superseded by the NIH RePORT Expenditures and Results (RePORTER) query tool, which allows for the search of a repository of both intramural and extramural NIH-funded research projects. Jablonski’s Dictionary also proved to be of little use. Upon further investigation, it was found that it is actually titled Jablonski’s Dictionary of Medical Acronyms and Abbreviations, and even if it was useful in building the RCDC thesaurus, it would not be needed for coding informatics research grants in this context.

### MeSH Thesaurus

Searching the MeSH Browser worked well for expanding the code definition of *Bioinformatics*, and it seemed that the term-mapping word bank method would prove to be a valid and comprehensive measure. However, when the same process was applied to the other terms, the results were not as robust. Some terms were in the Thesaurus, but their definitions were far too brief. Such was the case with *Public Health Informatics* and *Clinical Informatics*:

**Public Health Informatics**

**MeSH Heading**: Public Health Informatics

**Scope**: The systematic application of information and computer sciences to public health practice, research, and learning.

Figure 5: MeSH Tree Structure for ‘Public Health Informatics’

**Clinical Informatics**

**MeSH Heading:** Medical Informatics

**Scope**: The field of information science concerned with the analysis and dissemination of medical data through the application of computers to various aspects of health care and medicine.

Figure 6: MeSH Tree Structure for ‘Clinical Informatics’

Some of the terms were far too broad, as with *Information Science*:

**Information Science**

**MeSH Heading**: Information Science

**Scope**: The field of knowledge, theory, and technology dealing with the collection of facts and figures, and the processes and methods involved in their manipulation, storage, dissemination, publication, and retrieval. It includes the fields of COMMUNICATION; PUBLISHING: LIBRARY SCIENCE; and informatics.

**Tree Structure**:

**Information Science**

Book Collecting

Chronology as Topic

Classification

Communication

. . .

Informatics

Information Centers

Information Management

Information Services

. . .

Medical Informatics

Pattern Recognition, Automated

Publishing

Systems Analysis

Figure 7: MeSH Tree Structure for ‘Information Science’

Two of the terms, *Translational Bioinformatics* and *Consumer Health Informatics*, are such new fields of informatics research that they did not have any corresponding terms in the MeSH Thesaurus. This does not mean that definitions do not exist for those terms, but that finding those definitions would necessitate seeking further resources.

### Other Resources

The AMIA definitions serve the purpose of providing a broad understanding of the specific fields. However, they lack specificity. Therefore, it is virtually impossible to create a comprehensive working definition or word bank based on these definitions. Similar issues were found when searching for topic definitions from other resources, such as the CDC’s NCPR and the NCBO’s BioPortal. The NPCR definitions are specific, but they, just like those provided from AMIA, are not universal enough to serve as comprehensive definitions for the purpose of this project. Since individual institutes’ definitions seemed not to serve the project needs well enough, it was hoped that a wider search of ontologies available from the BioPortal’s library would produce better results, but that was not the case. Based on these issues, it was clear that a word bank would not be feasible, and a new approach would be needed.

## NATURAL LANGUAGE PROCESSING

While the concept of using natural language processing to create machine-learned maps from grant titles to codes was theoretically sound, it did not provide reliable results. Cosine similarity was weak in all of the tested grants, with the highest observed score being 0.31 in *User Sciences* for ‘Promoting Asthma Self Care in Inner City Patients with Low Health Literacy’. Appendix E lists a selection of the similarity test results.

## SECOND CODING TEST

Results from the second coding test were initially poor at best, for a number of reasons. Upon discussion, it was found that a few of the testers created a seventh code category: *CRI* for *Clinical Research Informatics*. Additionally, not all of the testers were clear on which two-letter combinations to use for code designations. For example, one tester used *CI* for *Clinical Informatics* rather than *HC*. Another tester transposed the codes for *Clinical Informatics* and *Consumer Health Informatics*. Any grant coded as *Consumer Health Informatics* was labeled with *CH* rather than the previously designated *US*. Several of the grants received two codes from individual testers by way of split decision. Each of these factors contributed to a very weak inter-rater reliability. Only 7 of the 27 items (26%) had full or near agreement across all coders (see Appendix F).

Based on these discoveries, the codes were revised to correct errors. Single, corrected codes were assigned to the grants, which significantly improved results. Once the results were corrected, inter-rater reliability increased to 81%, with 22 of the 27 items rating full or near agreement across all coders (see Appendix G).

# DISCUSSION and RECOMMENDATIONS

The results from the second coding tests warranted much further discussion. Overall, the group determined that *Consumer Health Information* was not a broad enough category. *Information* *Science* was used far too often as a dumping ground for grants that had no other clear code assignation, and it should be used to categorize education projects. *Clinical Research Informatics*, although an important concept, should be assimilated into another category as it is still a fairly new idea.

The coding team decided to alter a few of the investment category definitions to include coding instructions based on use cases. The *Consumer Health Informatics* definition remained largely the same. However, the team added a caveat to the concept of ‘consumer’. If the consumer in question is a caregiver or scientist, public health official or other professional, that particular grant should not be categorized as *Consumer Health Informatics*, but rather placed in one of the other categories. The *Information Science* category was adjusted to include instructional technologies for academic programs and medical schools. Yet if more than 50 percent of a basic information science grant project falls into the realm of health care or any other use case, it should be coded in the category that applies to the project’s data source or audience. *Clinical Research Informatics* grants, rather than being given their own new code like had been done in the second coding test, would be divided into two grant categories. If the projects draw data from electronic health records or other health care records, they belong in the *Clinical Informatics* category. If the data is drawn from biological specimens or other non-human sources, they should be coded as *Translational Bioinformatics*. If given more time, it would be interesting to see results from a third test that used the revised code definitions.

The natural language processing approach, even though it garnered poor results, would also benefit from further exploration after some corrections. Dr. Antonio Jimeno, who ran the cosine similarity test, noted that some of the tested grant titles were not complete, which weakened the similarities considerably. He also reported that providing linked grant abstracts would strengthen the similarities by providing a larger context for automatic labeling. However, since these are unfunded grants, their abstracts are considered confidential and are unable to be released for testing. After discussion of these findings, the team considered providing Antonio with a list of the awarded grants from 2012 with their accompanying abstracts in order to better teach the system. The similarities in complete grant titles with abstracts would most likely be significantly greater, and thus the machine learning method would provide more reliable automatic mapping results.

One testing issue, however, could not be so easily corrected. The investment categories and their corresponding two-letter codes changed over time throughout testing (see Figure 8). Without a consensus on the correct categories and codes, this caused inconsistencies in scoring the second coding test. This led to difficulties in comparing the data in the two tests, and it caused confusion for the coders.

| First Coding Test | | AMIA Definitions | | Second Coding Test | | Second Coding Test (Revised) | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Bioinformatics | BB | Bioinformatics | BB | Bioinformatics | BB/BI | Bioinformatics | BB |
| Health Care | HC | Clinical Informatics | HC | Clinical Informatics | HC/CI | Clinical Informatics | HC |
| User Sciences | US | Consumer Health Informatics | US | Consumer Health Informatics | CH | Consumer Health Informatics | CH |
| Basic Informatics | BI | Information Science | BI | Information Science | IS | Information Science | IS |
| Public Health Informatics | PH | Public Health Informatics | PH | Public Health Informatics | PH | Public Health Informatics | PH |
| Translational Informatics | TR | Translational Bioinformatics | TR | Translational Bioinformatics | TR/TB | Translational Bioinformatics | TR |
|  |  |  |  | Clinical Research Informatics | CRI |  |  |

Figure 8: Changing Investment Category Names and Codes

Once the codes for the second coding test were corrected, it showed significantly more reliable results than the first test. However, since the overall definitions the coders used had not changed, it is impossible to determine why the results from the second coding test were so much better than the results from the first coding test.

After the coders all reviewed the results from the tests and agreed upon the revised code definitions, they agreed to periodically run tests in the future to check on the strength of inter-rater reliability and the comprehensiveness of the investment category definitions.

Overall, the text mining and semantic natural language processing experiments proved ineffective. However, the discussion of these results and those of the two coding tests highlighted specific areas that needed improvement, as well as fostering discussion and analysis of the category terms by the group, which then led to unanimously agreed upon revision and acceptance of the new category definitions that would not have occurred had the experiments not taken place. It seems that the experiments were successful in that they served to prove that a complete redesign of the grant categories would not be necessary at this time.

# REFERENCES

1. Charting a Course for the 21st Century – NLM’s Long Range Plan 2006-2016 [Internet]. Bethesda (MD): National Library of Medicine (US); 2007 May 14 [cited 2012 Dec 28]. Available from: <http://www.nlm.nih.gov/pubs/plan/lrp06/report/default.html>
2. Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition [Internet]. 1993 [cited 2012 Dec 28]; 5(2):199-220. Available from: <http://tomgruber.org/writing/ontolingua-kaj-1993.htm>
3. Translational Bioinformatics [Internet]. Bethesda (MD): American Medical Informatics Association; c2012 [cited 2012 Dec 27]. Available from: <http://www.amia.org/applications-informatics/translational-bioinformatics>
4. Consumer Health Informatics [Internet]. Bethesda (MD): American Medical Informatics Association; c2012 [cited 2012 Dec 27]. Available from: <http://www.amia.org/applications-informatics/consumer-health-informatics>
5. Masys DR. American College of Medical Informatics Fellows and International Associates, 2005. J Am Med Inform Assoc [Internet]. 2005 [cited 2013 Jan 3]; 13:360-364. Available from: <http://jamia.bmj.com/content/13/3/360.extract>
6. Rindflesch TC. Natural language processing. Annu Rev Appl Linguist [Internet]. 1996 [cited 2013 Jan 3]; 16:71-85. Available from: <http://skr.nlm.nih.gov/papers/references/aral96.pdf>

# APPENDIX A: Original Grant Code Definitions

**Bioinformatics** is research, development, or application of computational tools and approaches for expanding the use of biological data, including those to acquire, store, organize, archive, analyze, or visualize such data. Bioinformatics is rooted in life sciences as well as computer and information sciences and technologies. Its interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful.

**Clinical Informatics** is the application of informatics in delivery of healthcare services. Informatics when used for healthcare delivery would be essentially the same regardless of the health professional group involved (whether dentist, pharmacist, physician, nurse, or other health professional). Clinical Informatics is concerned with information use in healthcare by clinicians. Included topics range from clinical decision support to integration and analysis of visual images (e.g. radiological, pathological, dermatological, ophthalmological, etc.); from clinical documentation to provider order entry systems; and from system design to system implementation and adoption issues.

**Consumer Health Informatics** is the field devoted to informatics from multiple consumer or patient views. These include patient-focused informatics, health literacy and consumer education, personal health records. The focus is on information structures and processes that empower consumers to manage their own health--for example health information literacy, consumer-friendly language, personal health records, and Internet-based strategies and resources. Researchers in this field analyze consumers' needs for information; study and implement methods for making information accessible to consumers; and model and integrate consumers' preferences into health information systems.

**Information Science** is an interdisciplinary field primarily concerned with the collection, classification, analysis, manipulation, storage, retrieval, visualization, and dissemination of information related to health. Research and development projects within the field study usage of knowledge in organizations, along with the interaction between people, organizations and information systems with the aim of creating, enhancing and understanding information systems. Information science is cross cutting across the subdisciplines of biomedical informatics with the core results generalizing to more than one domain.

**Public Health Informatics** is the application of informatics in areas of public health, including surveillance, prediction of epidemics, reporting, and health promotion. Public health informatics, and its corollary, population informatics, are concerned with groups rather than individuals. It also includes the application of informatics to ecology, climate change, health disparities and environmental factors in community health. Public health informatics projects enable the development and use of interoperable informatics systems for public health functions such as biosurveillance, disaster response, and electronic laboratory reporting.

**Translational Bioinformatics** is the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health. It is a field where bioinformatics meets clinical medicine. Translational bioinformatics focuses on applications of bioinformatics innovations within a clinical context and touches nearly all areas of biological, biomedical, and clinical research. Work in translational bioinformatics will typically include informatics methodology, clinical concepts (drugs, diseases, symptoms, diagnosis), and molecules (genes, proteins, DNA, RNA, small molecules, drugs).

# APPENDIX B: Specific Long Range Plan Goals

| Investment Category | Code | Link to NLM LRP Goals |
| --- | --- | --- |
| Bioinformatics | BB | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.2, 3.3 |
| Clinical Informatics | HC | Goal 3: Rec 3.1, 3.2, 3.3 |
| Consumer Health Informatics | US | Goal 2: Rec 2.1, 2.4, 2.5 |
| Information Science | BI | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.3 |
| Public Health Informatics | PH | Goal 3: Rec 3.2, 3.3 |
| Translational Bioinformatics | TR | Goal 2: Rec 2.5  Goal 3: Rec 3.1, 3.2, 3.3 |

Charting a Course for the 21st Century – NLM’s Long Range Plan 2006-2016

Goal 2: Trusted Information Services that Promote Health Literacy, Improve Health Outcomes, and Reduce Health Disparities Worldwide

Recommendation 2.1. Advance new outreach programs by NLM and NN/LM for underserved populations at home and abroad; work to reduce health disparities experienced by minority populations; share and actively promote lessons learned.

Recommendation 2.4. Test and evaluate digital infrastructure improvements (e.g., PDAs, intelligent agents, network techniques) to enable ubiquitous health information access in homes, schools, public libraries, and work places.

Recommendation 2.5. Support research on the application of cognitive and cultural models to facilitate information transfer and trust building and develop new methodologies to evaluate the impact of health information on patient care and health outcomes.

Goal 3: Integrated Biomedical, Clinical, and Publix Health Information Systems that Promote Scientific Discovery and Speed the Translation of Research into Practice

Recommendation 3.1. Develop linked databases for discovering relationships between clinical data, genetic information, and environmental factors.

Recommendation 3.2. Promote development of Next Generation electronic health records to facilitate patient-centric care, clinical research, and public health.

Recommendation 3.3. Promote development and use of advanced electronic representations of biomedical knowledge in conjunction with electronic health records.

# APPENDIX C: Initial Coding Test Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KEY** |  |  | of 25 items | |
| all 4 | agreement across coders | | | 7 |
| 3 of 4 | near agreement across coders | | | 5 |
| 2 and 2 | split agreement | | | 2 |
| 4 diff | different for each coder | | | 2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Title** | **#1** | **#2** | **#3** | **#4** |
| Bayesian Methods in Signal Transduction Network Analysis | BB | BB | BB | BB |
| Computational Methods for Expression Image Analysis | BB | BB | BB | BB |
| VizBi: A Conference on Visualization in Biology | BB | US | BB | BI |
| Large-scale evaluation of text features affecting perceived and actual text diffi | US | US | US | BI |
| Biocomputation across distr. private datasets to enhance drug discovery | BI | TR | BB | BI |
| Delivering Geospatial Intelligence to Health Care Professionals | HC | HC | HC | HC |
| Automated matching of relevant research studies to pt records for EBM | HC | HC | TR | HC |
| Exploring the Feasibility of Approximate Sequential Pattern Discovery in Massive | HC | HC | HC | HC |
| Multi-Institutional Pediatric Epilepsy Decision Support | HC | HC | HC | HC |
| Interactive Search and Review of Clinical Records with Multi-layered Semantic Ann | HC | TR | HC | HC |
| POET-2: High-performance computing for advanced clinical narrative preprocessing | BI | BI | HC | HC |
| Integrating Machine Learning and Physician Expertise for Breast Cancer Diagnosis | HC | HC | HC | HC |
| Development of a clinical robotic device for diagnosis, rehabilitation and treatm | BI | HC | BI | HC |
| A Mixed Reality Conscious Sedation Simulator for Learning to Manage Variability | BI | US | HC | HC |
| Secure Sharing of Clinical History & Genetic Data: Empowering Predictive Pers. Me | BI | TR | HC | PH |
| New Technology to Preserve Patient Privacy and Data Quality in Health Research | BI | TR | HC | PH |
| Informatic profiling of clinically relevant mutation | BB | BI | BB | TR |
| Bayesian Rule Learning Method for Disease Predict and Biomarker Disc | TR | BI | BB | TR |
| Informatics for Integrative Brain Tumor Whole Slide Analysis | TR | BB | BB | TR |
| Ontology-Driven Methods for Knowledge Acq. and Knowledge Disc. | TR | BI | BB | TR |
| From GWAS to PheWAS: Scanning the EMR phenome for gene-disease associations | TR | TR | TR | TR |
| Speech Therapy Robot (STR) to assist in the administration of evidence based speech | BI | HC | BI | US |
| Patient-Specific Simulation for Surgical Rehearsal | BI | US | HC | US |
| Development of a mobile robot with an affective interface and human activity trac | US | US | BI | US |
| Toward intelligent display of health data: A qualitative study of use patterns | US | US | HC | US |

# APPENDIX D: Bioinformatics Expanded MeSH Tree Structure

**MeSH Heading:** Computational Biology

**Scope:** A field of biology concerned with the development of techniques for the collection and manipulation of biological data, and the use of such data to make biological discoveries or prediction. This field encompasses all computational methods and theories applicable to molecular biology and areas of computer-based techniques for solving biological problems including manipulation of models and datasets.

**Entry Terms:** Bio-Informatics; Bioinformatics; Biology, Computational; Computational Molecular Biology; Molecular Biology, Computational

**See Also:** Medical Informatics  
**Tree Structure:**

Natural Science Disciplines

Biological Science Disciplines

Biology

**Computational Biology**

Genomics+

Epigenomics

Glycomics

HapMap Project

Human Genome Project

Nutrigenomics

Proteomics

Metabolomics

Systems Biology

Genomics

MeSH Heading: Genomics

Scope: The systematic study of the complete DNA sequences (genome) of organisms.

See Also: Computational Biology; Human Genome Project; Proteomics; Sequence Analysis, DNA

Epigenomics

MeSH Heading: Epigenomics

Scope: The systematic study of the global gene expression changes due to epigenetic processes and not due to DNA base sequence changes.

Entry Term: Epigenetics

Epigenetic Processes

MeSH Heading: Epigenesis, Genetic

Scope: A genetic process by which the adult organism is realized via mechanisms that lead to the restriction in the possible fates of cells, eventually leading to their differentiated state. Mechanisms involved cause heritable changes to cells without changes to DNA sequence such as DNA methylation, histone modification, DNA replication timing, nucleosome positioning, and heterochromization which result in selective gene expression or repression.

See Also: DNA Methylation; Morphogenesis

Glycomics

MeSH Heading: Glycomics

Scope: The systematic study of the structure and function of the complete set of glycans (the glycome) produced in a single organism and identification of all the genes that encode glycoproteins.

Entry Term: Glycobiology

See Also: Carbohydrates; Metabolism

HapMap Project

MeSH Heading: HapMap Project

Scope: A coordinated international effort to identify and catalog patterns of linked variations (haplotypes) found in the human genome across the entire human population.

Entry Terms: HapMap; Human Haplotype Map; International HapMap Project

See Also: Haplotypes

Haplotypes

MeSH Heading: Haplotypes

Scope: The genetic constitution of individuals with respect to one member of a pair of allelic genes, or sets of genes that are closely linked and tend to be inherited together such as those of the major histocompatibility complex.

See Also: Genotyping Techniques

Human Genome Project

MeSH Heading: Human Genome Project

Scope: A coordinated effort of researchers to map (chromosome mapping) and sequence (sequence analysis, DNA) the human genome.

Entry Terms: Genome Project, Human; Human Genome Diversity Project

See Also: Genomics

Nutrigenomics

MeSH Heading: Nutrigenomics

Scope: The study of the relationship between nutritional physiology and genetic makeup. It includes the effect of different food components on gene expression and how variations in genes affect responses to food component.

Entry Terms: Nutrigenetics; Nutritional Genetics; Nutritional Genomics

See Also: Metabolomics

Proteomics

MeSH Heading: Proteomics

Scope: The systematics study of the complete complement of proteins (proteome) of organisms.

See Also: Genomics

Metabolomics

MeSH Heading: Metabolomics

Scope: The systematic identification and quantification of all the metabolic products of a cell, tissue, organ, or organism under varying conditions. The metabolome of a cell or organism is a dynamic collection of metabolites which represent its net response to current conditions.

Entry Term: Metabolomics

See Also: Nutrigenomics

Systems Biology

MeSH Heading: Systems Biology

Scope: Comprehensive, methodical analysis of complex biological systems by monitoring responses to perturbations of biological processes. Large scale, computerized collection and analysis of the data are used to develop and test models of biological systems.

See Also: Systems Theory

Systems Theory

MeSH Heading: Systems Theory

Scope: Principle, models, and laws that apply to complex interrelationships and interdependencies of sets of linked components which form a functioning whole, a system. Any system may be composed of components which are systems in their own right (sub-systems), such as several organs within an individual organism.

Entry Term: General Systems Theory; Queuing Theory

See Also: Models, Theoretical; Systems Analysis; Systems Biology

# APPENDIX E: Cosine Similarity Results

Uncovering and reducing health literacy barriers to tobacco cessation

US|0.2654243731222853

PH|0.22478935866813277

HC|0.1074430618700507

BI|0.03479445003196102

TR|0.02933573244244292

BB|0.0

What Did the Doctor Say? What did the Patient Hear?

US|0.10136060675992287

HC|0.0410304969931109

BI|0.0

PH|0.0

BB|0.0

TR|0.0

Computer Agents to Promote Walking in Older Adults with Low Health Literacy

US|0.23408229439226114

PH|0.22027286681836322

BB|0.10256410256410253

HC|0.09475587393582685

BI|0.030685820596610736

TR|0.02587168419021113

Promoting Asthma Self Care in Inner City Patients with Low Health Literacy

US|0.31035220822588805

PH|0.23363465675799888

HC|0.17588161767036214

BI|0.03254722774520591

TR|0.027441064997422604

BB|0.0

Bench to Book: A Vertically Integrated Infrastructure for System-Level Bioscience

HC|0.12562972690740148

BI|0.09764168323561795

US|0.04138029443011837

BB|0.027196414661021073

PH|0.023363465675799833

TR|0.0

Development and Validation of Decision Models

HC|0.10660035817780522

BI|0.09205746178983232

BB|0.07692307692307687

US|0.058520573598065284

TR|0.038807526285316696

PH|0.033040930022754544

Intelligent Histories: Detecting Personalized Risk with Longitudinal Surveillance

US|0.06635609328057135

PH|0.0499531908151406

HC|0.026860765467512704

BI|0.0

BB|0.0

TR|0.0

Artificial Expert: Making Neuropsychiatric Decision Support Models Automatically

HC|0.15075567228888187

BB|0.05439282932204215

US|0.04138029443011837

BI|0.03254722774520591

TR|0.027441064997422604

PH|0.0

Homology Modeling of 3D Structures of Protein-Protein Complexes

TR|0.05488212999484521

BB|0.05439282932204215

US|0.04138029443011837

BI|0.03254722774520591

PH|0.0

HC|0.0

A software library and toolkit for genomic analyses of transcriptional regulation

TR|0.02933573244244292

US|0.0

BI|0.0

PH|0.0

BB|0.0

HC|0.0

PH|0.029552706228277104

# APPENDIX F: Second Coding Test Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KEY** |  |  | of 27 items | |
| all 5 | agreement across coders | | | 2 |
| 4 of 5 | near agreement across coders | | | 5 |
| 3 of 5 | split agreement | | | 8 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Title** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Modeling Transcriptional Reprogramming by Markov Chain Monte Carlo Sampling | BB | BB | BB | BI | PH |
| Mass Casualty Management System (DIORAMA-II) | PH | PH | PH | PH | PH |
| An Information Fusion Approach to Longitudinal Health Records | HC/CRI | HC | HC | TB/CRI | HC |
| Informatic tools for predicting an ordinal response for high-dimensional data | IS/CRI | BB | TR | IS | BB |
| Mining Social Network Postings for Mentions of Potential Adverse Drug Reactions | PH | CH | IS | CH | PH |
| Mapping the Genetic Architecture of Complex Disease via RNA-seq and GWAS Data | TR | TR | TR | TB | BB |
| Methods for Accurate and Efficient Discovery of Local Pathways. | TR | TR | TR | BI | TR |
| Improving Network Analysis and Visualization for Infectious Disease Control | PH | PH | PH | PI | IS |
| Active Patient Participation in a Disease Registry for Comparative Effectiveness | TR/CRI | HC | HC | CH | CH |
| Enhancing Genome-Wide Association Studies via Integrative Network Analysis | TR | TR | TR | TB | BB |
| RUMI: A patient portal for retrieving understandable medical information | CH | CH | CH | CH | CH |
| Assist Patients with Medication Decisions | CH | HC | CH | CH | HC |
| Bioinformatics Strategies for Multidimensional Brain Imaging Genetics | TR | TR | TR | TB | BB |
| Leveraging the EHR to Collect and Analyze Social, Behavioral & Familial Factors | HC/CRI | HC | HC | CI | HC |
| Scalable and Robust Clinical Text De-Identification Tools | HC/CRI | HC | HC | CI/IS | IS |
| Best Practices in Telemedicine Symposium-Workshop | HC | CH | HC | CI/IS | IS |
| Challenges in Natural Language Processing for Clinical Narratives | IS | HC | HC | CI/IS | BB |
| Exploring the Feasibility of Computational Markers to Predict Atrial Fibrillation | IS/HC | HC | HC | CI | TR |
| Exploratory evaluation of homomorphic cryptography for confidentiality protection | IS | PH | IS | IS | IS |
| A machine learning approach for fine-scale genome wide DNA methylation analysis | BB | HC | BB | BI | BB |
| A Paper-Digital Interface for Time-Critical Information Management | HC | HC | HC | CI | IS |
| Applying NLP to Free Text as an EHR Data Capture Method to Improve EHR Usability | HC | HC | HC | CI | HC |
| The Climate Change and Health Gateway (CCHG) for Enhancing Biomedical, Health, an | PH | PH | PH | PH | TR |
| Research Platform Integrating Patient Reported and Clinical Outcomes Data Sources | IS/CRI | HC | TR | CI/CRI | TR |
| Mobile Cadaver Lab: An Innovative Platform to Supplement Medical Education for Mo | IS | IS | IS | IS | HC |
| Tools for Coordination Among Caregivers of Alzheimers Disease Patients | HC | CH | CH | IS | CH |
| Bridging the Semantic Gap Between Research Eligibility Criteria and Clinical Data | IS/CRI | HC | HC | CI/CRI | HC |

# APPENDIX G: Second Coding Test Results (Revised)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KEY** |  |  | of 27 items | |
| all 5 | agreement across coders | | | 8 |
| 4 of 5 | near agreement across coders | | | 14 |
| 3 of 5 | split agreement | | | 3 |
| 5 diff | different for each coder | | | 2 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Title** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Modeling Transcriptional Reprogramming by Markov Chain Monte Carlo Sampling | BB | BB | BB | BB | PH |
| Mass Casualty Management System (DIORAMA-II) | PH | PH | PH | PH | PH |
| An Information Fusion Approach to Longitudinal Health Records | HC | HC | HC | HC | HC |
| Informatic tools for predicting an ordinal response for high-dimensional data | TR | BB | TR | IS | BB |
| Mining Social Network Postings for Mentions of Potential Adverse Drug Reactions | PH | CH | IS | CH | PH |
| Mapping the Genetic Architecture of Complex Disease via RNA-seq and GWAS Data | TR | TR | TR | TR | BB |
| Methods for Accurate and Efficient Discovery of Local Pathways. | TR | TR | TR | TR | TR |
| Improving Network Analysis and Visualization for Infectious Disease Control | PH | PH | PH | PH | IS |
| Active Patient Participation in a Disease Registry for Comparative Effectiveness | HC | HC | HC | CH | CH |
| Enhancing Genome-Wide Association Studies via Integrative Network Analysis | TR | TR | TR | TR | BB |
| RUMI: A patient portal for retrieving understandable medical information | CH | CH | CH | CH | CH |
| Assist Patients with Medication Decisions | CH | CH | CH | CH | HC |
| Bioinformatics Strategies for Multidimensional Brain Imaging Genetics | TR | TR | TR | TR | BB |
| Leveraging the EHR to Collect and Analyze Social, Behavioral & Familial Factors | HC | HC | HC | HC | HC |
| Scalable and Robust Clinical Text De-Identification Tools | HC | HC | HC | HC | IS |
| Best Practices in Telemedicine Symposium-Workshop | HC | HC | HC | HC | IS |
| Challenges in Natural Language Processing for Clinical Narratives | IS | HC | HC | HC | BB |
| Exploring the Feasibility of Computational Markers to Predict Atrial Fibrillation | HC | HC | HC | HC | TR |
| Exploratory evaluation of homomorphic cryptography for confidentiality protection | IS | PH | IS | IS | IS |
| A MACHINE LEARNING APPROACH FOR FINE-SCALE GENOME WIDE DNA METHYLATION ANALYSIS | BB | BB | BB | BB | BB |
| A Paper-Digital Interface for Time-Critical Information Management | HC | HC | HC | HC | IS |
| Applying NLP to Free Text as an EHR Data Capture Method to Improve EHR Usability | HC | HC | HC | HC | HC |
| The Climate Change and Health Gateway (CCHG) for Enhancing Biomedical, Health, an | PH | PH | PH | PH | TR |
| Research Platform Integrating Patient Reported and Clinical Outcomes Data Sources | HC | HC | TR | HC | TR |
| Mobile Cadaver Lab: An Innovative Platform to Supplement Medical Education for Mo | IS | IS | IS | IS | HC |
| Tools for Coordination Among Caregivers of Alzheimers Disease Patients | HC | CH | CH | IS | CH |
| Bridging the Semantic Gap Between Research Eligibility Criteria and Clinical Data | HC | HC | HC | HC | HC |