Biomedical Data Retrieval Utilizing Textual Data in a Gene Expression Database

By

Richard Lu, MD

SUBMITTED TO THE DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN BIOMEDICAL INFORMATICS AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 17, 2010
[June 2010]

Signature of Author: _____
Division of Health Sciences and Technology
May 17, 2010

Certified by: _____
Ronilda Lacson, MD, PhD

Accepted by: _____
Ram Sasisekharan, Ph.
Director, Harvard-MIT Division of Health Sciences and Technology; Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

Biomedical Data Retrieval Utilizing Textual Data in a Gene Expression Database

Submitted to the Division of Health Sciences and Technology
on May 17, 2010 in partial fulfillment of the requirements for the
Degree of Master of Science in Biomedical Informatics

ABSTRACT

***Background**: The commoditization of high-throughput gene expression sequencing and microarrays has led to a proliferation in both the amount of genomic and clinical data that is available. Descriptive textual information deposited with gene expression data in the Gene Expression Omnibus (GEO) is an underutilized resource because the textual information is unstructured and difficult to query. Rendering this information in a structured format utilizing standard medical terms would facilitate better searching and data reuse. Such a procedure would significantly increase the clinical utility of biomedical data repositories.* **Methods:** *The thesis is divided into two sections. The first section compares how well four medical terminologies were able to represent textual information deposited in GEO. The second section implements free-text search and faceted search and evaluates how well they are able to answer clinical queries with varying levels of complexity. Part I: 120 samples were randomly extracted from samples deposited in the GEO database from six clinical domains—breast cancer, colon cancer, rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type I diabetes mellitus (IDDM), and asthma. These samples were previously annotated manually and structured textual information was obtained in a tag:value format. Data was mapped to four different controlled terminologies: NCI Thesaurus, MeSH, SNOMED-CT, and ICD-10. The samples were assigned a score on a three-point scale that was based on how well the terminology was able to represent descriptive textual information. Part II: Faceted and free-text search tools were implemented, with 300 GEO samples included for querying. Eight natural language search questions were selected randomly from scientific journals. Academic researchers were recruited and asked to use the faceted and free-text search tools to locate samples matching the question criteria. Precision, recall, F-score, and search time were compared and analyzed for both free-text and faceted search.* **Results:** *The results show that the NCI Thesaurus consistently ranked as the most comprehensive terminology across all domains while ICD-10 consistently ranked as the least comprehensive. Using NCI Thesaurus to augment the faceted search tool, each researcher was able to reach 100% precision and recall (F-score 1.0) for each of the eight search questions. Using free-text search, test users averaged 22.8% precision, 60.7% recall, and an F-score of 0.282. The mean search time per question using faceted search and free-text search were 116.7 seconds, and 138.4 seconds, respectively. The difference between search time was not statistically significant ($p=0.734$). However, paired t-test analysis showed a statistically significant difference between the two search strategies with respect to precision ($p=0.001$), recall ($p=0.042$), and F-score ($p<0.001$).* **Conclusion:** *This work demonstrates that biomedical terms included in a gene expression database can be adequately expressed using the NCI Thesaurus. It also shows that faceted searching using a controlled terminology is superior to conventional free-text searching when answering queries of varying levels of complexity.*

Thesis Supervisor: Ronilda Lacson, MD, PhD
Title: Assistant Professor of Computer Science and Engineering & Health Sciences and Technology

Acknowledgments

The author would like to thank the following people:

Ronilda Lacson, MD, PhD for her mentorship during the informatics fellowship and for serving as thesis advisor.

The National Library of Medicine for NLM Training Grant 5T15-LM007092-17

Boston-Area Biomedical Informatics Research Training Program

Decision Systems Group, Brigham and Women's Hospital

Lucila Ohno-Machado, MD, PhD
Chris Hinski, MD, MS
Jihoon Kim, MS
Kumiko Oohashi, PhD
Erik Pitzer, PhD

Center for Evidence-Based Imaging, Brigham and Women's Hospital

Ramin Khorasani, MD, MPH
Kathy Andriole, PhD
Joaquin Blaya, PhD
Esteban Gershanik, MD, MPH
Shanta Griffin, PhD
Ivan Ip, MD, MPH
Cleo Maehara, MD
Luciano Prevadello, MD, MPH

Table of Contents

Introduction

Several methods have been employed to better organize and extract relevant textual information from vast databases on demand. These include natural language processing, manual annotation, faceted categorization, and semantic web technologies. All of these information-structuring and information-extraction techniques work together to accomplish the task of improving accessibility to large amounts of data.

Recognizing that biological information will primarily be consumed and deposited through the web and that finding relevant information will become more important than ever, this paper evaluates existing biomedical terminologies' ability to express biomedical terms inside gene expression repositories and identifies an optimal strategy to search through terminology-compliant, annotated samples.

Specific Aims

This paper aims to accomplish three tasks:

1. To evaluate the ability of various established medical terminologies to express and capture the clinical textual content deposited within a gene expression database, the Gene Expression Omnibus (GEO).

2. To build web-based, faceted and free-text search tools for locating annotated biological samples deposited within the Gene Expression Omnibus.

3. To compare faceted search to traditional free-text search in identifying annotated biological samples.

Chapter 1: Comparative Analysis of Four Controlled Medical Terminologies for Expressing Biomedical Data in a Gene Expression Database

Background

1.1 Microarray Technology

The completion of the thirteen-year, $4.3 billion Human Genome Project in 2003 [1, 2] was a seminal moment in biology. Knowing the base sequences that make up an entire human being forms the foundation upon which all other genomic discoveries are based. However, raw sequences can be rendered more informative, and subsequent areas of research are well underway that build upon what the Human Genome Project made possible. Such research areas include identifying gene function, investigating protein-protein interaction, and correlating single nucleotide polymorphisms (SNP) with disease. [2] These research fields work towards fully describing the steps in the pathway from nucleic acid sequences to physical characteristics that we can observe clinically.

One of the key tools involved in illuminating the genotype to phenotype pathway are microarrays, a technology that has revolutionized genomic research. In fact, the largest source of genomic data currently comes from analyzing microarrays.[3] The now routine process of assaying large numbers of genes simultaneously has reduced the time and cost that it takes to decipher

gene function. Concurrent with the growth in microarray adoption is the surge in the amount of gene expression data that is available.

1.2 Gene Expression Databases

In the field of genomics, centralized, online gene expression databases have been created in an effort to organize the vast amounts of data being generated. Further encouraging the submission of expression data is the fact that most journals mandate that gene expression data be submitted as a prerequisite for publication. The primary aim of these databases is to realize the general benefits of aggregating data in a centralized place—increased visibility, data sharing, and data mining.

Over seventy-five gene expression databases or tools for analyzing them are currently available online.[4] Selected examples include: ArrayExpress, Center for Information Biology Gene Expression Database (CIBEX), Gene Expression Omnibus (GEO), A Database Of Heterogenous Gene Expression Data Based on A Consistent Gene Nomenclature (CleanEx), Database for Annotation, Visualization, and Integrated Discovery (DAVID), Database of Gene Expression in Normal Adult Human Tissues (GeneNote), Gene Expression Database (GXA), the Stanford Tissue Microarray Database (TMAD), OncoMine, and the Reference Database For Human Gene Expression Analysis (RefExA).

Nevertheless, in practice the Microarray Gene Expression Data (MGED) Society recommends the following three repositories for storing gene expression data: ArrayExpress, CIBEX, and GEO. ArrayExpress was created by the European Bioinformatics Institute (EBI) and went online in 2002. ArrayExpress has the following three core features: a web-based

interface for uploading gene expression data, a query tool for finding normalized and curated gene expression data, and a data visualization and analysis tool.[5] ArrayExpress allows users to query the database by species, author, platform, gene attributes, gene names, gene function, gene classification, and sample properties.[5]

CIBEX was developed in an effort to collect expression data from researchers in mainly Asian countries. Like ArrayExpress, CIBEX is standards-compliant and allows researchers to upload expression data as well as to query and visualize it. One can filter CIBEX samples according to experimental and biological conditions, authors, gene names, and even according to the hardware platform used. [6] A unique feature of CIBEX is its spot-based visual viewer. When a user searches for an experimental or biological condition, matching spot images for the condition are displayed. The user can then click on the spot for detailed information. [6]

The largest gene expression database,[7, 8] the Gene Expression Omnibus (GEO), is run by the National Cancer and Blood Institute (NCBI) and serves as a "public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community." [9] GEO has undoubtedly been a valuable resource for bench researchers looking for gene expression data.

All the data inside GEO can be divided into two general categories: gene expression measurements and the metadata about each biological sample. [8] The NCBI recognizes the value of collecting as much metadata about GEO's biological samples as possible because such information is needed in order to effectively search for and compare samples with each other.

As a result, the NCBI has organized GEO's samples into logical groups of varying granularity called GEO Series (GSE) and GEO Datasets (GDS). GEO Series are "a set of related Samples considered to be part of a study, and describes the overall study aim and design." GEO

Datasets are both automatically and manually curated. They are a "collection of consistently processed, experimentally related Sample records, summarized and categorized according to experimental variables." [10] Most of these experimental variables relate to gene expression measurements or are characteristics of the experiment that was performed.

The next most-granular grouping of GEO samples that the NCBI provides is a unit called a GEO Series. GEO Series refers to GEO samples that all come from one particular study. GEO Series and GEO Samples (GSM) are the most suitable places to look for clinical metadata about biological specimens. GEO Series, because they consist of samples taken from the same study, typically contains descriptive information that applies to all samples. A specific GEO Series has two fields of interest: the study title and the study summary. The study title is the title of the journal article that the samples were used in. If the samples were not used in a study for publication, then the title field's value is provided by the person who uploaded the samples to GEO or assigned by NCBI staff. The study summary is typically the abstract from the published study that referenced the samples. In some cases, however, the summary is not the abstract from the published study but was likely written by someone associated with the original experiment when uploading samples to GEO.

The final and most granular organizational unit in GEO is the individual GEO Sample (GSM) itself. Samples consist of a "description of the biological material and the experimental protocols to which it was subjected, and ... may hold very large volumes of text to allow elaborate descriptions of the biological source..."[10] Since the sample is the lowest level view in GEO, it follows that most descriptive information for a sample would be included here. This is in fact only partly true. Browsing through the GEO database, one will notice that descriptive information is not uniformly deposited. On the GEO website, in their instructions to researchers

who plan to upload samples, GEO advises that descriptive information about a particular sample be deposited in a "characteristics" field, preferably in a tag: value format. However, this is not required.

At present, when submitting experimental data to GEO, the only requirement is that researchers adhere to the Minimum Information About A Microarray Experiment (MIAME) standard. [3] As helpful as the MIAME standard has been in standardizing the representation of technical and numeric data about microarray experiments, this standard was not designed to accommodate the clinical features that are known about biological samples.

The end result is that for researchers who want to perform research based on the clinical characteristics of GEO's biological samples, locating relevant samples using clinical metadata is problematic. Analyzing descriptive information deposited in GEO is not straightforward, and it is not feasible to filter the samples based on clinical and demographic criteria. Examples of such clinical metadata include what disease state a sample came from, disease severity, and what treatment was performed. Demographic features include a sample donor's race, gender, and age. However, no tools exist to effectively display these data. Currently, GEO Datasets offer two summary views of the samples they contain—an experiment-centered view and a gene-centered view.[10] No option exists to display samples based on clinical information at the sample level.

The three primary gene expression databases—ArrayExpress, CIBEX, and GEO—are all well designed for locating and visualizing gene expression data. However, finding clinically-oriented, textual information deposited in these repositories remains difficult. The following two steps are key to addressing this problem and will be addressed in this thesis: 1) identifying a suitable terminology that can express clinical terms used in gene expression experiments, and 2)

devising a way to efficiently and accurately search for descriptive terms in a large database like GEO.

1.3 Linking Genomic Data to Clinical Data

Taking descriptive information used within a gene expression database such as GEO and mapping them to existing medical terminologies is a key step towards being able to query the data efficiently. Doing this would lay the groundwork for various kinds of search queries, whether that be free-text, faceted searching, or semantic-based searching—all of which are useful for doing translational research.

The Stanford Biomedical Informatics group used a controlled terminology to standardize the expression of clinical terms within the Stanford Tissue Microarray Database (TMAD), and serves as a model for mapping the clinical terms inside GEO to an existing medical terminology.

In the TMAD, the Stanford group had an annotated, cancer-specific tissue microarray database. Along with the raw expression data, each tissue sample had a standard set of histopathological and clinical criteria describing it. Each sample's annotations contained the organ from which the tissue came from, the primary diagnosis, and up to four sub-diagnoses (subdiagnoses 1-4). For example, one tissue specimen might be annotated with breast, carcinoma ductal, and in situ. This pattern indicates that the organ is breast, primary diagnosis is ductal carcinoma, and the subdiagnosis is in situ.[11] The problem with this arrangement was that the terms "breast," "carcinoma ductal", and "in situ" were not standardized. Because of this, the group realized that a common category of questions such as "find all tissue samples that have a particular diagnosis" could not be answered because the words used to describe disease states

and diagnoses were heterogenous.[11] In addition, the lack of a backing ontology also hindered integrating the TMAD with other genomic repositories.

The Stanford group solved these problems by parsing all of the histopathological and clinical terms used in TMAD, generating all possible permutations, (over one million), and running their own heuristics to lower the number of permutations to twenty-thousand. Each of these terms was then mapped to the NCI Thesaurus with an 86% success rate.[11]

Because the tissue microarray database contains mostly samples derived from cancer patients or animal models of cancer,[11] the NCI Thesaurus was a logical controlled terminology to choose. The Gene Expression Omnibus, however, holds samples from the full spectrum of biology. Identifying the most appropriate controlled terminology is not as straightforward. After explaining differences between terminologies, thesauri, and ontologies, the relative strengths and weaknesses of several well-known, controlled medical terminologies' ability to express descriptive information inside microarray experiments will be discussed.

1.4 Terminologies versus Thesauri versus Ontologies

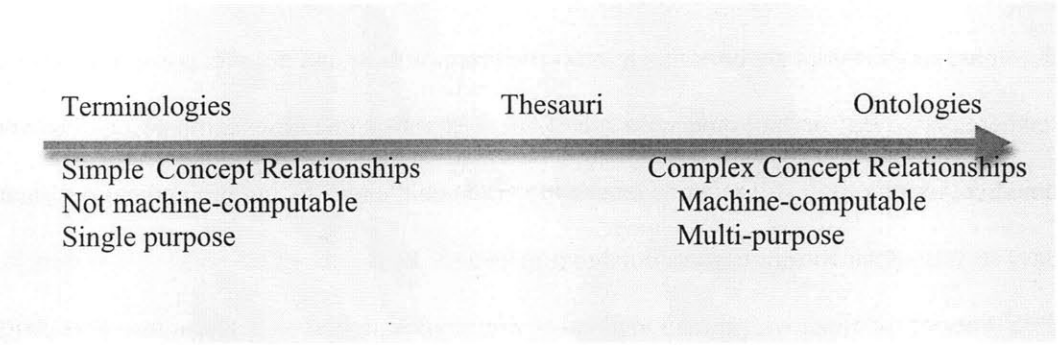
Although the terms "ontology," "thesaurus", "structured vocabulary", and "controlled terminology" are often interchanged, they are separate entities. Clear-cut, universally accepted definitions, however, are hard to find. As defined by Rosenfeld and Morville, a controlled vocabulary is "any defined subset of natural language."[12] Also known as structured vocabulary, controlled terminology or structured terminology, a controlled terminology is a standard group of words, usually agreed upon by consensus that are to be used to describe a domain of knowledge.

A thesaurus encompasses the definition of a controlled terminology and is defined as "a controlled vocabulary in which equivalence, hierarchical, and associative relationships are defined for purposes of improved retrieval."[12] Equivalence is the formal term for synonym support that most people associate with thesauri. Hierarchical relationships in thesauri allow terms to be grouped into categories and subcategories, and associative relationships allows connections between terms that are not handled by equivalence and hierarchical relationships.[12]

The most widely cited definition of an ontology is given by McGuiness and Noy: "a formal explicit description of concepts in a domain of discourse."[13] The line between an ontology and a thesaurus is less clear. An ontology subsumes the properties of a thesaurus, but in addition to equivalence, hierarchies, and associative relationships, ontologies add more relationships and also describe more comprehensive sets of attributes for each concept. Usually, an ontology is expressed in machine-computable language such as the resource description framework (RDF) or the web ontology language (OWL).[14] Ontologies can also be differentiated from thesauri by their broader range of applications: 1) to share common understanding of the

structure of information among people or software agents 2) to enable reuse of domain knowledge 3) to make domain assumptions explicit 4) to separate domain knowledge from operational knowledge 5) to analyze domain knowledge. [13] As one can see, terminologies, thesauri, and ontologies represent a spectrum of knowledge sources that share overlapping properties. Figure 1 helps to clarify the relationship of these three knowledge sources. [12, 15]

Figure 1: Relationship of Terminologies, Thesauri, and Ontologies



1.5 Controlled Medical Terminologies

The value of controlled terminologies, thesauri, and ontologies in biomedical research and clinical medicine have been recognized for decades.[16] Controlled terminologies can serve as the foundational layer underpinning a multitude of purposes, including capturing clinical/biologic findings, natural language processing, indexing medical records, indexing medical literature, and representing medical knowledge.[17]

Presently, over 100 controlled medical terminologies are in use,[18] but far fewer are widely used and established. Unfortunately, not all terminologies are created equal. Several evaluation studies have established that although one might assume that most controlled

terminologies can be used for multiple purposes, this is not the case. [17] Therefore, picking the right terminology for a given situation is a non-trivial task. Further complicating matters is the fact that many controlled medical terminologies exhibit flaws in their logical consistency or adherence to accepted design principles when examined under close scrutiny.[19]

In light of this heterogeneity, Cimino has synthesized a list of best practices to consider when building and evaluating controlled terminologies. Quality controlled terminologies must be multipurpose, capture the full discourse of its intended domain, be based on concepts that are uniquely identifiable, display concept permanence, have a hierarchical arrangement, have formal definitions, support viewing concepts at multiple granularities, and not recognize the terms "not elsewhere classified" or "not otherwise specified."[17] Creating one all-encompassing controlled biomedical terminology still remains one of the grand challenges facing biomedical informatics today, 16 years after Sitting first articulated it in 1994.[20]

Four of the most widespread medical terminologies are the NCI Thesaurus, SNOMED-CT, MeSH, and ICD-10. NCI Thesaurus is included for evaluation because of its stated goal of unifying molecular and clinical information into a single biomedical informatics framework[21] is closest in line with the aims of this research. SNOMED-CT is under evaluation because it is the largest clinical medical vocabulary currently in use. [22] MeSH is included because of its ubiquity in biomedical research and because it is used by some of GEO's sample query tools. Last, ICD-10 is under consideration because it is the oldest controlled terminology and arguably the most popular.

NCI Thesaurus

The National Cancer Institute (NCI) Thesaurus is a controlled terminology that is designed to cover "vocabulary for clinical care, translational and basic research, and public information and administrative activities." Initiated in 1997, the thesaurus contains vocabulary for over 10,000 cancers and 8,000 therapies for cancer [23] and over 60,000 concepts. [24] Its designers list three primary goals for the thesaurus: 1) provide an up-to-date cancer terminology based on science 2) use best practices to formally connect concepts to each other in ways that support automated reasoning 3) include the newest concepts and relationships from clinical trials and bench research.[19]

Despite its name, NCI Thesaurus functions essentially as an ontology, as well as a controlled terminology and a thesaurus.[19] Further, while the focus is on the cancer domain, the ontology contains concepts for far more than just cancer. The ontology is composed of three fundamental units: concepts, kinds, and roles. A kind in the NCI Thesaurus is a set of concepts much like an abstract superclass. Examples of kinds are: Anatomy (4,320 concepts), Biological Processes, Chemicals and Drugs (3,351 concepts), Genes, Findings and Disorders (10,000 concepts) Techniques, Anatomy, and Diagnostic and Prognostic Factors.[18] [25] Concepts are atomic terms that express a discrete idea. Concepts can contain annotations such as synonyms, a preferred name, references to external resources, and a standard definition.[26] Roles signify the relationship between concepts, such as is_a and has_a relationships. [25] The thesaurus contains twenty kinds and fifty roles. NCI Thesaurus is written in OWL-Lite, which makes it amenable to machine-computation and semantic web compliant.

SNOMED-CT

Like NCI Thesaurus, the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) is an ontology as well as a controlled terminology. It was created to capture the language of clinical medicine, including laboratory result contents, procedures, anatomy, and diagnosis. [27] SNOMED-CT is actually a product of two controlled terminologies, SNOMED-RT and Clinical Terms V3, that were merged together beginning in 1999.[28] SNOMED-RT's origin can be traced back to the New York Academy of Medicine meeting in 1928 when it was agreed that diagnosis would become multiaxial; diagnoses would henceforth consist of an anatomic site and a pathologic process. For the 35 years prior to the merger with Clinical Terms, SNOMED was maintained by the College of American Pathologists while Clinical Terms was maintained by the United Kingdom's National Health Service (NHS).[18]

SNOMED-CT's core structure is like most ontologies, even if the names of the structures go by different names. The base units of SNOMED-CT are concepts, descriptions, and relationships. SNOMED-CT contains over 300,000 concepts, 450,000 descriptions[29] and exactly four categories of relationships.

According to the official documentation, a concept is "a clinical meaning identified by a unique numeric identifier (ConceptID) that never changes." A description is a term or name that provides more information about a concept. Multiple terms or names can be assigned to each concept. Relationships join concepts together, and the four types of relationships in SNOMED-CT are defining, qualifying, historical, and additional relationships.[30] Most relationships in SNOMED-CT are defining relationships, which includes the "is_a" superclass to subclass

hierarchy. Key base concepts in the root hierarchy are Clinical finding, Procedure, Observable Entity, Body structure, Organism, Substance, Specimen, Physical object, and Event.[30]

Like the NCI Thesaurus, SNOMED-CT supports and encourages compound term composition in order to express more complex concepts. This capability has been shown to make a significant difference in real-world situations. The Mayo Clinic found that SNOMED-CT could represent their master index of common clinical conditions with only 51% sensitivity without compound term composition but with 92% sensitivity using compound term composition.[28]

Compared to NCI Thesaurus, SNOMED-CT focuses more on the medical domain than on the molecular domain. It aims to be a "comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting." [30] While SNOMED-CT is written in a description logic, it technically machine-computable, the description logic is non-standard for application to the semantic web. Nevertheless, SNOMED-CT is viewed favorably by government agencies such as the National Committee for Vital and Health Statistics (NCVHS).[18]

MeSH

Medical Subject Headings (MeSH) is managed by the National Library of Medicine and backs the popular MEDLINE/PubMed database. This year marks the 50th anniversary of this landmark controlled terminology. It was created in 1960 to replace *Index Medicus*, which had served as the major index for medical journals since 1879.[31] At the time, MeSH was unlike any other bibliographic resource because the NLM intended to make itself the "single subject

authority . . . for both books and periodical articles . . . We take the view that subject cataloging and periodical indexing . . . identical processes." [32]

The first edition of MeSH was strictly a controlled terminology. It was organized into hierarchies and had 4,300 descriptors and 67 topical subheadings,[33] but from its inception MeSH was designed to accommodate new descriptors as a result of scientific discovery as well as to rearrange its hierarchies according to the usage patterns of researchers.[33] Over the last fifty years as the field of ontologies evolved, MeSH has changed its fundamental structure as well from being term-driven to being concept-driven.[32] MeSH developers decided to make this change because they realized MeSH had difficulty expressing relationships between terms and could not attach multiple attributes (definitions) to terms.[34] This transition has resulted in confusion because the same component names were used in the "modern" version of MeSH as in the earlier versions, but with different meanings.

The term-centric version of MeSH only had two core components: descriptors and entry terms. The MeSH descriptor is synonymous with the idea of concepts in the NCI Thesaurus and SNOMED-CT. It is a discrete unit of meaning. For example, "Exercise" is an example of a descriptor. MeSH entry terms are just synonyms of descriptors. They are alternate ways of conveying the same meaning as the descriptor.

The concept-driven version of MeSH, created in 2000, introduced the entities MeSH "concept" and MeSH "descriptor classes." A descriptor class is a group of related concepts, and a concept is a group of related terms.[34] A descriptor is no longer the base unit like it was in the term-centric MeSH. This role is assumed by the concept, with descriptors being reserved for more high-level roles. All of this reorganization was done to make MeSH less redundant, more

flexible, and maintainable. In all there are currently 25,186 descriptors and 160,000 entry terms in the 2010 edition of MeSH. [35]

To the average user, however, these changes are not paramount. MeSH users will interact with two main parts of MeSH: the subject headings themselves and their subheadings (qualifiers). Subject headings are similar to the idea of kinds in the NCI Thesaurus. Sixteen top-level examples exist, and selected examples are Organisms, Diseases, Chemicals and Drugs, and Phenomena and Processes.[35] For example, one MeSH subject heading is named "Kidney Calculi." While this is the preferred term, it has three accepted synonyms (entry terms)—"Kidney Stones," "Renal Calculi," and "Renal Calculus." Qualifiers modify subject headings and provide more detailed information and context about a subject heading. For example, Kidney Calculi is associated with qualifiers such as Diagnosis, Urine, and Microbiology.

ICD-10

The last controlled medical terminology, International Classification of Diseases 10, is the oldest.[18] Its origins date back to the 1850s when it was called the International List of Causes of Death and used primarily to keep track of mortality statistics. It is also often used for reimbursement purposes by governments and health insurance companies. The World Health Organization took over stewardship of ICD in 1948, and the most current revision, the tenth, was released in 1994. It is considered the "international standard diagnostic classification" for epidemiological, health management sand clinical use.[36]

ICD-10 is made up of three volumes. Volume 1 contains the main classifications and is the heart of ICD-10. Volume 2 assists users in coding for ICD, and Volume 3 is an alphabetical index of classifications.

Volume 1 is divided into a series of twenty-one "Chapters," each of which is hierarchical. The twenty-one chapters are grouped into five general categories that William Farr believed should be used to classify diseases: epidemic diseases, constitutional or general diseases, local diseases arranged by site, developmental diseases, and injuries.[37] Half of the chapters follow major body systems such as Diseases of the circulatory system and Diseases of the digestive system.

Within each chapter, ICD-10 reduces clinical conditions to three- and four-character codes arranged by categories and subcategories. For instance, the three-character code "K70" stands for alcoholic liver disease. These three-character classifications are called "core" classifications, and the first character is always associated with a particular chapter. The four-character code "K70.3" stands for alcoholic cirrhosis of the liver. The fourth digit is optional for reporting to the WHO, but should be included if possible.

Compared to SNOMED-CT, ICD encodes medical concepts less granularly.[28] It also does not allow compound term composition that SNOMED-CT, NCI Thesaurus, and MeSH do. ICD-10 prefers instead to explicitly enumerate each possible permutation of one disease. For example, code I60 denotes subarachnoid haemorrhage. Codes I60.1 – I60.6, though, denote subarachnoid haemorrhage in each of the possible arteries: subarachnoid haemorrhage from carotid siphon and bifurcation, subarachnoid haemorrhage from middle cerebral artery, subarachnoid haemorrhage from anterior communicating artery, and so forth. ICD-10 also does not possess synonym support.

Since 1900, the ICD has been updated about once a decade. Between the ninth and tenth revisions, the WHO realized that ICD needed to begin thinking about revising its fundamental structure to facilitate stable and flexible classification for the years ahead.[38] Compared to ICD-9, ICD-10 nearly doubles the number of categories from 4,000 to 8,000.[39] Other changes to ICD-10 are that it uses alphanumeric codes instead of just numeric codes and increases the causes of death list from 72 to 113.[39]

Table 1 summarizes the similarities and differences between each of the four controlled terminologies.

Table 1: Comparison of NCI Thesaurus, SNOMED-CT, MeSH, ICD-10

Feature	NCI Thesaurus	SNOMED-CT	MeSH	ICD-10
Controlled Terminology?	√	√	√	√
Thesaurus?	√	√	√	
Ontology?	√	√		
Hierarchies?	√	√	√	√
Synonym support?	√	√	√	
Concept relationships?	√	√	√	
Compound Term Composition?	√	√	√	
Browsable web interface?	√	√	√	√
Programmatic Access?	√	√	√	
Primary Domain	Cancer	Clinical Medicine	Clinical Medicine + General Biology	Clinical Medicine
Primary Purpose	Unifying molecular and clinical terminology and concepts of cancer	Representing the whole of clinical medicine	Bibliographic retrieval	Reimbursements + Mortality Statistics
Available Formats	OWL-Lite, DL	DL	XML	
Publisher	NCI	SNOMED®	NLM	WHO
Year	1997	1999	1960	1850s
Cost	Free	Free for academic use	Free	Free

Methods

In order to facilitate efficient and flexible searching of gene expression repositories, two barriers need to be addressed; 1) Structured representation of descriptive information and 2) Evaluation of the ability of existing medical terminology to map structured data to a standardized terminology.

The first step was recently completed and is detailed in another paper.[40] In that work, the authors built DSGeo, a previously completed web-based annotation tool that pulls existing samples out of GEO for annotation by a team of physician and student curators. [40, 41] The annotators read through free-text descriptions given about each GEO Sample, associated GEO Series, and associated GEO Datasets to identify salient features of the samples. These features were condensed into tag: value pairs. Annotator consistency, accuracy, and comprehensiveness were compared, and the results demonstrated that manual annotation was consistent among curators, accurately captures descriptive information, and is efficient enough to be performed on a large scale.[41] While the annotations agreed with each other, evaluating how well these annotations map to a controlled terminology will be described in the next section.

1.6 Evaluating Four Controlled Medical Terminologies across Six Clinical Domains

The following process was used to assess the ability of medical terminologies to express clinical terms within GEO. The four previously described medical terminologies were chosen for evaluation: the National Cancer Institute Thesaurus, Medical Subject Headings, SNOMED-CT, and ICD-10. Next, a randomly obtained a subset of GEO samples was generated among six

representative clinical conditions. Third, a three-point scoring system was devised, and scores were calculated for each sample using each medical terminology. Scores reflect how well each terminology represented the clinical terms that describe a given sample. Last, these scores were compared across clinical conditions and across terminologies.

The evaluation was limited to six common clinical domains since evaluating the entire scope of clinical domains contained in GEO would be too broad. The six domains chosen were the ones that had been previously annotated by a team of physicians and university biology students: breast cancer, colon cancer, rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type I diabetes mellitus (IDDM), and asthma.

Once the domains were chosen, the next step was to figure out the best way to structure clinical terms and then try to match it into each of the four ontologies. As mentioned earlier, clinical or phenotypic terms describing each GEO sample was structured in a tag:value format (i.e., "Estrogen Receptor Positive: Yes"). Domain-specific tags were decided upon iteratively by identifying frequently occurring clinical or phenotypic terms describing the samples and from knowledge learned by consulting with domain experts. For example, breast cancer samples have thirty-eight possible tags to be annotated from the descriptions accompanying each sample.

In order to map each tag and value to an appropriate concept or term in each of the four medical terminologies, the corresponding web- based tools that each of these controlled terminologies make available were utilized manually. These search interfaces allow for searching and hierarchically browsing the ontology tree for the desired terms. In order to evaluate how well a terminology could represent a given tag, a three-point scoring system was devised. Tags that exactly matched each term received a point (example: "atopic"). Tags that could be matched after combining two or more atomic concepts using compound term

composition received two points (example: "systemic steroid"). Tags that could not be matched at all received three points (example: "time series").

Because of the structure of DSGeo, one must search for all the studies within a given domain first (i.e., search for breast cancer studies) before searching at the sample level. Studies were selected at random from the resulting studies using a random number generator. Once a study was selected, one DSGeo sample was selected from that study at random, again using a random number generator. Twenty samples were selected for each of the domains without replacement. In all, 480 evaluations were performed across six clinical domains, comparing four controlled terminologies.

The last step was to assign a raw score for each sample based on the actual tags and values that were present. The scores were then normalized to take into account descriptive terms that were actually present. This adjustment was necessary because a substantial amount of clinical information was frequently omitted. For example, if a breast cancer sample contained only five tag:value pairs (instead of the full thirty-eight that DSGeo considers to be a full annotation); its score would be divided by five, for the five tags or values that were actually present in the sample. Ultimately, each sample received a normalized score for tags and values.

Results

Overall, the NCI Thesaurus was the terminology that provided the most comprehensive coverage of all clinical terms used in the evaluated samples. ICD-10 consistently provided the least coverage, which is to be expected since the vocabulary is used primarily for billing

purposes rather than research purposes. Nevertheless, it was formally evaluated since it is in such widespread use.

1.7 Aggregate Scores

The average actual tag scores (Table 2) and value scores (Table 3) across all six domains for all of the samples are listed below:

Table 2
Mean Tag Scores, Combined Over All Domains

Terminology	Mean Score	Ranking
NCI	1.29	1
MeSH	1.81	2
SNOMED-CT	1.87	3
ICD-10	2.96	4

Table 3
Mean Value Scores, Combined Over All Domains

Terminology	Mean Score	Ranking
NCI	1.18	1
MeSH	1.19	2
SNOMED-CT	1.21	3
ICD-10	1.98	4

Comparing the mean actual tag scores with the mean actual value scores, one can see that they are uniformly lower, although the rankings remain the same. This is due to the fact that the tag fields are often phrased to expect "yes" or "no" answers. For example, in the diabetes annotation form one of the tags is "kidney affected?" as opposed to "extent of kidney disease." Because of this propensity for the value fields to contain either "yes" or "no," a score of one was assigned (a perfect match) for all such fields that expected binary answers.

As is almost always the case with large databases, many of the samples had a significant number of unpopulated tags (and therefore values). Because of this, a total tag score was

calculated in addition to an actual tag score, which can be thought of as a reflection of the medical terminology's ability to represent the full set of tags had all of the tags been present.

Table 4
Mean Total Tag Scores, Aggregated Across All Domains

Terminology	Mean Score	Ranking
NCI	1.43	1
MeSH	1.88	3
SNOMED-CT	1.73	2
ICD-10	2.92	4

1.8 Scores According To Clinical Domain

When separating the overall results by clinical domain, the relative rankings changed slightly across clinical domains for both tags and values. For the tags (Tables 5 and 6),

Table 5
Mean Sample Actual Tag Scores by Clinical Domain

Domain	NCI Thesaurus	MeSH	SNOMED-CT	ICD-10	Ranking
Breast Cancer	1.21	1.30	1.37	1.95	NCI, MeSH, SNOMED-CT, ICD-10
Colon Cancer	1.28	1.27	1.28	1.95	MeSH, NCI/SNOMED-CT, ICD-10
SLE	1.23	1.23	1.20	2.20	SNOMED-CT, NCI/MeSH, ICD-10
RA	1.17	1.15	1.26	2.21	MeSH, NCI, SNOMED-CT, ICD-10
IDDM	1.05	1.06	1.05	2.09	NCI/SNOMED-CT, MeSH, ICD-10
Asthma	1.11	1.12	1.11	1.49	NCI/SNOMED-CT, MeSH, ICD-10

NCI Thesaurus still provided the most comprehensive coverage and ICD-10 the least coverage in all clinical domains. The second and third ranks changed from MeSH to SNOMED-CT for rheumatoid arthritis. Otherwise, the relative ranks remained the same.

Table 6
Total Tag Scores by Clinical Domain

Domain	NCI Thesaurus	MeSH	SNOMED-CT	ICD-10	Ranking
Breast Cancer	1.13	1.75	1.97	2.97	NCI, MeSH, SNOMED-CT, ICD-10
Colon Cancer	1.28	1.84	1.89	2.81	NCI, MeSH, SNOMED-CT, ICD-10
SLE	1.44	2.02	2.11	3.00	NCI, MeSH, SNOMED-CT, ICD-10
RA	1.29	1.72	1.64	3.00	NCI, SNOMED-CT,MeSH, ICD-10
IDDM	1.29	1.77	1.86	3.00	NCI, MeSH, SNOMED-CT, ICD-10
Asthma	1.30	1.75	1.75	3.00	NCI, SNOMED-CT/MeSH, ICD-10

For the values separated by clinical domain (Table 7), NCI Thesaurus was either the most robust or tied for being the most robust, and ICD-10 was the least robust in all cases. It is worth nothing that even though the NCI Thesaurus was designed as a cancer terminology, it still had the best performance, even in non-cancer domains.

Table 7
Mean Sample Actual Value Scores by Clinical Domain

Domain	NCI Thesaurus	MeSH	SNOMED-CT	ICD-10	Ranking
Breast Cancer	45/38	76/38	68/38	110/38	NCI, SNOMED-CT, MeSH, ICD-10
Colon Cancer	41/32	62/32	50/32	84/32	NCI, SNOMED-CT, MeSH, ICD-10
SLE	51/32	58/32	52/32	96/32	NCI, SNOMED-CT, MeSH, ICD-10
RA	27/18	32/18	33/18	54/18	NCI, MeSH, SNOMED-CT, ICD-10
IDDM	29/21	33/21	33/21	63/21	NCI, SNOMED-CT/MeSH, ICD-10
Asthma	30/18	39/18	36/18	66/18	NCI, SNOMED-CT, MeSH, ICD-10

1.9 Unrepresented Tags

Despite efforts to find a suitable mapping for each term in a given terminology, numerous terms could not be represented in just one terminology. A selected sample is provided in Table 8.

Tags that could not be represented in one terminology could be represented using a different terminology (usually the NCI Thesaurus.) Overall, 100% of the tags could be represented if searching across all four of the terminologies was allowed.

Table 8
Selected Problematic Tags (Tags with a score of 3)

Tag	Domain	Terminology
Time series	Breast Cancer	MeSH, SNOMED-CT
Disease state	Breast Cancer	MeSH
Genetically Modified	Breast Cancer	SNOMED-CT
Her2/Neu	Breast Cancer	ICD-10
CD Class	RA	NCI
Diagnosis Criteria	RA	SNOMED-CT
Treated	Asthma	MeSH
Atopic	Asthma	ICD-10

Table 9 shows that the NCI Thesaurus had the least number of unrepresented tags (tags that received a score of 3), and Table 10 shows the result of three pairwise Chi-Square Tests comparing NCI Thesaurus with the other three terminologies with regards to the number of unrepresented tags. From Tables 2 and 4, one can see that the NCI Thesaurus ranked first in both actual and total tag scores. Because of this, pairwise comparisons using MeSH, SNOMED-CT, and ICD-10 as the primary comparator were not needed. Table 10 demonstrates that the number of unrepresented tags is significantly less using the NCI Thesaurus compared to any of the other terminologies evaluated.

Table 9
Comparison of Unique, Unrepresented Tags by Terminology

Terminology	Number of Unrepresented Tags (87 Unique Tags)
NCI Thesaurus	6
MeSH	30
SNOMED-CT	17
ICD-10	81

Table 10
Pairwise Chi-Square Tests

Comparisons	p-value
NCI Thesaurus versus MeSH	<0.001
NCI Thesaurus versus SNOMED-CT	0.0138
NCI Thesaurus versus ICD-10	<0.0001

Discussion

The first part of this project evaluated four medical terminologies' ability to cover the breadth of descriptive information used to describe GEO samples. The primary observation that became clear when analyzing the results was that descriptive terms used by researchers vary widely with regard to quality, organization, and comprehensiveness. At present, the contents of descriptive fields deposited in GEO are entirely at the researchers' discretion. Because of this,

there was concern that evaluating samples containing incomplete information might influence the performance of the terminologies being evaluated. Indeed, when comparing actual tag scores, (which disregarded unused tags) with total tag scores, the rankings were slightly different. NCI Thesaurus still had the best score, but MeSH and SNOMED-CT alternated between second and third place.

NCI Thesaurus consistently provided more comprehensive coverage of descriptive textual information. This was likely due to the NCI Thesaurus' higher concept granularity and larger number of concepts terms compared to some of the other terminologies, especially ICD-10. Because of this, NCI Thesaurus was more flexible and allowed for the creation of composite concepts that were not expressible in other controlled terminologies, an idea called compound term composition. An example of this was the term "other affected organs". This was a tag in the type I diabetes annotation form, and the NCI Concept IDs C17649 + C64917 + C13018 could have been combined to compose this term, while SNOMED-CT, MeSH, and ICD-10 were unable to represent it precisely. The primary advantage of compound term composition is that it averts the need for a large, monolithic and strictly hierarchical taxonomy. Instead, large numbers of valid concept or term indexes can be created with the added benefit of occupying minimal storage space. [42]

The fact that ICD-10 did not perform robustly can be attributed to this terminology's designated purpose - ICD-10 was developed primarily for diagnostic and reimbursement purposes. [43] As a result, it summarizes a clinical encounter or condition in one atomic term, such as "asthmatic bronchitis NOS." This works adequately for reimbursement, but can not provide the level of detail that is often necessary in translational research. Nevertheless, ICD codes remain commonly used in clinical research. ICD-9, the previous version of ICD-10, was found

in one study to only have 37% coverage for clinical terms, whereas SNOMED-CT successfully mapped 92% of terms in a different study. [27] This is likely because in addition to not supporting compound term composition, ICD also does not support synonyms for common terms.

Even though it would have been possible to capture all descriptive terms by combining the four terminologies being evaluated, no single terminology provided 100% coverage. One way to achieve broader coverage would be to use the Unified Medical Language System (UMLS), a composite of over 150 medical terminologies. This stands the best chance of representing the most number of clinical terms. The UMLS Metathesaurus would have helped in difficult cases because it is not only comprehensive, but it also maps synonymous concepts from disparate terminologies into one UMLS identifier while still preserving each individual terminology's meaning for the concept.[18]

Chapter 2: Comparative Effectiveness of Free-text Versus Faceted Search for Retrieving Relevant Biological Samples from a Gene Expression Database

A fundamental benefit of ontologies is that they provide a consensus standard terminology for a domain. The results above indicate that the NCI Thesaurus may be the ideal controlled medical terminology to use for expressing the clinical terms used inside GEO, at least for the six domains sampled. However, confirming that the tag: value pairs created by the human annotators can be successfully mapped to the NCI Thesaurus' standardized terminology, does not yield any practical benefit unless these annotations are leveraged to enable the efficient searching of samples.

Beyond standardizing terminology, ontologies provide two other concepts that depend on this property: concept hierarchies and concept properties. Concept hierarchies specify an atomic term's relationships in a parent-child or sibling-sibling fashion to other terms. This tree-like structure enables terms to be searched with varying levels of granularity. Concept properties describe the relationships between individual concepts. Together these two properties give ontologies the potential to represent phrases and terms in combinations that may not have been envisioned by the original designers. This way knowledge does not have to be exhaustively enumerated in order for the ontology to be comprehensive.

Concept hierarchies and concept relationships lay the ground work for flexible and standard knowledge representation, and these benefits will extend to descriptive textual

information inside gene expression databases as long as the descriptive textual information is expressible using the ontology.

As GEO's size grows, researchers looking for samples with similar clinical characteristics to their own will be more likely to find them and conduct studies without having to collect samples anew. [11, 40] This ability to reuse data carries the potential to greatly accelerate the rate at which new discoveries are made.

However, once descriptive textual information inside GEO is expressed using controlled terminology, a search strategy that leverages the metadata that has been condensed into tag: value pairs is still needed. Two generally accepted search strategies are free-text search and faceted search, and the benefits and drawbacks of each are described next.

Background

2.1 Free-text Search

Free-text search is the search method with which most computer users are familiar. Google.com, the world's busiest website, [44] is the most well-known example of free-text search today. In the biomedical domain, free-text search is a near-universal feature of all the major online medical databases, including Ovid, Pubmed, Web of Science, and Google Scholar.

The basic idea is straightforward: a user simply types his/her question into an empty text field and presses the <Enter> key. The primary advantage of free-text search is the lack of constraints on the search parameters. Some users prefer this freedom since they can type any text string using 'regular' English questions, which makes the interaction similar to having a

conversation with the computer. Moreover, free-text search also allows users to express their queries using personally synthesized information. Users are able to interface with the computer using their preferred terminology instead of being forced to examine long lists of select boxes that contain a structured set of terms. The primary disadvantage of free-text search is that little guidance is offered when searches return unsatisfactory results,[45] and users do not get a general sense of usable information inside the search space. What typically happens as users gain experience, though, is that they start reducing their natural language search queries to a series of keywords when they realize that most free-text search engines do not understand English grammar. Many free-text search engines actually provide advanced features that allow the user to limit the search space and/or express terms in semi-structured ways. However, a study of over 60,000,000 search engine queries concluded that many users either do not know how to use these advanced features or are not motivated to use them, resulting in an average query length of only 2.4 words.[46]

The largest online bibliographic resources have their own dedicated free-text search interfaces and employ proprietary free-text search algorithms. However, numerous small- and large-scale microarray-focused projects are powered by the MySQL database management system. This list includes software such as the analysis and visualization tool BASE, the cDNA database NOMAD, and the general gene expression databases maxdSQL and LIMas.[8, 47, 48] Regardless of the backend used, free-text searches usually fall into four general categories, all of which MySQL supports.

The four major types of free-text search available in MySQL are: string matching, natural language, Boolean, and automatic query expansion. String matching is the most basic free-text search capability, and it matches raw character sequences in the query to the character

sequences in the database. Natural language free-text search attempts to emulate human natural language queries.[49] Natural language search works using a similarity algorithm between the query and the target columns to be searched in the database. Unlike string matching, partial matches are allowed, and matches are displayed in descending order by relevance score.

The relevance score is calculated by using a vector-space model that has as many dimensions as there are unique words in the columns to search over the entire database. [49, 50] Regarding GEO sample searches, the title, description, and abstract are the searchable columns. Each unique word in these three searchable columns within each sample is assigned a weight that together forms a vector. If there are 300 samples, then there are 300 sample vectors. The terms in each query also form a vector. The sample vectors that are nearest in vector-space to the query's vector are considered to be the most similar, and these are the samples that are returned.

Boolean search is the third major type of free-text search that the search tool implements. In this type of search, the operators "+","-", "*", "~", and "()" are available to give the user finer-grained control than string matching. [51] The "+" means that the word must be present in the sample, "-" means that it must not be present, and "~" means that the word is preferred not to be present, but if it is, matching samples appear lower in the results. The parentheses enable grouping the search terms using standard Boolean order of operations, and the "*" is a wildcard which can be useful for matching multiple word stems.

The last type of free-text search available is automatic query expansion, also known as automatic relevance feedback. Whereas natural language queries tend to work better on relatively long queries, [49] automatic query expansion is useful for improving the number of matches for short queries, which generally contain little information content.[52] Automatic query expansion was developed more than forty years ago and [53] assumes that the reason that the query

is so short is that the user is relying on implied knowledge. This search method tries to make explicit this implied knowledge by making educated "guesses" and adding these terms to the original query.

The algorithm consists of making two queries for every query. The first involves the original query, with subsequent return of matching documents. The matching terms in the retrieved documents are given more weight and a second query is enabled using the original query plus the additional terms. This expanded search is expected to return more relevant documents, based on the initial query.

2.2 Faceted Classification

At the most basic level, classification can be thought of as "the meaningful clustering of experience." [54] Classification attempts to structure knowledge so that it is more accessible and flexible. The underlying representations are usually that of a hierarchy, tree, or a faceted system—and often a combination of all three.

Hierarchies attempt to put all members of a given domain in its proper place with respect to each other and the world. Aristotle believed that all of nature functions as one unified whole, and part of the reason that humans innately try to order the world around them is that only when an entity is properly classified does one truly *know* it.[54] In addition, for a classification scheme to be considered a hierarchy, it should possess several properties. They are: inheritance, transitivity, systematic and predictable rules for association and distinction, mutual exclusivity, and necessary and sufficient criteria. Inheritance and transitivity refer to property of every subclass and every subclass's subclass possessing at least the properties of its parent class. The requires

for hierarchies to possess systematic and predictable rules for association and distinction and necessary and sufficient criteria means that there should be formal criteria for where to place entities in a hierarchy and formal criteria for testing class membership. Last, mutual exclusivity only applies to pure hierarchies. In a pure hierarchy, a given entity can only belong to one class, and multiple hierarchies are not allowed. These formal properties confer several advantages to storing knowledge in hierarchical form: comprehensiveness of information, economy of notation, and inference.[54]

Trees are quite similar to hierarchies, but theorists consider them distinct from hierarchies because trees do not display the inheritance property. Like hierarchies, trees do progressively subdivide its members as one goes deeper into the tree. However, a tree does not assume an "is-a" relationship between members and submembers. The ordering of the members in a tree structure is done to distribute members along one specific type of non-inherited relationship.[54] Using GEO as an example, the primary navigation links on the homepage of the GEO repository could be represented as a tree with one level. The relationship modeled would be a part-to-whole relationship, and such a tree's structure would resemble a table of contents:

GEO

- Home
- Search
- Site Map
- GEO Publications
- FAQ
- MIAME
- Email GEO

The terms "Home", "Search", "Site Map", etc. are not subclasses of GEO; they are more accurately described as parts of GEO. Further, sibling terms such as "FAQ" and "MIAME" do not share common traits with each other, as one would expect in a hierarchy. Yet the relationship of each of these entities is that they are all individual features that GEO offers. Despite their differences, in daily usage, however, the terms "hierarchy" and "trees" are often used synonymously.

Faceted classification involves recording observations about an entity from a number of different angles. Taken together, facets characterize information about items in a collection.[55] Distilling the descriptive textual information inside GEO samples into tag: value pairs like the human annotators did in the first section of this thesis is an example of faceted classification. Some synonyms for facet are "perspective", "aspect", or "category."[56] [55] The credit for formalizing the notion of faceted classification is usually given to S.R. Ranganathan, who did so in India in 1967.[54]

In its simplest form, faceted classification is quite different from hierarchies and trees because each facet is regarded as completely independent from other facets.[54, 57] This lack of structure (relative to hierarchies) is regarded as one of faceted classification's strengths because it enables a dataset to be viewed from multiple perspectives.[54] For example, viewing a biological sample in GEO from different perspectives means that it can be understood in terms of the different roles that the sample might play. One role would be the sample's role serving as a control to an experimental sample. Such a facet might be "study_group." Another way to view the same sample might be for its role being run on a certain piece of hardware. Such a facet might be named "platform." Yet another way to view the sample is as part of the group of samples that were obtained within the last month. Such a facet could be named "date_obtained."

These facets (study_group, platform, date_obtained) may or may not overlap or have an identifiable relationship with each other, but in pure faceted classification the distinction is not important.

The important point is that facets can be freely combined in myriad ways, depending on the vantage point that a researcher wants to take. This mixing and matching of facets is formally known as postcoordination.[54] Usually, faceted search interfaces allow users to combine multiple facets to progressively refine a set of matches in a drill-down fashion. [46]

Faceted classification was the chosen representation scheme for sample annotations because of its flexibility and hospitality, as well as because faceted classification has previously been shown to be effective and easy to comprehend.[46, 54] "Hospitality" in knowledge representation refers to the ability of a classification method to accommodate new terms.[54] Since no inherent relationship or order between facets exists, new facets can be added in the future without the need to rearrange the previous structure. The notion of flexibility encompasses the fact that faceted classification does not require any unified theory about a domain, nor does it depend on having complete knowledge of a domain like hierarchies do. New facets may be added as they are discovered since a facet is simple observation or fact about an entity without any implied meaning from its position in the list of facets. [54]

The properties of hospitality and flexibility make faceted classification well-suited for representing the genomics domain because of the rapid pace at which new information is discovered and because of the myriad combinations of criteria that researchers might use to locate samples in GEO.

2.3 Prior Implementations of Free-text & Faceted Search Tools in GEO

The team responsible for GEOmetadb has produced the most comprehensive effort at making the metadata inside GEO more easily accessible for ordinary biologists, statisticians, and bioinformaticians.[8] Their approach has been to create a powerful, web-based search tool that combines elements of both free-text and faceted search.

GEOmetadb allows one to search at the GEO at multiple levels of detail, including the GEO Dataset, GEO Series, and GEO Sample levels. At the sample level, one can search over thirty different fields (facets) of metadata, including the sample "characteristics" field of the MIAME specification. The query tools provided by GEO itself also feature the same basic capabilities; the difference with GEOmetadb is that more fields can be searched with more specificity. In addition, the tool also supports querying within results, creating lists, personalized display options, and downloading results.[8]

The primary limitation of GEOmetadb for the purpose of identifying samples according to clinical characteristics is that GEOmetadb only supports searching the characteristics fields using free-text.

Because of the high-quality, detailed annotations contained in DSGeo and their organization into tag: value pairs, [40, 41] an opportunity exists to identify samples at an even more granular level. In order to do that, however, a tool that takes advantage of the tag: value pair structure needed to be built.

Methods

The methods in Chapter 1 were aimed at finding a terminology that could represent the various clinical terms that were used to describe GEO samples. Identifying a suitable terminology was necessary in order to make sure that the terms that the human annotators were curating with came from a standard vocabulary.

Once all of the knowledge about the samples was condensed into tag: value form, the next step was to devise a suitable scheme that would enable accurate and efficient retrieval of the samples. Given the annotations' organization into tag: value pairs (e.g., disease state = rheumatoid arthritis) there were two obvious search strategies to employ: faceted search and the traditional free-text search.

2.4 Building Geosearch: A Faceted Search and Free-Text Search Comparison Tool

The two search strategies, faceted search and free-text search, were implemented with five general features. These features are listed in Table 11. The first feature is the ability to add and annotate samples, which was useful when adjustments to the samples such as adding annotations or correcting errors were needed. The second feature is the ability to browse through all of the samples in the database. This way, on one screen, the samples' clinical contents could be displayed as they are in GEO alongside their annotated tag: value pairs.

Table 11: Faceted Search & Free-text Search Tool Requirements

General Requirements	Details			
Sample Addition and Annotation				
Sample Browser				
Faceted Search	Provides AND/OR functionality	Allows simple linear, chained queries	Allows for operator precedence, i.e., simulates parentheses	
Free-Text Search	Can search different sample fields: title, abstract, description	Can free-text search multiple different ways: string-matching, Boolean, natural language, automatic query expansion		
Statistics	Precision	Recall	F-score	Search Time

Regarding the faceted search feature, the tool needed to provide basic Boolean operator functionality so that simple clauses could be joined together with an "AND" or an "OR" conjunction. A significant proportion of faceted searches are likely to be simple, requiring nothing more than a series of atomic clauses joined together by "AND" / "OR." Simple clauses are read left to right with any "ANDs" and "ORs" evaluated in the order that they come. An example of such a query is the search question, "Identify all of the samples that came from female SLE patients." This translates into searching for the following facet, (tag) : value pair "gender = female" AND "disease state = systemic lupus erythematosus."

For more complex queries, however, the tool needed to support operator precedence (by default, OR is always evaluated prior to AND). That is, the tool needed to support grouping

simple clauses by priority. I defined a complex query to be search criteria that required deviating from the standard order of operations in which clauses are evaluated in order to express the searcher's intended meaning. In other words, the tool needed to support "parentheses."

For example, consider the question, "Find samples that came from stage 4 breast cancer patients who were treated with tamoxifen, as well as samples that came from metastatic colon cancer patients." This translates into: ("stage = 4" AND "disease state = breast cancer" AND "treated = tamoxifen") OR ("disease state = colon cancer" AND "metastatic = yes"). If this query was processed strictly from left to right, the result set would be different. The matches would answer the question "Locate samples from colon cancer patients or stage 4 breast cancer patients who were treated with tamoxifen. Of those patients, find those who also had metastatic disease."

The way that the tool simulates parenthetic grouping of clauses is by having the user enter in simple clauses (clauses in which the correct meaning can be constructed by processing tag: value pairs from left to right), saving each one as a subquery, and then joining the subqueries together. The results of these joined subqueries can themselves be saved as another subquery, etc., so that in theory an infinite number of nested queries can be computed.

Concerning the free-text search feature, the goal was to provide basic free-text search functions. The first function needed is the ability to search either all fields or just selected fields. In the case of GEO samples, the tool searches the title, abstract, and description fields or just the description field itself.

As mentioned earlier, searching in various fields is relevant because descriptive textual information in GEO is scattered across more than one field across GEO Samples and GEO Series. In general, descriptive terms are most often found in GEO Series titles, GEO Series

summaries (identical to the abstract from the published study), and descriptions within GEO samples. Searching all three fields (title, abstract, and description) is often necessary in order to piece together enough clinical information, while for other samples, investigators followed GEO's recommended protocol and deposited descriptive terms only in the sample description field, thus making a search through all three fields unnecessary. To accommodate this flexibility, the free-text search tool was programmed to provide the option of searching through all relevant fields or just the sample description.

Giving the user four different free-text search choices (string matching, natural language, Boolean, automatic query expansion) and two different columns in which to search (title-abstract-description, description) was done to help ensure that any difference between faceted search and free-text search performance would be more likely due to the inherent differences in faceted versus free-text searching itself, rather than on any one specific implementation of free-text searching.

The fifth software feature is the results module. This module saves the search history and [54]faceted search. The statistics calculated are precision, recall, F-score, and search time for each individual question. The average precision, recall, and F-score across all search questions were also calculated.

Implementation

Django (version 1.1 http://www.djangoproject.com/), a model-view-controller toolkit using Python, was chosen as the web development framework, and MySQL Community Server 5.1.46 (http://www.mysql.com/downloads/mysql/) was chosen as the database server. Both

pieces of software are open source. An additional benefit of using Django to write the search tool was that DSGeo was also written in Django. As a result, integrating the search tool with DSGeo would be easier in the future.

2.4 Geosearch: Evaluation

Faceted search was evaluated and compared to free-text search. This evaluation was performed utilizing a randomly selected subset of annotated samples, comparing each search strategy's performance in regards to answering multiple questions of varying complexity.

First, 300 previously annotated samples were loaded into the search tool. Each sample number's title, description, abstract, and annotations were all imported. Next, a list of search questions was compiled (Table 12). Three types of questions were chosen: simple, compound, and complex. They represented increasingly specific (and presumably harder to answer) questions that a researcher might ask.

A simple question was defined as a question that only required one tag: value pair to answer correctly using faceted search. The first question in Table 12 provides an example of a simple question: "Locate the samples that were obtained from the condition breast cancer." A compound question was defined as a search question that required sequentially chaining together multiple tag: value pairs in order to correctly answer the question. Question 4 provides an example of this type of search question.

A compound question that required multiple tag: value pairs chained together but evaluated in a non-linear sequence, i.e., evaluation according to parentheses, was termed a complex question. Question 7 is an example of this type of question. If this question was

answered by joining together tag: value pairs sequentially from left to right, the resulting matches would be incorrect.

The last two questions were designed to illustrate one of the main benefits of faceted search: guided navigation. Organizing data with facets allows the user to see an overview of the data by categories. This is helpful when the user is interested in browsing to see what the database contains and gives the user an idea of what types of questions that he/she could realistically ask. Answering these types of questions would be almost impossible using free-text search.

Table 12: List of Search Questions

Question Number	Question	Question Type
1	Locate the samples that were obtained from the condition breast cancer.	Simple Query
2	List samples that have a p53 mutation	Simple Query
3	Which samples came from African-Americans?	Simple Query
4	List the samples that came from breast cancer patients with a positive family history	Compound Query
5	List the samples that came from either breast cancer or colon cancer that were metastatic	Compound Query
6	Locate specimens that came from the mononuclear cells of 7-year-old insulin dependent diabetics who have been treated with insulin.	Compound Query
7	List samples that came from systemic lupus erythematosus patients who were treated with po and iv steroids or that came from RF negative rheumatoid arthritis patients	Complex Query
8	List samples that came from Duke Stage B or C patients who are either female or Caucasian.	Complex Query
9	List the tissue types that the biological specimens came from	Not included in statistical analysis
10	List all the available diseases represented in the database	Not included in statistical analysis

Once the question list was made, test users were gathered. The test subjects were all post-doctoral fellows (MDs or PhDs) who are familiar with clinical terminology, but less so with terms related to genomic research. Each volunteer was given a brief tutorial on the search tool. They were instructed on the four types of free-text searching available as well as on the capabilities of the faceted search tool to not only chain together tag: value pairs with Boolean operators, but also to save partial result sets and chain those together in a recursive fashion. Each test user was given the list of ten search questions and instructed to first answer each question using the free-text search tool and then using faceted search.

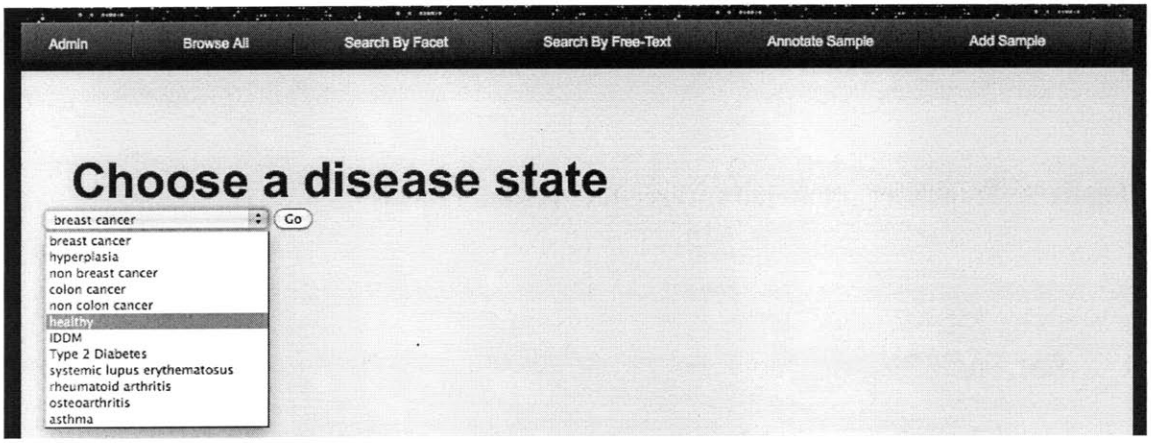
Regarding free-text search, users were allowed to specify any combination of search terms, free-text search method (string matching, natural language, Boolean, or automatic query expansion), and search fields that they wished (Figure 2). They were allowed to iterate until they believed that their result set was the best that they could obtain using free-text search. This final result was recorded along with the time that it took to obtain this result set.

Figure 2: Geosearch Free-text Search Interface

Next, a similar process was used for faceted search (Figure 3). The user was asked to locate relevant samples using faceted browsing. A query was a "list of facet-value pairs that jointly specify the required properties of a matching document." [46] Geosearch users were allowed to try as many combinations of tag: value pairs and inspect the result sets for accuracy until they were satisfied. The matching samples were recorded along with the time that it took to obtain these results. The user then proceeded to answer the next question in the list, and the process was repeated until all questions had been answered. For each question, each user always searched using free-text before being allowed to use the faceted search interface. Users were given multiple tries to come up with the correct result sets because of the learning curve of using

a new interface. Moreover, in real world situations getting the right answer immediately is not critical, while being able to retrieve relevant biological specimens is.

Figure 3: Geosearch Faceted Search Interface



At that point, the user's results and statistics were displayed (Figure 4). For each question, there was a set of "correct" matches. The correct answers were verified manually beforehand. The software tabulates each user's free-text and faceted matches and could classify each matched sample as a true positive, false positive, or false negative. (When describing search performance, true negatives are not usually tallied because one is typically interested in measuring performance based on the ability to find items that truly do match as opposed to items that truly do not match.) From the true positive, false positive, and false negative values, precision, recall, and F-scores could be calculated for each search question. In addition, the time spent searching was also calculated. The overall average precision, recall, and F-score for each person were also displayed. A visual representation of the testing process is shown in Figure 5.

Figure 4: Geosearch Results Module

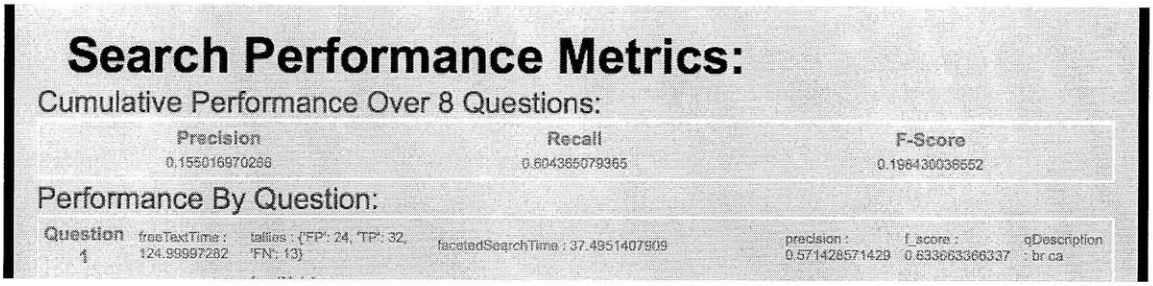
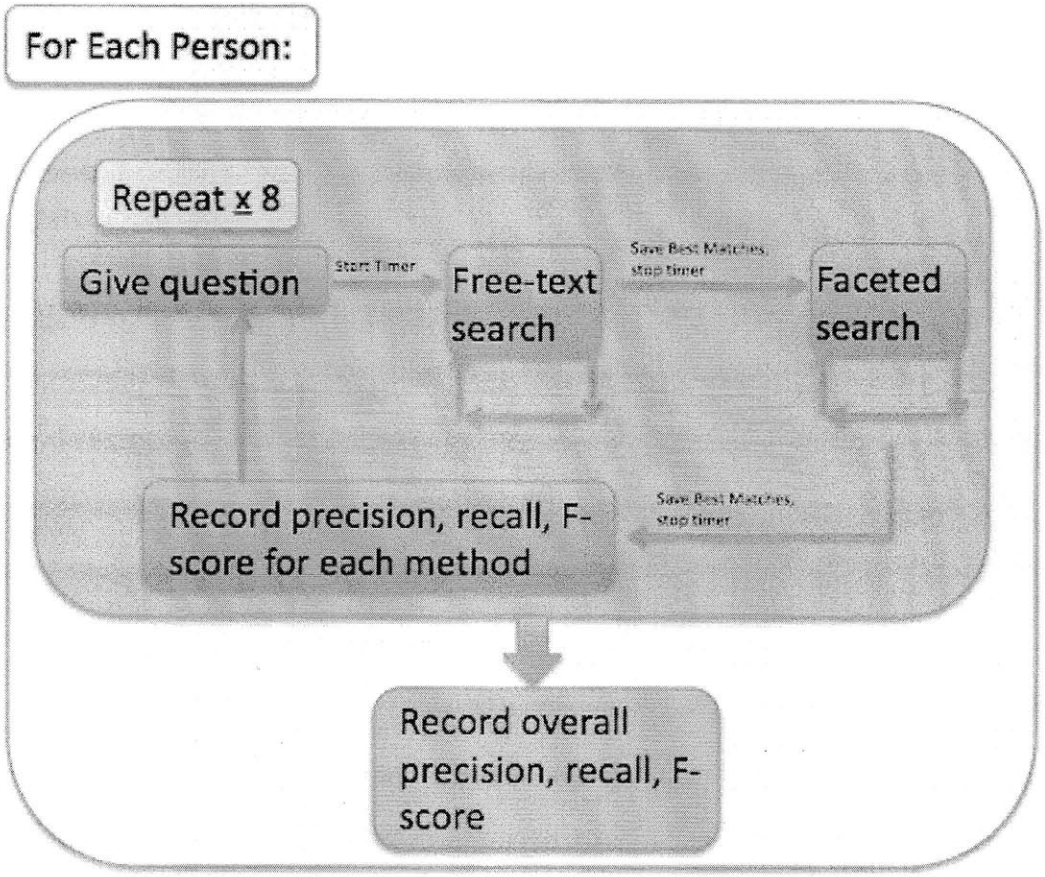


Figure 5: Flowchart of Faceted vs. Free-text Testing Process



Results

The two search strategies were compared using paired t-tests. Outcome measures included precision, recall, and F-scores.

Precision is also known as positive predictive value, and can be represented as:

True Positives / (True Positives + False Positives)

In this project, the positive predictive value measures the likelihood that a returned matched sample actually fits the query.

Recall is also known as sensitivity, and can be represented as:

True Positives / (True Positives + False Negatives)

In this project, recall refers to the ability of the tool to find all possible matches, even at the expense of introducing some inaccurate matches.

The F-score is an information retrieval statistic that takes into account both the precision and the recall. The F-score's range is 0 to 1, with 1 being the highest. It can be represented as:

2 * precision * recall / (precision + recall)

It is essentially a weighted average of the precision and recall, and it gives a better global view of search performance. [58]

2.6 Overall Performance of Faceted Search versus Free-text Search

The average precision, recall, F-score, and search time for the entire search question list when using free-text search is shown in Table 13. One can immediately see that the performance was not optimal.

Table 13: Average Performance for All Search Questions Using Free-Text Search

Person	Precision	Recall	F-score
1	0.257	0.481	0.287
2	0.233	0.576	0.303
3	0.194	0.763	0.256
Overall Averages	**0.228**	**0.607**	**0.282**

In table 14, one can see that, given enough time, each search tool user was able to locate matching samples with perfect precision, recall, and F-score. This was not always done on the first try, especially for complex questions. This was in part due to the fact that there was a learning curve for understanding how to navigate and join queries using the tool.

Table 14: Average Performance for All Search Questions Using Faceted Search

Person	Precision	Recall	F-score
Overall Averages	**1**	**1**	**1**

At first glance, the precision, recall, and F-scores may seem spuriously high, but the nature of the faceted searching tends to result in matches that can quickly be seen as "all right" or "all wrong". For example, consider a complex search task such as "List samples that came from either Duke Stage B or C patients who are either female or Caucasian." Using faceted search, if one chooses "Duke Stage = B", followed by "Duke Stage = C", and then decides to join these two simple clauses together with "AND" the result set will immediately be empty. Some users initially made this mistake because in everyday conversation, "and" has two meanings: set addition ("Give me a red and a blue ball") as well as set intersection, "Give me the red- and blue-colored shirt." In Boolean logic, however, "OR" denotes set union. The test users quickly noticed mistakes and corrected themselves when their Boolean operations resulted in unintended matches.

2.7 Overall Performance Grouped By Search Complexity

Table 15 and Figure 6 report the performance in numeric and graphical form, respectively, of free-text search in answering three increasingly difficult question types. Free-text search clearly performed best for simple questions in terms of precision and recall. As a reminder, a simple question only returns samples identifiable with a single tag: value pair, e.g. "Find diabetic patients." Once questions contained more criteria, however, free-text search's precision and recall decreased sharply. For complex questions, precision was less than 15% and recall was less than 27%. The F-scores mirrored precision and recall.

Table 15: Average Performance by Question Type Using Free-Text Search

Person	Question Type	Precision	Recall	F-score
1	Simple	0.380	0.652	0.476
	Compound	0.151	0.700	0.230
	Complex	0.138	0.258	0.153
2	Simple	0.434	0.467	0.445
	Compound	0.143	0.633	0.208
	Complex	0.160	0.276	0.169
3	Simple	0.378	0.667	0.479
	Compound	0.023	0.833	0.044
	Complex	0.116	0.800	0.251
Overall Averages	**Simple**	**0.397**	**0.595**	**0.467**
	Compound	**0.106**	**0.722**	**0.161**
	Complex	**0.138**	**0.444**	**0.191**

Figure 6: Free-text Search Precision, Recall, F-Scores by Question Type

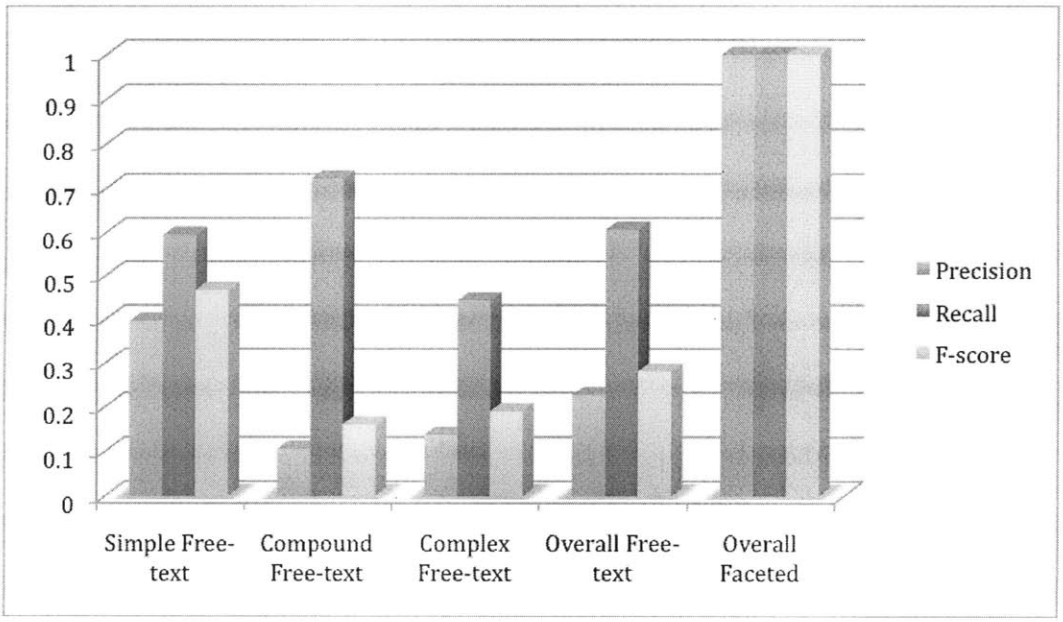


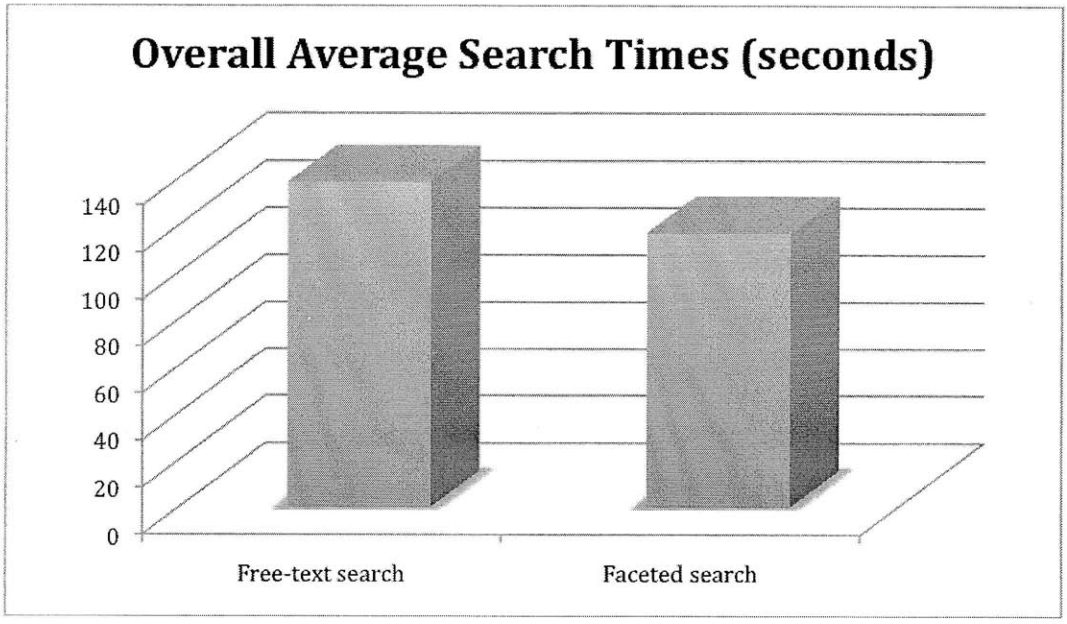
Table 16 and Figure 7 show the search times using each search strategy. It is not surprising that as question complexity grew, the average time that it took to find matches increased (Table 16). For simple questions that can be answered using only one tag: value pair, faceted search was far faster than free-text search. For more complex questions, however, the relationship is less clear. However, faceted search maintained better precision, recall, and F-scores.

The overall search time using faceted search averaged 116.7 seconds, while using free-text search averaged 138.4 seconds (Figure 7). While the search times were noted, two factors make them poor candidates for rigorous analysis. The first factor is that how long each person wanted to search before he/she was satisfied with the matches is subjective. The second factor is that search times that differ on the order of a few minutes are rarely consequential in practice.

Table 16: Average Search Times Comparison by Question Type

Person	Question Type	Free-text Search Time (seconds)	Faceted Search Time (seconds)
1	Simple	52.6	12.3
	Compound	72.1	113.3
	Complex	131.5	309.6
2	Simple	217.1	16.5
	Compound	62.1	90.9
	Complex	93.3	263.89
3	Simple	87.2	9.5
	Compound	264.1	50.2
	Complex	265.6	183.6
Averages By Type	**Simple**	**118.9**	**12.8**
	Compound	**132.8**	**84.8**
	Complex	**163.5**	**252.4**
Overall Averages		**138.4**	**116.7**

Figure 7: Comparison of Overall Average Search Times



Finally, results were analyzed by performing paired t-tests for each performance metric. The objective was to see if precision, recall, F-score, and search time were statistically different using free-text compared to faceted search. Table 17 demonstrates that faceted search performed significantly better than free-text search, producing better precision ($p=0.001$), recall ($p=0.042$), and F-scores ($p=0.0003$). With regard to search time, faceted search was also faster. However, paired t-test analysis does not indicate statistically significant time savings for either search method.

Table 17: Paired T-test Results Free-text Versus Faceted Search Overall Performance

Performance Metric	T-score	p-value
Precision	**42.050**	**0.001**
Recall	**4.748**	**0.042**
F-score	**52.044**	**0.0003**
Search time	**0.389**	**0.734**

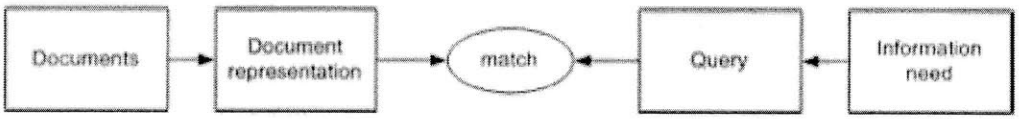
Discussion

Most information retrieval techniques fall along a continuum of free-text search on one end and strictly controlled vocabularies on the other end. [59] The second part of this research looked at both extremes and evaluated the hypothesis that the information needs of biomedical researchers who use a large gene expression database are better filled using faceted browsing than conventional free-text searching.

In classic information retrieval, the process is typically split into five parts: the search space (documents), document representation, the search results (matches), the query, and the original information need of the information seeker (Figure 8). [60]

This search tool evaluated the robustness of the second component of the classic information retrieval model, document representation. The results show that if one has the resources to condense the biomedical information contained in disparate GEO sample fields into tag: value pairs, then searching along those facets gives more accurate results than free-text search.

Figure 8: Traditional Information Retrieval Model



(From Garcia E, MA S. User Interface Tactics In Ontology-based Information Seeking. PsychNology Journal 2003;1:242-55.)

One observation that the testers made concerned the length of the facet lists as more and more facets are added. Deciding how many facets to present to users, where to present them (such as off in a sidebar), and when to present them is a continual challenge facing faceted search researchers. One project whose primary aim is to optimize the presentation of facets is UC Berkeley's open source Flexible Information Access Using Metadata in Novel Combinations (Flamenco) Project. Sponsored by the National Science Foundation, the project tries to help "users move through large information spaces without feeling lost" by using faceted search.[61] Towards this end, project designers have tried displaying breadcrumbs, a "sequence of actions that a user has done within the query session,"[55] displaying a hand-selected group of facets that they believe users will find the most helpful, displaying the most frequently selected facets, and displaying facets in alphabetical order.[46] Koren, et al., have devised yet another method of getting the most relevant facets to each user by creating what they term "personalized interactive faceted search." Their method involves using explicit user ratings of facets to deliver the most

probabilistically relevant facets to users.[46] Geosearch displays facets in the chronological order in which they were added to the database.

Limitations

A main limitation of this research can be attributed to the less than optimal quality of descriptive information deposited within GEO. Without comprehensive information being supplied by researchers who upload samples to GEO, it is more difficult to evaluate the performance of search tools since poor results could be the result of simply not having enough information.

A second limitation is that the GEO browser already interfaces with NLM's rich resource of query tools, including MeSH. However, the GEO website's browser is not designed to find information at the sample level like the search tool in this project.

A third limitation concerns the scalability and sustainability of the manual annotation process as GEO continues to grow. If the accuracy that only human annotation can provide is the overriding concern, then recruiting the biomedical informatics community at large to annotate could be done. The benefits of this approach are the low cost and a perpetual pool of annotators. The main drawback would be the reduced ability to guarantee quality annotations. Requiring user registration, implementing a zero to five-star rating system for annotations, and allowing users to flag problematic entries would enforce quality. Ultimately, offloading the work to the biomedical informatics community would be an exercise in trust.

A fourth limitation concerns the backgrounds of the test users. Most of them had knowledge of informatics but did not use the GEO database regularly. Thus, finding users who

come from less technical backgrounds or are unfamiliar with Boolean operators could lessen the performance advantage that faceted search has over free-text search, since answering more complex questions using faceted search requires more training.

Future Directions

The next step after confirming that faceted searching is superior to free-text searching is to incorporate terminologies into the search tool. How can terminologies help faceted search? First, standard terms are needed when naming the facets themselves. Making sure that facet names are from a controlled terminology is crucial for sharing data with other databases. The second way that terminologies can enhance faceted search is through query expansion.[60] The NCI thesaurus, in addition to providing a standard terminology, also has two properties that can be used for this purpose: synonymy and class hierarchies. It has been established that using synonyms of search terms improves search results. [62] Recognizing synonyms, however, is a first step towards improving precision and recall, because augmenting search terms with synonyms indicates that the system would then be beginning to search according to the meaning of the query. Without synonyms, searching individual tags and values—even if they are standardized—still amounts to string matching on the facets themselves. For example, without synonyms, searching for male patients amounts to: "gender = male". With synonym support, a user could use that same tag: value pair and receive matches for samples that have been annotated with "sex = boy" as well as "gender = man." Synonym support is important because the keywords used by indexers to describe facets do not match the keywords that users expect.

In general, "users do not understand an information space in terms of the same facets as the indexers who designed it."[45]

The second way that some terminologies can enhance search is through their class hierarchies, when available. This organization allows searching for terms with varying levels of detail as well as for searching according to terms' relationship with each other. These properties give terminologies the potential to represent phrases and terms in combinations that may not have been envisioned by the original terminology designers.

For example, suppose that a researcher is interested in finding samples from humans with a hormone receptor mutation, but that the available annotations are only: "gender = female" and "progesterone receptor= mutation." In order to match samples without adding more facets such as "species = human" and hormone receptor status = mutation," knowledge that "female" is a subclass of "human" enables the appropriate inference that a female is also a human. So, the search tool would be sure to include samples matching the tag: value pair "gender=female" when returning matches even though the original query never mentioned the word "female."
These features have been implemented successfully with the Tissue Microarray Database described earlier and with Amigo, an ontology-backed browsing interface for the Gene Ontology.[11]

Future versions of the free-text search component of the search tool could implement and tune pre-packaged search solutions to work with biomedical data. Two prime examples in this area are the Apache Software Foundation's Lucene, an open source, text search engine library, [57] [63] and Google Custom Search.

Conclusion

This research addresses the increasingly important problem of information retrieval within gene expression databases as high throughput methods continue to generate larger and larger volumes of data. The clinical descriptive information that already exists within gene expression repositories such as GEO is an untapped resource for translational research because it is stored in neither a structured nor standardized format. Transforming the text into a computer-interpretable format by using standard terminologies would lay the groundwork for future insights into the relationship between genomic data and clinical data by facilitating data reuse.

Today clinical researchers who are interested in correlating genomic data with phenotypic data would rely on a manual process. In GEO, one has to look in a GEO Sample's characteristics field, a GEO Series' title, description, or summary field, in the abstract of the published study itself, or in a GEO Dataset's description on a sample-by-sample basis.

The first part of this thesis confirmed that clinical descriptive information can be effectively represented using current terminologies. The second part involved implementing two search strategies, free-text and faceted search, and comparing these two search strategies' performance in searching for samples using descriptive information.

Today the main barrier towards making this goal a reality lies at the point of data entry—that is, researchers who upload data to GEO need incentives to include more comprehensive clinical characteristics of their samples. This current research demonstrated that once descriptive textual information is deposited in GEO, structured and standardized representation is possible using existing medical terminologies.

Once sample data is annotated into tag: value structure, using faceted search to locate samples of interest is a feasible search strategy since it demonstrates high precision and recall when compared to normal free-text searching. Identifying samples in this manner would ultimately enhance the ability to correlate genomic data with clinical data.

References

1. Human Genome Project Budget. 2009. (Accessed 10-23, 2009, at http://www.ornl.gov/sci/techresources/Human_Genome/project/budget.shtml.)

2. Human Genome Project Information. 2009. (Accessed 10-23, 2009, at http://www.ornl.gov/sci/techresources/Human_Genome/faq/faqs1.shtml.)

3. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365-71.

4. HSLS: Microarray, SAGE, and other gene expression databases. 2010. (Accessed at http://www.hsls.pitt.edu/guides/genetics/obrc/gene_expression/gene_expression_databases.)

5. Parkinson H, Sarkans U, Shojatalab M, et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005;33:D553-5.

6. Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y. CIBEX: center for information biology gene expression database. C R Biol 2003;326:1079-82.

7. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods in enzymology 2006;411:352-69.

8. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics 2008;24:2798-800.

9. GEO FAQ. 2009. (Accessed 10/23, 2009, at http://www.ncbi.nlm.nih.gov/geo/info/faq.html.)

10. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research 2009;37:D885-90.

11. Shah NH, Rubin DL, Espinosa I, Montgomery K, Musen MA. Annotation and query of tissue microarray data using the NCI Thesaurus. BMC Bioinformatics 2007;8:296.

12. Rosenfeld LaM, Peter. Information Architecture for the World Wide Web. In: LeJeune L, ed. second ed. Sebastopol, CA: O'Reilly & Associates; 2002.

13. Noy N, McGuinness D. Ontology Development 101. To be downloaded from http://www cs man ac uk/~ carole/old/GGF% 20Tutorial% 20Stuff/ontology101 pdf Stanford University, Stanford, CA, USA(March 2001) Last accessed 2005;30.

14. Wielinga B, Schreiber Aea. From Thesaurus To Ontology. Proceedings of the 1st International Conference On Knowledge Capture 2002:194-201.

15. Controlled Vocabulary / Terminology Concepts. 2002. (Accessed at http://www.dlese.org/Metadata/vocabularies/term_expln.php.)

16. Cimino JJ. Review paper: coding systems in health care. Methods Inf Med 1996;35:273-84.

17. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394-403.

18. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform 2006:124-35.

19. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med 2005;44:498-507.

20. Sittig DF. Grand challenges in medical informatics? Journal of the American Medical Informatics Association : JAMIA 1994;1:412-3.

21. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics 2007;40:30-43.

22. SNOMED CT User Guide. In; 2008:1-70.

23. Terminology Resources--National Cancer Institute. 2009. (Accessed 10-29-09, 2009, at http://www.cancer.gov/cancertopics/terminologyresources.)

24. Mougin F, Bodenreider O. Auditing the NCI Thesaurus with semantic web technologies. AMIA Annual Symposium Proceedings 2008;2008:500.

25. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Stud Health Technol Inform 2004;107:33-7.

26. Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and Utilization of the NCI Thesaurus. Comp Funct Genomics 2004;5:648-54.

27. Hoskins H, Hildebrand P, Lum F. The American Academy of Ophthalmology Adopts SNOMED CT as Its Official Clinical Terminology. Ophthalmology 2008;115:225-6.

28. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006;81:741-8.

29. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-6.

30. Solutions CST. SNOMED CT User Guide. 2008:1-70.

31. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. Journal of the American Medical Informatics Association : JAMIA 2001;8:317-23.

32. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000;88:265-6.

33. Schulman J-L. History of MeSH. In; 2010:1.

34. Redefining a Thesaurus: Term-centric No More. NLM, 2001. (Accessed 2010, May 16, at http://www.nlm.nih.gov/mesh/redefine.html.)

35. MeSH Fact Sheet. National Library of Medicine, 2009. (Accessed 10-29, 2009, at http://www.nlm.nih.gov/pubs/factsheets/mesh.html.)

36. WHO | International Classification Of Diseases (ICD). World Health Organization, 2009. (Accessed 10-29, 2009, at http://www.who.int/classifications/icd/en/.)

37. Organization WH. ICD-10 Second Edition. In: WHO, ed. 2 ed. Geneva; 2004.

38. Microsoft Word - History of the development of the ICD.doc; 2004 Jul 6.

39. Health CDoP, Environment. New International Classification of Diseases (ICD-10): The History and Impact. 2001;41:1-4.

40. Pitzer E, Lacson R, Hinske C, Kim J. Towards large-scale sample annotation in gene expression repositories. BMC Bioinformatics 2009.

41. Lacson R, Pitzer E, Hinske C, Galante P, Ohno-Machado L. Evaluation of a large-scale biomedical data annotation initiative. BMC Bioinformatics 2009;10 Suppl 9:S10.

42. Tzitzikas Y, Analyti A, Spyratos N, Constantopoulos P. An algebraic approach for specifying compound terms in faceted taxonomies. Information Modelling and Knowledge Bases

XV, 13th European-Japanese Conference on Information Modelling and Knowledge Bases, EJC'03 2004:67-87.

43. Vikström A, Skånér Y, Strender L, Nilsson G. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. BMC Medical Informatics and Decision Making 2007;7:9.

44. Alexa Top 500 Global Sites. 2010. (Accessed at http://www.alexa.com/topsites.)

45. Tunkelag D. Dynamic Category Sets: An Approach For Faceted Search. SIGIR '06 Workshop on Faceted Search Conference 2006:1-5.

46. Koren J, Zhang Y. Personalized Interactive Faceted Search. Proceeding of the 17th international conference on World Wide Web 2008:477-86.

47. Anderle P, Duval M, Draghici S, et al. Gene expression databases and data mining. BioTechniques 2003;Suppl:36-44.

48. Lemoine S. A Global Evaluation Of Microarray Dedicated Databases. Biology Department Genomic Service 2004:1-4.

49. Golubchik S. MySQL Fulltext Search. In: ComCon. Frankfurt; 2004.

50. Zhou W, Smalheiser NR, Yu C. A tutorial on information retrieval: basic terms and concepts. Journal of biomedical discovery and collaboration 2006;1:2.

51. MySQL. MySQL 5.0 Reference Manual: 11.8.2 Boolean Full-Text Searches. In; 2010.

52. MySQL. MySQL 5.0 Reference Manual: 11.8.3 Full-Text Searches With Query Expansion. In; 2010.

53. Salton G, Buckley C. Improving retrieval performance by relevance feedback‡ Journal of the American Society For Information Science 1988:24.

54. Kwasnik B. The Role of Classification In Knowledge Representation and Discovery. Library Trends 1999;48:22-47.

55. Hearst M. Design Recommendations For Hierarchical Faceted Search Interfaces. ACM SIGIR Workshop on Faceted Search 2006:1-5.

56. How to Make a Faceted Classification and Put It On the Web. Miskatonic University Press, 2003. (Accessed February 19, 2010, at http://www.miskatonic.org/library/facet-web-howto.html.)

57. Ben-Yitzhak O, Golbandi Nea. Beyond Basic Faceted Search. Web Search and Data Mining '08: Proceedings of the international conference on Web search and web data mining 2008:33-44.

58. Goutte Cyril GE. A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation. Proceedings of the 27th European Conference on Information Retrieval 2005:345-59.

59. Prieto-Diaz R. Implementing Faceted Classification For Software Reuse. IEEE 1990:300-4.

60. Garcia E, MA S. User Interface Tactics In Ontology-based Information Seeking. PsychNology Journal 2003;1:242-55.

61. Flamenco Search Interface Project: Search Interfaces That Flow. UC Berkeley. (Accessed May 14, 2010, at http://flamenco.berkeley.edu/.)

62. Bechara A, Machado ML. Applying Biomedical Ontologies On Semantic Query Expansion. Nature Precedings 2009:1-.

63. Apache Lucene - Overview. 2010. (Accessed at http://lucene.apache.org/java/docs/#Apache.)

Figures & Tables List

Figures

Tables