This is a preprint of an article published in *Journal of Library Metadata*.
Wu, S., Stvilia, B., & Lee, D. J. (2012). Authority control for scientific data: The case of molecular biology. *Journal of Library Metadata, 12(2-3)*, 61-82. http://www.tandfonline.com/doi/abs/10.1080/19386389.2012.699822

Authority Control for Scientific Data: The Case of Molecular Biology

Shuheng Wu, Besiki Stvilia, & Dong Joon Lee

School of Library and Information Studies, Florida State University,

Tallahassee, FL, 32306-2100, USA

sw09f@my.fsu.edu, bstvilia@fsu.edu, dl10e@my.fsu.edu

Authority Control for Scientific Data: The Case of Molecular Biology

Abstract

This paper analyzes the authority control practices in molecular biology using literature review and scenario analysis and makes a comparison with bibliographic authority control. The analysis indicates the absence of conceptual authority control model in molecular bioinformatics. In addition to traditional knowledge organization tools, authority control in molecular biology requires the use of reference sequences and version numbers to identify entities and keep track of entity changes. The identified authority control issues are conceptualized as quality problems caused by four sources. This study can inform librarians and educators of the needs for and approaches to authority control in molecular biology.

Keywords: authority control, molecular biology, metadata, scientific data management

Introduction

Research processes have become increasingly data driven, and there are growing needs as well as opportunities to share, reuse, and aggregate data from different contexts. The Institute for Museum and Library Services (IMLS, 2011), the National Endowment for the Humanities (NEH, 2011), the National Science Foundation (NSF, 2010), and the National Institutes of Health (NIH, 2010) now require applicants to submit data management plans, including plans for disseminating and providing access to research data and related metadata. To maintain data in a usable/reusable state for ongoing research, education, reporting, verification, and evaluation, it is essential to assure the quality of related metadata, including entity metadata (i.e., entity profiles). Effective reuse and aggregation of data may require knowledge of community, disciplinary and

cultural differences in metadata quality requirements, rules, norms, and references sources (Atkins et al., 2003; National Science Board, 2005; NSF, 2007; Stvilia, Gasser, Twidale, & Smith, 2007).

Entities are distinguishable objects that can be concrete or abstract (Elmasri & Navathe, 2000). Examples of entities are books, authors, proteins or genes. A set of important attributes that characterize a particular entity constitutes the entity's metadata profile, which can be included in reference databases (e.g., authority databases) and used for entity determination and disambiguation. Effective management of entity metadata, the ability of entity resolution and disambiguation are essential for scientific research processes, as well as for scientist productivity and impact evaluation. In biology, taxonomists may need to determine whether a particular specimen belongs to an established taxon, or if it represents a new taxon. Genomics researchers may need to distinguish the sample's identity in order to identify genotype-phenotype relationships at the individual or population level. Librarians, and in particular catalogers, need to resolve different entities in bibliographic databases in order to link and collocate related works. Likewise, administrators and bibliometrics/scientometrics researchers may need to resolve author names to evaluate the productivity and impact of individual scientists, groups, or institutions.

There have been distinct domain-specific approaches to entity metadata management. Libraries have managed entity metadata through authority databases and controlled vocabularies. Similarly, life sciences rely on elaborate manually constructed and maintained taxonomies, keys, and ontologies to make entity determination. In addition, there have been efforts to automate entity resolution and disambiguation in large-scale text collections (see Smalheiser & Torvik, 2009, for a recent review). Different communities have proposed different conceptual

frameworks, metadata schemas, and data structures for entity identifiers and metadata profiles (e.g., FRAD, URI, ISBN, DOI, LSID, PURL, MARC21 for Authority). As Semantic Web technologies become more widely accepted, libraries, institutions, governments, and communities are starting to disseminate their data and the reference sources used in entity resolution as linked data for open access and use (e.g., DBpedia,[i] LinkingOpenData[ii]).

The efforts of providing open access to data and integrating data from different contexts also highlight the need for better reasoning about the quality and interoperability of identifier and reference/knowledge organization systems used for data referencing and entity resolution. Needs and requirements for entity metadata control, and their operationalizations—what entities are controlled, what sets of attributes are used for each entity—may change in time and space. Different domains may control for different entities using different sets of attributes. Furthermore, these sets may evolve and change over time as the amount of data grows and more attributes are needed for entity resolution. Finally, entities and their instances are dynamic in that they move in space and time. Authors may change names, affiliations, disciplines, and residences. Data, too, are often "works in progress." Old knowledge becomes obsolete as new knowledge becomes available, and can be reused and updated by different actors (e.g., genome annotation data).

Problem Statement

Most biology journals now require submission of newly sequenced DNA to one of the public nucleotide repositories (e.g., GenBank[iii]) before publication. This policy has led to great success in the progress of biology, and exponential increase in the size and usage of nucleotide sequences. Since the publication of the human genome in 2001, the world has entered into the "post-genome age" (Higgs & Attwood, 2005, p. 4). With advances in sequencing technologies,

high-throughput experimental techniques have been developed to study large numbers of genes or proteins simultaneously. Microarrays, proteomics, and structural genomics are examples of high-throughput techniques. The exponential increase in the size of nucleotide sequences, the availability of whole genome sequencing, and the large amount of data generated from high-throughput experiments—data encoded in different formats using different vocabularies and stored in different databases—pose challenges for organizing, storing, retrieving, analyzing, and managing data in biological repositories. Furthermore, the increasingly data driven science, along with funding agencies and publications requiring scientists to provide access to research data, puts pressure on scientists' home institutions and their libraries to provide appropriate infrastructure and expand the scope of their traditional services to meet the changing data management and dissemination needs of their constituents.

Metadata management for entity and instance determination, disambiguation, and referencing, referred to in libraries as authority control, is an essential part of data management in any domain. Libraries need to have better understanding of the data practices of different disciplines to be effective in assisting their faculty with the management and dissemination of research data. Although there is significant prior research of the disciplinary practices of authority control in Library and Information Science (LIS) and other disciplines, there has been relatively little examination of the similarities and differences of authority control, and the reusability of authority models, tools, and data across different domains.

This paper analyzes the authority control practices in the area of molecular biology, and compares those to the authority control practices in libraries. This can inform librarians and library educators about the requirements for authority data in molecular biology, and help align

library authority models, vocabularies, and data with the needs of scientific data curation and research tasks.

Overview of Research Questions and Methodology

This paper explores the needs and requirements for data referencing, entity resolution, and authority control in molecular biology. It helps illuminate the following research questions: What are the needs and requirements for data referencing and authority control in molecular biology? How is authority control currently implemented in molecular biology? What are some of the frameworks, models, controlled vocabularies and schemas used? What are some of the issues and problems with the current practices of authority control in molecular biology? What solutions have been sought? How do the models and practices of authority control from molecular biology compare to bibliographic authority control and how one field can inform the other? A detailed examination of the literature and entity resolution and authority control frameworks, models, and systems is provided.

In addition to literature analysis, the study's methodology includes the use of specific data use scenarios to illustrate the needs for and issues surrounding data referencing, entity determination, and disambiguation at different levels and in different activities within molecular bioinformatics. Scenarios and scenario based task analysis (Go & Carroll, 2004a, 2004b) are particularly helpful when there is a need to identify and develop an inquiry into non-routine or future possible uses of technology, and thus nicely complement the analysis of established frameworks, models, and standards from the literature that are tailored to the routine tasks of a particular domain. The scenario development in this study was informed by an examination of the datasets generated from three NSF funded research projects at the Florida State University,

and were collected from the American Chemistry Society Publications Database[iv] and the Web of Knowledge.[v]

Authority Control in Molecular Biology

Molecular biology is a branch of biology that seeks to explain the structure, function, and interaction of biological molecules, primarily nucleic acids (i.e., DNA, RNA) and proteins (MacMullen & Denn, 2005; Turner, McLennan, Bates, & White, 2005). Molecular bioinformatics is concerned with developing and applying computer-information technologies for studying and organizing data and knowledge about these entities and relationships. The concept of authority control in molecular biology is associated with three tasks: *named entity recognition*, *disambiguation*, and *unification*. The named entities in molecular biology include RNA, DNA (e.g., genes, gene clusters, genomes), proteins, species, organisms, and others (Blaschke, Hirschman, & Valencia, 2002; de Bruijn & Martin, 2002; Krallinger, Erhardt, Valencia, 2005; Krallinger, Valencia, & Hirschman, 2008). Biological scientists usually report newly found entities in the literature, and deposit their information to related databases as required by scholarly journals. Each record of these databases usually provides references to the publications that discuss the data in the record. Biological scientists search these databases for information about biological entities using their names and alternates, which are usually incomplete with respect to those found in the literature (Cohen & Hersh, 2005).

The purpose of named entity *recognition* in biology is to identify all the instances of a name for a specific type of biological object within a collection of text (Cohen & Hersh, 2005). Due to the dynamic properties of biological objects, biological terminology constantly changes. Hence many biological entities have multiple names and abbreviations used and referenced interchangeably in databases and the literature. In addition, names can sometimes refer to

different concepts dependent on context. The purpose of name entity *disambiguation* is to link a recognized named entity to a correct concept in a taxonomy, controlled vocabulary, thesaurus, or ontology (Ananiadou, Friedman, & Tsujii, 2004). Overall, the purpose of named entity recognition and disambiguation is to identify key concepts of interest in literature or data, and represent these concepts in a consistent, formalized, and related form. However, there does not exist a complete dictionary for the standardized use of most types of named entities (Cohen & Hersh, 2005). The purpose of named entity *unification* is to produce a unification of biological entities to conceptualize the shared biological objects among communities, standardize the nomenclature and use of these entities, and enable interoperability of biological databases (Gene Ontology Consortium, 2000).

Entities and Relationships: Central Dogma of Molecular Biology

The main theoretical model that conceptualizes relationships among entities in molecular biology is a principle known as central dogma theory. It states that genetic information passes from DNA to RNA to proteins (Crick, 1970). Transcription refers to the process from DNA to RNA, where synthesis of RNA involves rewriting or transcribing the DNA sequences in the same language of nucleotides. Translation refers to the process from RNA to proteins, where synthesis of proteins involves translating the language of nucleotides to the language of amino acids. In addition to DNA (e.g., genes, gene clusters, genomes), RNA, amino acids, proteins, and traditional entities of authority control (such as persons and organizations), some of the other entities that are important to knowledge organization in molecular biology are cells, tissues, species, populations, organisms, drugs, and diseases (Blaschke et al., 2002; Krallinger et al., 2005; Krallinger et al., 2008). Based on these entities, the properties (e.g., protein functions, cellular locations) of and relationships (e.g., protein-protein interactions, protein-drug

interactions) between these entities are also recognized and extracted from the literature, and used to construct thesauri, controlled vocabularies, and ontologies (Blaschke et al., 2002).

Authority Control Tools in Molecular Biology

The authority control infrastructure of molecular biology consists of several kinds of metadata and knowledge organization systems, such as nomenclatures, ontologies, reference sequence databases, and metadata and identifier schemas.

Nomenclatures. Nomenclature is concerned with the scientific naming of objects and establishing principles on which scientific names are based. Through the standardized and unique naming of biological objects, nomenclature is essential for the literature search, entity representation and retrieval, and scientific communication among different communities. In the domain of molecular biology, a number of conventions have been developed to standardize gene and protein nomenclatures. The HUGO Human Gene Nomenclature Committee is the only authority to assign and approve unique and meaningful human gene names and symbols (Wain et al., 2002). The Committee published the Guidelines for Human Gene Nomenclature as early as 1979, with later updates to evolve with new technology (e.g., high-throughput techniques). The Guidelines recommend that gene names be brief, specific, use American spelling, and convey the function of the gene. The Committee stores all approved human gene nomenclature in a publicly accessible database, genenames.org[vi] (Seal, Gordon, Lush, Wright, & Bruford, 2011). Each gene in the database has a symbol report that contains approved nomenclature; previous symbols, names, and aliases; and a unique identifier that remains stable even if the nomenclature changes. Manually curated and reviewed by the Committee editors, the genenames.org Web site serves as an authority file for approved human gene nomenclature. Responding to the need to report and describe changes (mutations) in DNA and protein sequences, the Human Genome Variation

Society developed the nomenclature for sequence variants, suggesting, for example, that the description should be at the most basic level (i.e., DNA) and be related to a reference sequence (den Dunnen & Antonarakis, 2000).

Many species-specific communities have also established gene nomenclature committees to assign and approve gene names and symbols, such as the International Committee on Standardized Genetic Nomenclature for Mice. In addition, some model organism databases provide guidelines for establishing gene names and symbols, such as FlyBase[vii] and WormBase.[viii] However, there are no specialized organizations establishing protein-naming rules and standardizing protein names across species. Some protein repositories, such as UniProt,[ix] have developed local naming guidelines to standardize the nomenclature for a given protein across related organisms (UniProt Consortium, 2011). This is accomplished through ongoing efforts to assign a recommended name to existing proteins with a list of alternative names as references in the repository based on the UniProt guidelines.

Reference sequences. As any scientist can submit data to GenBank or other sequence databases, there are cases that several entries in these databases are representing the same sequence or presenting alternate views of protein or entity names. To resolve the data redundancy problem in GenBank, the National Center for Biotechnology Information (NCBI)[x] established a publicly accessible nucleotide and protein sequence database—Reference Sequence (RefSeq)[xi]—by collocating, synthesizing, validating, and summarizing the sequence data available in GenBank (Pruitt, Tatusova, & Maglott, 2005). The goal of RefSeq is to provide a non-redundant collection of genomic and protein sequence data for any given species. In addition to the annotation propagated and validated from GenBank records, NCBI staff may provide supplementary annotation to each record in RefSeq with support from collaborative

nomenclature committees, model organism databases, user feedback, and other scientific communities (Pruitt et al., 2005; Pruitt, Tatusova, Klimke, & Maglott, 2009). In particular, RefSeq assigns current entity names and symbols approved by collaborative nomenclature committees to represent the current view of entities. Scientists use RefSeq as an international authority for genome annotation and a stable genomic sequence standard for reporting sequence variants that might be of clinical significance (Pruitt et al., 2009). Likewise, NCBI built RefSeqGene[xii] to store reference sequences for well-characterized individual genes, which is used as the "gold standard" for determining and describing gene variants (Gulley et al., 2007, p. 862).

Bio-ontologies. Recently, there has been a trend towards the development and adoption of bio-ontologies in the biomedical and biological communities, attempting to: (a) represent current biological knowledge, (b) annotate and organize biological data, (c) improve interoperability across biological databases, (d) turn new biological data into knowledge, and (e) assist users in analyzing data across different domains (Bard & Rhee, 2004; Gene Ontology Consortium, 2000). Bard and Rhee (2004) define bio-ontologies as "formal representations of areas of knowledge in which the essential terms are combined with structuring rules that describe the relationship between the terms" (p. 213). Unlike thesauri and taxonomies, ontologies are more flexible. The relationships among terms in a thesaurus are loosely specified, and usually include broader terms (BT), narrower terms (NT), related terms (RT), and synonymous terms (ST) (Allen, 2011; Hodge, 2000). However, the relationship between two concepts in an ontology can be of any type, not limited to those BT, NT, RT, and ST relationships in thesauri or the "is-a" relationship in taxonomies (Lambe, 2007). For example, the Gene Ontology[xiii] uses three types of relationship between terms ("is-a," "part of," and "regulates") to encode

knowledge about genes and gene products related to biological processes, molecular functions, and cellular components (Gene Ontology, 2011).

Among many bio-ontologies that have been developed, the Gene Ontology and the Unified Medical Language System (UMLS)[xiv] are the most influential in molecular biology and biomedicine, and have been widely used for text mining and information extraction (Blaschke et al., 2002). The UMLS is a large-scale repository of biomedical vocabularies developed by the U.S. National Library of Medicine, aiming to enhance access to biomedical literature and improve interoperability of biomedical databases by solving the problem of a variety of names being used for the same concept (Bodenreider, 2004). The UMLS consists of three knowledge sources: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. Metathesaurus is a repository integrating over 2.6 million concepts and their relationships from 135 source vocabularies, including the Medical Subject Headings (MeSH),[xv] NCBI Taxonomy,[xvi] Gene Ontology, and HUGO Gene Nomenclature (Bodenreider, 2004; Unified Medical Language System, 2011). Besides relationships inherited from source vocabularies, Metathesaurus editors assign one or more semantic categories to each concept in the Metathesaurus from the Semantic Network, which is a catalog of 133 semantic categories linked by 54 relationships. Most of the relationships in the Semantic Network are hierarchical (e.g., “is a,” “part of”), but some of them are associative (e.g., “spatially related to,” “temporally related to,” “functionally related to”). Independent of the structure of source vocabularies, the Semantic Network serves as an authority of semantic categories and relationships for concepts in the Metathesaurus, and enables cross-references of biomedical concepts from different source vocabularies. Therefore, the UMLS may be viewed as a large-scale biomedical ontology (Bard & Rhee, 2004).

Entity identification system. In order to promote a consistent identification mechanism for assigning and recognizing identifiers in the scientific community, the Interoperable Informatics Infrastructure Consortium published the life science identifier (LSID) specification (Martin, Hohman, & Liefeld, 2005). The LSID is a special form of universal resource name and has six components delimited by colon, including an authority ID (e.g., "ncbi.nlm.nih.gov") that identifies an authority which assigned the LSID, an authority namespace ID (e.g., "GenBank.accession") that identifies an authority-specific namespace within which the LSID lives, a unique object ID, and an optional revision ID (Clark, Martin, & Liefeld, 2004). The LSID specification requires any LSID to be location-independent, globally unique, and permanent; it can specify only one object at a time and can never be reassigned. A change of even a single bit to the object identified by an LSID should result in a new LSID. This is known as the byte-identity contract (Martin et al., 2005). It is recommended that the new LSID be based on the original one, but with an increment to the revision ID. The use of the revision ID allows users to retrieve different versions of the data object, and indicates the number of times the object has been changed (Dalgleish et al., 2010).

Compared with other identifiers, the LSID can provide semantics or context to make it recognizable to machines, and easier to parse (Clark et al., 2004). For example, the LSID "URN:LSID:ncbi.nlm.nih.gov:GenBank.accession:NC_003428.1" is an identifier for a GenBank record stored at the NCBI database. The existing identifiers can be wrapped into or included within LSIDs (as the GenBank example above, where "NC_003428.1" is a current GenBank identifier), and thus data providers do not need to create new identifiers and discard their current ones. Furthermore, as location-independent and globally unique identifiers, LSIDs can be

associated with concepts in ontologies, taxonomies, and controlled vocabularies, and serve as the foundation for the biological Semantic Web.

Metadata schemas. Previous studies (e.g., MacMullen & Denn, 2005; San Gil, Hutchison, Frame, & Palanisamy, 2010) have identified the need for metadata standardization to support interoperability among disparate biological databases. There is also a need for added metadata in existing biological databases to fulfill different users' needs. For example, contextual metadata describing the environment where a gene or an organism was collected (in terms of space, time, and habitat characteristics) is a prerequisite to understanding the function of unknown genes and organisms (Yilmaz, Gilbert et al., 2011; Yilmaz, Kottmann et al., 2011). However, contextual metadata, usually found in the literature, is missing in sequence databases (Field et al., 2008; Yilmaz, Kottmann et al., 2011). With the exponential increase in the quantity of genome sequences, it is imperative to provide adequate contextual metadata in a standardized format to extend the existing sequence databases and support genomic analysis. In response to the need to enhance the classic GenBank metadata, the Genomic Standards Consortium (GSC) published the Minimum Information about a Genome Sequence (MIGS) specification to define a set of core (required) metadata for genomes, including information about the environment where the sample was collected (contextual metadata), taxonomic groups of the sequence, and the experimental process (Field et al., 2008). The GSC implements MIGS in extensible markup language (XML)[xvii] as Genomic Contextual Data Markup Language, specifying the use of particular identifier systems (e.g., PubMed identifier, digital object identifier), controlled vocabularies, and ontologies (e.g., the Environment Ontology[xviii]) for most genomic metadata in the standard.

More recently, the GSC published the Minimum Information about a Metagenome Sequence (MIMS), which is an extension of MIGS to include habitat contextual metadata to describe metagenome sequences (GSC, 2011). Based on the results of community-led surveys about marker gene descriptors and analysis of contextual data in published rRNA gene studies, the GSC proposed the Minimum Information about a Marker Gene Sequence (MIMARKS) and the environmental packages to standardize descriptions for a more comprehensive range of environmental parameters (Yilmaz, Kottmann et al., 2011). The primary reason for introducing the environmental packages is that the existing keyword search in sequence databases cannot retrieve sequences originated from certain environments or particular locations (e.g., freshwater lakes). The environmental packages can be combined with any GSC standard to enhance sequence description. In order to have a single entry for all the GSC standards, the GSC created the Minimum Information about Any (x) Sequence (MIxS) specifications as an overarching framework to include MIGS, MIMS, MIMARKS, and the environmental packages (Yilmaz, Kottmann et al., 2011).

Authority Control Issues in Molecular Biology

The issues and problems of authority control can be conceptualized as data and metadata quality problems rather than unexpected phenomena. In general, quality is defined as "fitness for use," and it is contextual (Juran, 1992; Wang & Strong, 1996). The issues of authority control in molecular biology can be analyzed under four facets or categories of quality problem sources: (a) problems of inaccurate, inconsistent, or incomplete mapping; dynamic quality problems such as (b) problems caused by context changes; (c) problems caused by changes in the entity; and (d) problems caused by changes in the entity's metadata (Stvilia et al, 2007; Wand & Wang, 1996).

Inconsistent Mapping

Biological researchers usually need to consult or collect data from multiple sources to conduct experiments, interpret results, and make predictions. For example, in order to study the structure of an unknown protein, researchers need to take into account several types of data, ranging from gene sequences to protein structures. However, most publicly accessible databases curate only one type of data, and no resources are available to provide one-stop shopping for all information (Khatri et al., 2005). In order to gain a complete picture of the problem under study, researchers need to navigate from one resource to another. Therefore, it is necessary to create cross-references among related databases: a gene database needs to link to a genome database to signify the location of a gene on its genome; an mRNA database needs to link to a gene database and a protein database to indicate a gene from which an mRNA is transcribed, and a protein to which an mRNA translates; and a protein database needs to link to a gene database and a protein structure database. Even though many public databases now provide cross-references to related databases, each of them has its own metadata schema and identifier system. Most of the time, users have to manually convert an identifier from one database to another, or use the online converters (e.g., X-REF Converter[xix]).

Scenario 1

Biomedical researchers want to analyze proteins in Androgen-repressed human prostate cancer (ARCaP) cells. They have the names and International Protein Index (IPI) accession numbers (identifiers) of these proteins, and want to identify their cellular locations, functions, and pathways by searching the NCBI databases and the Pathway Interaction Database.[xx] However, these databases do not identify proteins by the IPI accession number. Considering various names used for these proteins in the databases, researchers have to use an identifier

converter to convert the IPI accession numbers to the GenBank protein accession numbers and UniProt protein accession numbers that are recognized in these databases.

Solutions to Inconsistent Mapping

The biological text-mining community has created dictionaries to aggregate and resolve various gene and protein names to improve entity recognition and retrieval in the literature (Goll et al., 2010). For example, Liu, Hu, Zhang, and Wu (2006) constructed a BioThesaurus by collecting gene and protein names from 13 databases and mapping them to protein entries in UniProt. The synonymous protein names in the Thesaurus can be used for query expansion when doing database or literature searches. For each protein, the Thesaurus also includes information about protein classifications and source organisms that can help disambiguate homonymous protein names used for different organisms. Fundel and Zimmer (2006) demonstrated that combining different gene and protein databases could result in a broader coverage of the dictionary, which considerably increased the number of terms and decreased the ambiguities of gene and protein names in different databases and with common English words.

The other approach to dealing with mapping problems is to construct a single new bio-ontology by combining terms and relationships from multiple orthogonal ontologies. However, the difficulties of this approach are in determining which overlapping terms should be eliminated and analyzing new concepts generated by the combination of related terms. Although the UMLS covers concepts from a variety of domains (e.g., anatomy, clinical genetics, nursing, psychiatry), it still preserves views and architecture from diverse source vocabularies (National Library of Medicine, 2009). Instead of constructing a general new bio-ontology, the UMLS combined terms from source vocabularies through identifying and mapping synonymy relationships among the terms (Smith et al., 2007).

Incomplete Mapping

Biologists may have difficulty finding and reusing data underlying published research due to incomplete metadata (Greenberg, 2009). Missing the metadata necessary for discovering, interpreting, using, and reusing existing data—such as spatial and temporal coverage, revision ID, specimen identity information, or contextual metadata—may hamper the research process. Publication without mentioning the version number (revision ID) of sequences can result in ambiguous interpretations and inconsistent descriptions of the data (Dalgleish et al., 2010). Geneticists and clinicians may find current reference sequences are missing annotations of clinically relevant transcripts (i.e., RNA sequences produced from transcription) that are essential for reporting sequence variants. Bioinformaticians doing phylogenetic analysis (studying evolutionary relatedness among groups of organisms) or phylogeographic analysis (studying geographic distributions of organisms) may spend longer than expected collecting and identifying samples from existing sequence databases due to the lack of geographic or contextual metadata describing organisms.

Scenario 2

A graduate student is doing a bioinformatics course project to determine the taxonomy of the giant panda from the phylogenetic perspective. The student wants to collect the ND2 gene sequence of the giant panda and six different kinds of bears from the GenBank to do phylogenetic analysis. In particular, the student is trying to determine whether the giant panda belongs to Ursidae as true bears or is an independent species. To answer this question, the student wants to collect the ND2 gene sequence from multiple individuals within a species. Even though they have their own identifiers, several records retrieved from the GenBank are repetitive or redundant, representing the same sequence. As no identity information about the specimen

(individual) is provided in the GenBank, the student has to examine the publication information in each record or read the publications that discuss the data to determine if the sequence was collected from the same individual or not.

Solutions to Incomplete Mapping

In order to extend current sequence databases, the International Nucleotide Sequence Database Collaboration (INSDC)[xxi] has recently adopted the MIxS specifications to include contextual metadata in the sequence records (Yilmaz, Kottmann et al., 2011). Additionally, the GSC recommends authors of genome and metagenome publications submit a MIGS report (about contextual metadata) after depositing sequence data in the INSDC databases (Field et al., 2008). To enrich data service and enhance metadata in biological databases, Patterson et al. (2010) proposed a taxon name-based infrastructure: a linked data cloud consisting of taxon names interconnected to an array of data including nomenclature and taxonomies, publication data, georeferences, and social network data. For example, the latitude-longitude metadata in a sequence record of a rare spider allows access to maps that display where the spider was found; nomenclature and taxonomies enable reconciliation and disambiguation of the spider, and access to worldwide distributional data of the spider in the Global Biodiversity Information Facility;[xxii] publication data (e.g., keywords) enables retrieving more publications about the spider in digital libraries; and author's social network data can provide access to all the publications by the author and the author's collaborators and colleagues.

Greenberg (2009) proposed and demonstrated the applicability of automatic metadata propagation, inheritance, and adoption from outside (non-biology) standardized value systems (e.g., the Library of Congress Subject Headings, the Library of Congress Name Authority File) to enhance biological repositories. For example, a data object in a repository can inherit

keywords from its original research article, which can be used as seeds to harvest more keywords for the data object from outside controlled vocabularies. The author of an article is usually the creator of or contributor to the data object represented in the article, and therefore the author metadata can be automatically propagated as the creator metadata of the data object.

Dynamic Quality Problems

As stated previously, proteins and genes are recommended to be named based on their functions and homology to known proteins (Goll et al., 2010; Wain et al., 2002). Scientists, however, are still learning more about protein functions, and thus protein names need to be changed frequently to reflect newly found or revised knowledge of functions. With the large scale of data generated by high-throughput techniques, the manual correction of existing problematic names is not feasible (Goll et al., 2010). As a consequence, several names (synonyms) are in use for the same genes and their corresponding proteins across databases and the literature. Researchers should consider all available gene or protein names when doing database or literature searches; otherwise they risk missing information. Furthermore, the number of cross-references among databases has increased significantly since many of them began to collaborate and share data (Fundel & Zimmer, 2006). Cross-referencing, however, may lead to data redundancy and inconsistency since some data are stored in multiple databases and might be updated or changed asynchronously (Khatri et al., 2005).

Scenario 3

Biomedical researchers are interested in the function of a protein named FAM20C in humans. They want to know about previous research on this protein by doing database and literature searches. Since protein names are usually based on their functions, this protein might have various names if different researchers have identified different functions over time.

Researchers use the protein name "FAM20C" to find related articles from databases, and analyze the reference list of these articles to know about other names (e.g., dentin matrix protein 4) for this protein. They then use these names to find more related articles.

Solutions to Dynamic Quality Problems

The difficulties of maintaining bio-ontologies lie in gaining community acceptance and integrating new knowledge. One solution to these problems is to create forums (Bard & Rhee, 2004) or Wikis for bio-ontologies, allowing those with specialized domain knowledge or interested in ontology development to provide feedback and contribute new concepts. Community involvement in the maintenance of ontologies can not only help gain public support and facilitate public ownership, but also ensure that only a single ontology is used in any particular domain. Because of the proliferation of bio-ontologies, the Open Biological and Biomedical Ontologies (OBO) consortium was founded in 2001 to establish principles to standardize the format of bio-ontologies, foster interoperability, and ensure a reference ontology for any particular domain (Smith et al., 2007). Built on the success of the Gene Ontology, the OBO principles specify that ontologies must be open access without any constraint; be expressed in a shared syntax, either the OBO syntax or OWL;[xxiii] possess a unique identifier space; be receptive to community feedback and modification; and be orthogonal without overlap in content (OBO, 2011; Smith et al., 2007).

Discussion

The literature analysis indicates that molecular bioinformatics does not define a high level conceptual model of information systems with relationships among tasks and entities. Although the Central Dogma theory defines the relationships among DNA, RNA, and proteins, there is no overall community-agreed model that defines relationships among entity metadata

and information tasks. Entity databases have their own data models, nomenclatures, and identifier schemas. Users are often forced to do complex mapping and translation of entity names and identifiers to search and aggregate data from these databases. In contrast, libraries have already developed and somewhat adopted clear conceptual models for organizing and providing access to library materials. The International Federation of Library Associations and Institutions (IFLA) developed the Functional Requirements for Bibliographic Records (FRBR), which uses the Entity-Relationship Diagram (ERD) conceptual modeling technique and language to conceptualize main bibliographic entities and relationships. First, FRBR defines a data model for library catalogs, which consists of three groups of entities—products of intellectual or artistic endeavor (Group 1); those responsible for the content, production, or custodianship of the products (Group 2); and entities that may serve as subjects of the entities (Group 3) (IFLA, 2008). Next, FRBR links these entities to four user tasks that "are defined in relation to the elementary uses that are made of the data by the user": (a) *Find* entities using entity attributes or relationships; (b) *Identify* entities, or distinguish entities with similar attributes; (c) *Select* entities corresponding to the user's needs; and (d) *Obtain* access to online electronic entities, or acquire physical entities (IFLA, 2008, p. 79).

IFLA has also developed a conceptual model for authority records, the Functional Requirements for Authority Data (FRAD), which specifies attributes of and relationships between entities (e.g., subject headings, personal names, etc.) and the authority records for those entities are based on another set of four user tasks: (a) *Find* entities using stated criteria or explore using entity attributes or relationships; (b) *Identify* the attributes of an entity to be used as an access point, or validate the attributes; (c) *Contextualize* or clarify the relationship between entities used as access points; and (d) *Justify* or document the reasons for the choices made by

the authority data creator (Patton, 2009, p.83). The three tasks associated with the concept of authority control in molecular biology—*named entity recognition*, *disambiguation*, and *unification*—have parallels to the users' tasks in FRAD. Similar to *named entity recognition* and *disambiguation* in molecular biology, bibliographic authority data creators *identify* versions of entity names, and *validate* or *establish* authorized versions of entity names. Similar to *named entity unification* in molecular biology, bibliographic authority data creators *justify* the choice of authorized versions of entity names, and *contextualize* entity names, collocating and relating access points where relationships exist. The bibliographic conceptual models and the best practices of model implementations in libraries could benefit the molecular biology communities, and help them develop their own aggregate data and authority control models.

The absence of overall conceptual data and task models in molecular bioinformatics could be attributed to the complexity of the field and the data. Molecular bioinformatics is a relatively new field developing computational methods to study the structure, functions, and relationships of multiple entities, such as RNA, genes, and proteins (Higgs & Attwood, 2005). The curators of molecular biology databases and knowledge organization tools have to collect data about entities and standardize the descriptions of these entities that are independent from each other and stored in different databases maintained by different communities. In contrast, libraries, until recently, have been organizing and providing access to mostly one entity: the item entity from Group 1 of the FRBR model (books, serial publications, maps, etc.). Catalogers are responsible for describing these materials and creating access points. Cataloging also includes authority control to ensure the consistency of access points through terminological control (Gorman, 2004; Svenonius, 1986; Tillett, 2004). These access points—typically name and subject terms and phrases—are then used to help the user find, identify, select, and obtain

relevant resources via the library catalog. Although librarians create a separate authority record for each entity serving as a controlled access point, the creation of the authority record is triggered by creation of the information resource and/or its introduction into a collection. Most importantly, the completeness of the data model for an entity (i.e., the set of attributes) is determined by how the user tasks of finding, identifying, and selecting a publication in a library catalog need to be supported. This limits the use or reuse of bibliographic entity data for the data tasks (e.g., annotation) or for research focused on an access point entity rather than on publication (see Scenario 4).

Scenario 4

Researchers want to determine whether team demographic characteristics (e.g., affiliation, discipline, gender, and seniority) are correlated with team publication productivity and impact in a community of scientists gathered around a specialized national scientific laboratory. The researchers use the Web of Knowledge[xxiv] to identify the number of publications produced by each team in the community and the number of citations received by those publications within a fixed time window. However, the Web of Knowledge provides little authority control and presents no demographic information about authors beyond providing institutional affiliation for only some of the publications. Hence, the researches have to resort to manual search, collection, disambiguation and triangulation of author identities and their demographic information from other sources on the Web, such as institutional and lab websites.

Interestingly, the successor to the Anglo American Cataloguing Rules, 2nd edition (AACR2), the Resource Description and Access (RDA) standard, allows catalogers to extend the scope of attributes, relationships, and access point control data associated with the entities of FRBR and FRAD (Joint Steering Committee for Development of RDA, 2009). Through the

extension of entity descriptions, not only may more detailed entity profiles and larger number of access points for bibliographic resources be produced, but library entity metadata may become more usable for non-library tasks and for different communities.

Entity metadata in molecular biology is different from traditional library metadata in that biological entities and their attributes are dynamic and can change or mutate with time and space (e.g., exposure to radiation). In addition to linguistic description of an entity, researchers and curators need a "data" representation—a reference sequence—to determine the entity and variants of the entity (e.g., mutated genes). Hence, authority control in molecular biology requires not only the knowledge of naming standards (i.e., nomenclatures), metadata schemas, and ontologies, but also significant subject knowledge and the knowledge of sequencing techniques and alignment tools to identify identical or similar sequences. Different from creating a single accumulative authority record for the entity (e.g., person, corporate body) in library catalogs, the biology community uses version number (revision ID) to keep track of sequence changes to the entities, creates separate records for each version of the entity linked by identifiers, and enables retrieving metadata associated with each version of the entity. Hindered by the semantic ambiguities in terminology, the biology community resorts to natural language processing techniques, such as text-mining and information extraction, to perform named entity recognition and disambiguation in the literature. Furthermore, the complexity of biological entities and their relationships leads to the development and adoption of bio-ontologies to represent, disambiguate, and unify entities and to manage biological knowledge.

Conclusion

This paper examined the authority control practices for scientific data in the area of molecular biology, and made a comparison with bibliographic authority control. The literature

analysis and data use scenarios were used to illustrate the types of authority data quality problems and issues in molecular biology as well as the solutions sought. Similarly, a data use scenario was also used to illustrate the limited use of library metadata as data in research and non-library tasks. Comparing the two practices of authority control suggests that managers and curators of molecular biology data repositories could benefit by following cataloging librarians' approach of developing systematic conceptualizations of authority metadata and task based relationships within bibliographic databases. Likewise, the analysis of data management practices in molecular biology can inform libraries how they could extend their existing authority data model and systems to enable more effective reuse of library metadata outside of the traditional library context, as well as develop new models and services for authority control for scientific data.

With academic libraries increasingly involved with scientific data curation through institutional data repositories, understanding authority control needs and practices in different disciplines becomes important. By improving our understanding of the needs for data referencing, entity determination, and disambiguation across different domains, we can better understand how to support the development of more effective data management systems as well as to enable more effective reasoning about the interoperability of these systems, and reuse of entity metadata across different domains. Future research could include examining concepts, entities, and relationships of and needs for authority control in other scientific disciplines, such as condensed matter physics.

Acknowledgements

We express our gratitude to Paul Stewart and Dr. Qingxiang Sang for providing datasets. We thank Debbie Paul, Adam Worrall, and Drs. Jianzhong Wen and Hong Huang for helpful conversations. We also thank the reviewer for helpful comments and feedback.

Notes

i. DBpedia. http://dbpedia.org/About

ii. LinkingOpenData.

http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

iii. GenBank. http://www.ncbi.nlm.nih.gov/genbank/

iv. American Chemistry Society Publications Database.

http://pubs.acs.org/doi/suppl/10.1021/pr2000144

v. Web of Knowledge. http://wokinfo.com/

vi. http://www.genenames.org/

vii. FlyBase. http://flybase.org/

viii. WormBase. http://www.wormbase.org/

ix. UniProt. Universal Protein Resource. http://www.uniprot.org/

x. National Center for Biotechnology Information (NCBI). http://www.ncbi.nlm.nih.gov/

xi. NCBI Reference Sequence (RefSeq). http://www.ncbi.nlm.nih.gov/RefSeq/

xii. RefSeqGene. http://www.ncbi.nlm.nih.gov/refseq/rsg/

xiii. Gene Ontology. http://www.geneontology.org/

xiv. Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/

xv. Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/

xvi. NCBI Taxonomy. http://www.ncbi.nlm.nih.gov/Taxonomy/

xvii. Extensible Markup Language (XML). http://www.w3.org/XML/

xviii. Environment Ontology. http://environmentontology.org/

xix. X-REF Converter. http://refdic.rcai.riken.jp/tools/xrefconv.cgi

xx. Pathway Interaction Database. http://pid.nci.nih.gov/

xxi. International Nucleotide Sequence Database Collaboration (INSDC). http://www.ncbi.nlm.nih.gov/projects/collab/

xxii. Global Biodiversity Information Facility. http://www.gbif.org/

xxiii. OWL. http://www.w3.org/TR/owl-ref/

xxiv. Web of Knowledge. http://apps.isiknowledge.com/

References

Allen, R. B. (2011). Category-based models for knowledge representation. In *Information: A fundamental construct.* Manuscript in preparation. Retrieved from http://boballen.info/ISS/

Ananiadou, S., Friedman, C., & Tsujii, J. (2004). Introduction: Named entity recognition in biomedicine. *Biomedical Informatics, 37*, 393-395. doi:10.1016/j.jbi.2004.08.011

Atkins, D., Droegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., Messerschmitt, D., & Wright, M. H. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

Bard, J. B. L., & Rhee, S. Y. (2004). Ontologies in biology: Design, applications and future challenges. *Nature Review Genetics, 5*, 213-222. doi:10.1038/nrg1295

Blaschke, C., Hirschman, L., & Valencia, A. (2002). Information extraction in molecular biology. *Briefings in Bioinformatics, 3*, 154-165.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research, 32*, D267-D270. doi:10.1093/nar/gkh061

de Bruijn, L., & Martin, J. (2002). Literature mining in molecular biology. In R. Baud & P. Ruch (Eds.), *Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications* (pp. 1-5). Ottawa, Canada: National Research Council of Canada.

Clark, T., Martin, S., & Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics, 5*, 59-70. doi:10.1093/bib/5.1.59

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6*, 57-71. doi:10.1093/bib/6.1.57

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563. doi:10.10381227561a0

Dalgleish, R., Flicek, P., Cunningham, F., Astashyn A., Tully, R. E., Proctor, G., … Maglott, D. R. (2010). Locus Reference Genomic sequences: An improved basis for describing human DNA variants. *Genome Medicine, 2,* 24. doi:10.1186/gm145

den Dunnen, J. T., & Antonarakis, S. E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation, 15*, 7-12.

Elmasri, R., & Navathe, S. (2000). *Fundamentals of database systems* (3rd ed.). Reading, MA: Addison-Wesley.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., … Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26,* 541-547. doi:10.1038/1360

Fundel, K., & Zimmer, R. (2006). Gene and protein nomenclature in public databases. *BMC Bioinformatics, 7,* 372. doi:10.1186/1471-2105-7-372

Gene Ontology. (2011). *GO ontology relations*. Retrieved from http://www.geneontology.org/GO.ontology.relations.shtml

Gene Ontology Consortium. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics, 25*, 25-29. doi:10.1038/75556

Genomic Standards Consortium. (2011). *Minimum Information about a (Meta)Genome Sequence.* Retrieved from http://gensc.org/gc_wiki/index.php/MIGS/MIMS

Go, K., & Carroll, J. (2004a). Scenario-based task analysis. In D. Diaper & N. Stanton (Eds.), *The handbook of task analysis for human-computer interaction* (pp. 117–133). Mahwah, NJ: Lawrence Erlbaum Associates.

Go, K., & Carroll, J. (2004b). The blind men and the elephant: Views of scenario-based system design. *Interactions, 11*(6), 44-53. doi:10.1145/1029036.1029037

Goll, J., Montgomery, R., Brinkac, L. M., Schobel, S., Harkins, D. M., Sebastian, Y., … Sutton, G. (2010). The Protein Naming Utility: A rules database for protein nomenclature. *Nucleic Acids Research, 38*, D336-D339. doi:10.1093/nar/gkp958

Gorman, M. (2004). Authority control in the context of bibliographic control in the electronic environment. *Cataloging & Classification Quarterly, 38*(3-4), 11-22. doi:10.1300/J104v38n03_03

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: An analysis of the Dryad Repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly, 47,* 380-402. doi:10.1080/01639370902737547

Gulley, M. L., Braziel, R. M., Halling, K. C., His, E. D., Kant, J. A, Nikiforova, M. N., … Versalovic, J. (2007). Clinical laboratory reports in molecular pathology. *Archives of Pathology & Laboratory Medicine, 131*, 852-863.

Higgs, P. G., & Attwood, T. K. (2005). *Bioinformatics and molecular evolution*. Malden, MA: Blackwell Publishing Company.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files* (CLIR Report 91). Washington, DC: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub91/pub91.pdf

Institute for Museum and Library Services (2011). *Specifications for projects that develop digital products*. Retrieved from http://www.imls.gov/assets/1/AssetManager/DigitalProducts.pdf

International Federation of Library Associations, IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR). (1998). *Functional requirements for bibliographic records: Final report*. (As amended and corrected through February 2009.) Retrieved from http://archive.ifla.org/VII/s13/frbr/frbr_2008.pdf

Joint Steering Committee for Development of RDA. (Ed.). (2009). *RDA scope and structure*. Chicago, IL: American Library Association. Retrieved from http://www.rda-jsc.org/docs/5rda-scoperev4.pdf

Juran, J. (1992). *Juran on quality by design*. New York, NY: The Free Press.

Khatri, P., Sellamuthu, S., Malhotra, P., Amin, K., Done, A., & Draghici, S. (2005). Recent additions and improvements to the Onto-Tools. *Nucleic Acids Research, 33*, W762-W765. doi:10.1093/nar/gki472

Krallinger, M., Erhardt, R. A., & Valencia, A. (2005, March). Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today, 10,* 439-445. doi:10.1016/S1359-6446(05)03376-3

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. *Genome Biology, 9*(Suppl. 2), S8-S8.14. doi:10.1186/gb-2008-9-S2-S8

Lambe, P. (2007). *Organising knowledge: Taxonomies, knowledge and organisational effectiveness*. Oxford, England: Chadons Publishing.

Liu, H., Hu, Z., Zhang, J., & Wu, C. (2006). BioThesaurus: A web-based thesaurus of protein and gene names. *Bioinformatics, 22,* 103-105. doi:10.1093/bioinformatics/bti749

MacMullen, W. J., & Denn, S. O. (2005). Information problems in molecular biology and bioinformatics. *Journal of the American Society for Information Science and Technology, 56,* 447-456. doi:10.1002/asi.20134

Martin, S., Hohman, M. M., & Liefeld, T. (2005). The impact of life science identifier on informatics data. *Drug Discovery Today, 10,* 1566-1572. doi:10.1016/S1359-6446(05)03651-2

National Endowment for the Humanities (2011). *Guidance for data management plans for NEH Office of Digital Humanities proposals and awards*. Retrieved from http://www.neh.gov/grants/guidelines/pdf/DataManagementPlans.pdf

National Institutes of Health. (2010). *NIH data sharing policy and implementation guidance* (NIH Publication No. 03-05-2003). Bethesda, MD: Author. Retrieved from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#goals

National Library of Medicine. (2009). *UMLS reference manual*. Bethesda, MD: Author. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK9675/

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century* (NSB Report No. 05-40). Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

National Science Foundation. (2007). *Cyberinfrastructure vision for 21st century discovery* (NSF Report No. 07-28). Arlington, VA: Author. Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

National Science Foundation. (2010). *Grant proposal guide* (NSF Publication No. gpg11001). Arlington, VA: Author. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpgprint.pdf

Open Biological and Biomedical Ontologies. (2011). *Archive of original principles*. Retrieved from http://www.obofoundry.org/crit_2006.shtml

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in Ecology and Evolution, 25*, 686-691. doi:10.1016/j.tree.2010.09.004

Patton, G. (Ed.). (2009). *Functional requirements for authority data: A conceptual model (FRAD)*. IFLA Working Group on the Functional Requirements and Numbering of Authority Records (FRANAR). IFLA Series on Bibliographic Control (Vol. 34). München: K.G. Saur.

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acid Research, 37*, D32-D36. doi:10.1093/nar/gkn721

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 33*, D501-D504. doi:10.1093/nar/gki025

San Gil, I., Hutchison, V., Frame, M., & Palanisamy, G. (2010). Metadata activities in biology. *Journal of Library Metadata, 10*, 99-118. doi:10.1080/19386389.506389

Seal, R. L., Gordon, S. M., Lush, M. J., Wright M. W., & Bruford, E. A. (2011). Genenames.org: The HGNC resources in 2011. *Nucleic Acids Research, 39*, D514-D519. doi:10.1093/nar/gkq892

Smalheiser, N., & Torvik, V. (2009). Author name disambiguation. *Annual Review of Information Science and Technology (ARIST), 43*, 1-43. doi:10.1002/aris.2009.1440430113

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*, 1251-1255. doi:10.1038/nbt1346

Stvilia, B., Gasser, L., Twidale M., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58*, 1720-1733. doi:10.1002/asi.20652

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*, 331–340.

Tillett, B. (2004). Authority control: State of the art and new perspectives. *Cataloging & Classification Quarterly*, *38*(3-4), 23-41. doi:10.1300/J104v38n03_04

Turner, P., McLennan, A., Bates, A., & White, M. (2005). *Molecular biology* (3rd ed.). New York, NY: Taylor & Francis.

Unified Medical Language System. (2011). *Metathesaurus release statistics*. Retrieved from http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

UniProt Consortium. (2011). *Protein naming guidelines.* Washington, DC: Author. Retrieved from http://www.uniprot.org/docs/nameprot

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., & Povey, S. (2002). Guidelines for Human Gene Nomenclature. *Genomics, 79*, 464-470. doi:10.1006/geno.2002.6748

Wand, Y., & Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-92.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–35.

Yilmaz, P., Gilbert, J. A., Knight, R., Amaral-Zettler, L., Karsch-Mizrachi, I., Cochrane, G., … Field, D. (2011). The genomic standards consortium: Bringing standards to life for microbial ecology. *The ISME Journal, 5*, 1565-1567. doi:10.1038/ismej.2011.39

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology, 29*, 415-420. doi:10.1038/nbt.1823