Literature Review

Table of Contents

2.1. Historical perspective

Early biological databases were created before the advent of the world wide web. That was in a world where biological data was shared by means of snail mailing data storage devices such as punched paper cards or magnetic tapes to scientific collaborators. Email and mass production of personal computers were yet to arrive on the scene. The first citation of digital data storage can be traced back to 1945 when the *memex* portmanteau was coined by Vannevar Bush using the words *memory* and *index* (Bush, 1945). This was done to give a name to a concept that envisioned people compressing and storing their books, communication and other records in a mechanized and easily retrievable manner in order to enhance human memory. Twenty years later the Cambridge Structural Database (CSD) started its transition from printed circulation to digital circulation (Attwood *et al.*, 2011). The CSD is recognized as one of the first scientific databases and as inspiration for the formation of the more widely know Protein Data Bank (PDB). The use of computers in the field of biology eventually led to the formation of the word *bioinformatics* (also known as computational biology or genomic data science) in the late 1970's. This marked the emergence of computer aided biological studies as a discipline in its own right. Shortly afterwards, in the early 1980's, scientists began to realize that the rate limiting step in nucleic acid sequencing was shifting from data acquisition to data management due to the emergence of faster sequencing technologies (Gingeras and Roberts, 1980). Collecting large amounts of scientific data and designing efficient search algorithms were no longer enough. Information resource interoperability became a central concern of bioinformatics during the mid 1980's when the volume of information being generated became more than databases of the time could handle (Robbins, 1996). The problems faced during biological data management has remained an area of concern since then. In the late 2000's the reason for this lack of progress in integration was ascribed to a lack of standardization and uncoordinated bioinformatic research rather than to the volume of information (Goble & Stevens, 2008). Currently, semantic web technology is being investigated as an integration architecture that would allow for the substitution of links between data documents with links between the underlying data, thereby decreasing search times and allowing abstraction of information (Machado *et al*., 2015). Abstraction is vital for the application of *black box medicine* in future clinical settings, that is, Decision Support Systems (DSS) that will match clinical data with broadly distributed and heterogeneous evidence based knowledge bases (Price, 2015) without requiring the user to have technical expertise in database querying.

2.2. Motivation

2.2.1. General observation

The volume of information from Next Generation Sequencing (NGS) is increasing at an exponential rate and so is the amount of information available downstream in the bioinformatics analysis pipeline. To make efficient use of the data becoming available to scientists and the general public, databases have to be integrated to improve user interaction by being readily accessible to multiple controller frameworks employing different standards and protocols. A conservative estimate that excluded non peer reviewed databases and commercial databases put the number of human related databases at 1550+ as at 2014 (Zou *et al*., 2015). These data sources are heterogeneous in content and globally distributed, which necessitates the need for providing Application Programming Interfaces (APIs) in addition to User Interfaces (UIs).

2.2.2. Regional context

The Southern African Human Genome Programme (SAHGP) is expected to produce South Africa's first bulk NGS data in the near future. Researchers will have access to this data in formats representing progressive stages of analysis. One of the final stages of NGS data processing is encapsulated in Variant Call Format (VCF) files. The information on genetic variation contained within VCF files will assist researchers not directly involved with the SAHGP in analyzing human specific experimental data and in the generation of hypotheses that could lead to research outputs benefiting the population of Southern Africa.

2.2.3. Problem statement

Conducting research using unstructured data is a time consuming process. As mentioned above, post genomic sequencing data is frequently processed and captured in VCF files. The content of files using this format is then relied upon by researchers to further investigate biological phenomena by querying, for examples, unstructured literature resources. However, VCF files are not presented in a user friendly format. Parsing of their content using software libraries such as VCFtools (written for Perl) and PyVCF (written for Python) is frequently required to extract useful information such as variant ID's that could serve as search query keywords. Moreover, literature retrieval search engines seem to (i) not provide an option to return relevant results based on the occurrence

frequency of a keyword, and (ii) not provide file upload facilities for bulk querying using keyword lists.

2.2.4. Research question

Can data contained within VCF files be integrated with the processes of keyword generation and search engine querying in order to automate bulk literature retrieval?

2.2.5. Point-of-view

Given the research question this project seeks to answer and the trend towards using semantic web technology, the current state of the literature will be critically analyzed from the standpoint that integration of biological databases with structured and unstructured data is necessary. The literature review that follows will argue that a gap exists in the process of biological literature retrieval that could be addressed by designing a tertiary data artifact with search engine functionality that integrates relevant information from *open access* scientific literature with parsed information from next generation sequencing data contained within VCF files.

2.3. Literature review methodology

2.3.1. Choice of academic search engine

In a comparison of Google Scholar (GS) with eleven other bibliographical databases, Walters (2009) found that GS outperformed its rivals by returning 41% of a set of preselected relevant records when all search results were taken into consideration. Although GS was not the top ranking academic search engine when examining hits of less than 75 using the same criteria, it should be noted that GS has a much larger indexed literature set. On average GS returned a total of 20400 records per search while the second most records returned by a competitor averaged at 311 records per search. A more specific examination of medical related record retrieval compared GS to three other major knowledge bases: Pubmed, Scopus and Web of Science (Falagas *et al*., 2008). These authors concluded that Pubmed remains an important resource for clinicians, but that the volume of information available through GS made it a better choice when searching for less known papers. However, they did point out that citation analysis using GS fell short of their expectations and a later study by Ortega and Aguillo (2014) determined that GS search results tended to be more

biased towards computing and information science. Given the scope of this project, the latter perceived shortcoming is not a cause for concern, but Ortega and Aguillo concluded that both GS and Microsoft Academic Search (MAS) should be used in conjunction with other citation indexes due to GS and MAS exhibiting technical limitations such as duplications and manipulations of results and citations. Their conclusion might have been too harsh given that Walters (2009) had already found GS to be a reliable indexer of relevant literature five years earlier. Moreover, it makes intuitive sense that paid access literature search engines like Scopus and WoS could be overtaken by free search engines such as GS and Pubmed because of an increase in awareness among researchers on how to obtain Open Access (OA) literature. Nevertheless, literature retrieved for the purposes of this study was sourced using both GS and Pubmed. Indeed, a subsection of this project relied on material acquired exclusively from Pubmed Central for the creation of a local database. One specific example of a Pubmed feature not available to GS users is the Medical Subject Headings (MeSH) term selector, which allows for filtering search results by selecting manually curated hierarchical keywords that are tagged to relevant articles.

2.3.2. Citation and rank analysis

Pubmed and GS have different approaches in terms of citation analysis and ranking of search results. These two properties have an effect on how article quality is perceived by the reader. Pubmed does not reflect citations in its search results, but instead lists articles in reverse chronological order by default (Falagas *et al*., 2008), while GS includes citation counts in its search results, but does not make its ranking criteria known to the public. In addition, GS offers citation analysis using the h-index formula, which is considered one of the two major citation indexes, the other being the Thomson Reuters Impact Factor associated with the Institute of Scientific Information (ISI) and the Web of Science (WoS) literature search engine. The h-index allows for both publications and authors to be ranked in terms of their productivity and apparent research output quality. Using the citation score of a journal as an indication of the average quality of an article in the journal could be appropriate, but caution should be exercised in using this metric for determining material relevant to a literature review. This is due to a well known skewness in individual journal quality observed within single publications (Garfield, 2006). Therefore, individual article citation counts was only used as a general indication of quality for GS results during this literature review. The preference for a more-recent-literature approach as seen in Pubmed's rankings was not strictly applied, although most articles were selected based on their date of authoring, especially while reviewing trends in technologies relevant to this project.

2.3.3. Keyword selection

While conducting searches using academic search engines, care should be taken during keyword submission to ensure that results will be efficient and reproducible. When entering search terms in GS and Pubmed the following guidelines should be kept in mind:

- Boolean operators must be in capital letters (NOT, AND, OR). Their order of execution is from left to right in the search string, but the order can be predetermined on Pubmed by using parenthesis.

- Search terms are case insensitive. It is therefore good practice to enter search terms using lowercase to distinguish them from boolean operators.

- The “AND” operator is implied and can be replaced by prefixing a plus sign to the term on GS or an ampersand to the term on Pubmed.

- The tilde sign (~) can be placed in front of a keyword to include synonyms in the search using GS. With Pubmed synonyms are automatically searched for when a term is suffixed with the MeSH operator “[mh]”.

- Avoid using adverbs, conjunctions and prepositions. These words, also called stop-words, are ignored by search engines unless they are preceded by a boolean operator.

- Quotes should be placed around phrases to search for keywords in their exact order while using GS. In contrast, Pubmed's search engine automatically searches for phrases, and researchers are advised to first attempt searching for phrases without using quotes.

- Enter plural word forms to search for the plural of a singular in GS. In Pubmed, word stems are automatically included in searches when using MeSH terms.

- Precede a keyword with “title:” when using GS to request that all results contain the keyword in the article's title. Follow the search term with “[ti]” to accomplish the same in Pubmed.

Primary keywords terms were identified by deconstruction of informative sentences in scientific articles acquired from publications such as the Bioinformatics journal (ISO4 abbr. *Bioinformatics)*. Secondary keyword terms were generated in a similar manner by using articles obtained from search results derived from primary keyword terms. The following sentence from a paper by Kamphans and Krawitz (2012) will be used for illustrative purposes:

A user that has a patient's informed consent to analyze the clinical data may upload sequence variants to GeneTalk in variant call format (VCF)

Three phrases and two individual keywords underlined in this sentence are possible choices for inclusion as search terms. However, additional qualifying keywords might be necessary to avoid ambiguity. For example, the keyword *vcf* should be followed by the keyword *bioinformatics* to retrieve results relevant to biological data instead of electronic business cards (vCards) that are coincidently stored in files with the extension *vcf*. Alternatively, the acronym can be replaced by the complete phrase, *variant call format*.

2.4. Scope and structure of the review

2.4.1. Scope

Although this survey of the literature has properties of a systematic review, the selection of relevant material was not trimmed until a manageable volume of material remained. Instead, an iterative approach was followed that consisted of:

1. keyword generation
2. search engine submission
3. intuitive selection based on title, blurb, citation scores, and publication date

The process was repeated until an argument could be sufficiently constructed from identification of:

- key trends and main theories
- proponents and opponents of themes
- examples from previous research and results

- all topics relevant to the project's implementation

2.4.2. Structure

The main body of the review that follows next will begin with a general discussion of major theories and models in database administration and biological data analysis. Critical analysis of topics and trends will be grouped together by theme. Benefits and drawbacks will be highlighted in the discussion of each topic or trend, and where relevant, followed by subdivision and clustering of author opinions into proponents and critics. Finally, a conclusion will follow that summarizes the main points of justification for the project.

2.5. Biological ontologies

2.5.1. Defining ontology

The science of ontology attempts to reconcile conflicting opinion regarding the difference between the abstract and the concrete by studying existence. The paradigm of ontology states that the problem of ambiguity in naming conventions has in its nature uncertainty brought about by conflicting schools of though regarding whether materialism or idealism best describes nature. However, the inherent nature of conflicting opinion already becomes apparent when attempting to define ontology. There is a multitude of definitions for this field of study with the following three equally correct definitions being put forward as applicable to biological database design in specific:

> *An ontology is a concrete form of a conceptualization of a community's knowledge.*
> (Stevens *et al*., 2000)

> *A precise explanation of one's terms and reasoning in some subject area, which can allow computers to help, is called an ontology.*
> (The World Wide Web Consortium, 2015)

> *An explicit, formal representation of concepts and relationships among them within a particular domain that expresses human knowledge in machine readable form.*
> (Martone, 2012)

A simple explanation for what an ontology is in the domain of bioinformatics can be derived from the above definitions: it's a concept dictionary that can be used by both humans and machines. Moreover, an ontology is composed of a vocabulary of words together with a specification of their meaning, and an ontology has to be encoded using a Knowledge Representation (KR) language so that the ontology can serve as a data template. The semantic clarity that an ontology provides through KR encoding has to be balanced with available time and resources because creating ontologies is a time consuming process (Martone, 2012). The following diagram (Figure 1) illustrates a simple biological ontology using the subject of this project, DeoxyriboNucleic Acid (DNA), as an example:

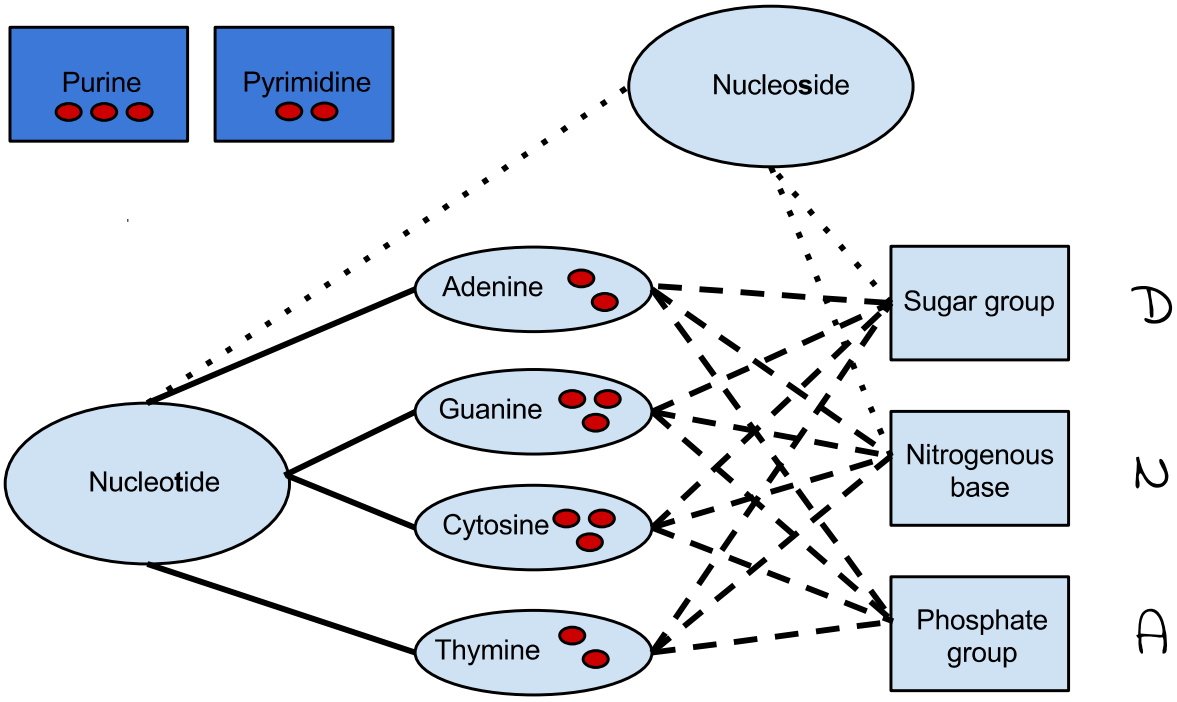

Figure 1: An ontology of the sub-molecular structure of nucleotides

In the above ontological illustration the difference between nucleotides and nucleosides are emphasized using solid lines to indicate sub-classes (is_a relationships), while the non-solid lines refer to compositional relationships (has_a relationships). This exercise in deconstruction assigns compositional members to the class nucleosides (dotted lines) and the class nucleotides (striped lines) by using three biochemical functional groups that can form part of either of these two types of organic molecules. The red dots (pairs and triplets) in some of the classes indicate the amount of hydrogen bonds that each member of the class have that can engage in non-covalent bonding. This exemplifies the value of having automated reasoners since if a relationship has not been explicitly declared such as illustrated with the lack of connecting lines for the Purine and Pyrimidine classes

(e.g. adenine is a member of the class pyrimidine), the relationship can be inferred by automatic reasoning based on the cardinality of the hydrogen bonding property. For example, if a cardinality rule states that all members of the class purine must have three and only three hydrogen bonds capable of forming non-covalent bonds in the context of DNA strand formation, then a computer will be able to determine that guanine and cytosine must be purines, while thymine and adenine cannot be purines. Automated reasoners are essential in the design of larger ontologies that may contain thousands of classes and will be discussed next.

2.5.2. Encoding an ontology

Ontologies can be prone to incongruences such as classes that cannot be instantiated or conflicting DL specifications. For example, imagine that a relationship between two entities has been declared as functional (being a single valued property). If a third entity is assigned the same relationship that exists between the first and second entity, the implication is that the second and third entities are the same entity. However, if the second and third entity have been explicitly declared as distinct entities, an inconsistency arises. Automated reasoners can be used to avoid these type of idiosyncratic artifacts as well as to infer relationships that were not manually declared. The specification of an ontology is often implemented with Descriptive Logic, although some reasoners use a more reduced Propositional Logic (PL). Fact++ (Tasrkov and Horrocks, 2006) is an example of a automated reasoner relying on *SHOIQ* description logic (DL notation is printed in a *grotesque* typeface by convention (Szeredi *et al*., 2014)). This well developed reasoner is packaged with the widely known Protégé ontology editor (http://protege.stanford.edu). A popular KR language is the Resource Description Framework (RDF) triple-based representation system, but it is considered by some to be semantically weak (Schulz and Jansen, 2013). Triple refers to the fact that relationships between concepts are represented by Subject-Predicate-Object (SPO) connectors where the object and subject are concepts, and the predicate is a relationship encoded in Description Logic (DL) syntax. The specification of an ontology usually relies on a combination of the following Boolean constructors and restrictions (Staab and Studer, 2013; Handke, 2015):

Conjunction ($\wedge$, $\&$, *and*)

Disjunction ($\vee$, *or*)

Negation (¬, ~)

Equivalence (≡, ↔)

Material implication (→...⊃, *if...then*)

Universal restriction (∀)

Existential restriction (∃)

2.5.3. Applying ontologies

Once the ontological template has been populated with data instances, it becomes a Knowledge Base (KB). KBs serve as the foundation of the semantic web and they are a distinct form of databases in that the relations between data in KBs are semantically enriched. Although ontologies are different from database schemata for the same reason, ontologies can be useful in databases design. In essence, ontologies are re-usable and database schemata are once-off blueprints. It follows that ontologies have multiple purposes. Stevens (2000) lists these purposes as: providing a community reference system, defining database schemata, providing ontology based search tools, and supporting natural language processing. Examples of applied ontologies in the biological sciences include: RiboWeb, EcoCyc, MBO, GO and TaO. BIOlogical PAthway eXchange (BIOPAX) structured language is an example of an effort to specifically control biological vocabulary to enhance communication between databases (Demir et al., 2010).

2.5.3. Upper ontologies

Research is also being conducted into creating what are called u*pper ontologies* (Soldatova and King, 2005). In contrast to subject specific ontologies such as biological ontologies, *upper ontologies* seek to anchor all other ontologies by serving as a root ontology. The Institute for Electronics and Electrical Engineering (IEEE) investigated the viability of such ontologies by creating the Suggested Upper Ontology Working Group (SUO-WG). They concluded their research by recommending further investigation into three candidate upper ontologies (Poli *et al*., 2010): the

Information Flow Framework (IFF), the Suggested Upper Merged Ontology (SUMO) and the Upper Cyc Ontology (UCO). An example of a browser that was subsequently designed to navigate SUMO terms during ontology construction is SIGMA-KEE (http://sigmakee.sourceforge.net/). There is however a distinction to be made between ontology construction assistants such as SIGMA-KEE and the Protégé ontology editor mentioned earlier. Employing both these tools could assist the user in determining relevant standardized vocabulary from SUMO in the case of SIGMA-KEE, which can then be used to create structural and composition facets of an ontology in Protégé.

2.5.4. Ontology design obstacles

Typical issues that can arise during ontology construction are highlighted by example with Protégé ontology editor's accompanying “Pizza” tutorial. These stumbling blocks include: determining valid membership of a class, avoiding ambiguity in assigning names to classes, defining classes using the correct quantifiers/properties, distinguishing a class from an instance, preventing duplication due to multiple inheritance, and separating structural from compositional relationships.

2.6. Biological database design

2.6.1. The need for data storage

Obtaining results in many lines of work depend on crunching data: from gathering valid raw data through observation in the field to making discoveries in the lab. Appropriate databased analysis is also a cornerstone of scientific progress. One of the most challenging issues to deal with in biological database design is the necessity for continued maintenance of integrated databases due to information and software evolution. IBM (2012) estimated that 2.5 quintillion bytes of data was generated on a daily basis and that 90% of the data in the world had been generated in the two year period leading up to the estimation. The bulk of new data is generated by the Internet of Things (IoT) as less human-to-human and human-to-machine interaction is required for information to be transferred over networks. This leads to an accumulation of stored information in every data warehouse imaginable including biological databases.

2.6.2. Structured, semi-structured and unstructured data

Blumberg and Atre (2003) estimated that structured data accounted for approximately 15% of data

in existence, but a more recent estimation by Cisco (2014) states that approximately 90% of data is either unstructured or semi-structured. Although there is some disparity about the percentage of structured data, there is a consensus that structured data is much less abundant than unstructured data. In computer science, data structures can be composed of primitive data types such as integers, composite types such as arrays, or abstract data types such as stacks. The type of a data element determines the set of values that the data element can represent (Wirth, 1985). Structured data is data that can be readily read, but not necessarily semantically interpreted, by machines. It is *this* type of data that usually resides in databases, data warehouses, customer relationship management systems, enterprise resource planing systems, or XML documents (Cisco, 2014). In contrast, semi-structured and unstructured data is usually formatted as word processing documents (such as this thesis), emails, audiovisual files, or social media feeds. Semi-structured data can be created by organizing unstructured data such as word processing documents into a hierarchy (taxonomy) or contextualizing unstructured data files with meta-data (Blumberg and Atre, 2003) in a process called data classification. Automated data classification has grown in popularity over expert curated text classification due to the considerable time-savings that accompanies automation (Sebastiani, 2002). This transition has led to an increased interest in the field of Machine Learning (ML), which is defined as automated general inductive processes, either supervised or unsupervised, that build classifiers by determining the characteristics of a category through analyzes of sets of preclassified documents. Software packages such as Scikit-learn (Pedregosa *et al*., 2011) built with the Python language has been created to assist non-technical users with employing ML techniques in their research since creating ML algorithms requires substantial knowledge of statistics.

2.6.3. Broad classification of biological databases

Biological databases can be classified according to data type and use. Sub-categories that emerge from classification include sequence, structure, proteomic, interactomic and genomic databases (Cannataro *et al*., 2014). The information they contain represents a data layer between bioinformatics and molecular biology. The flow of information from the laboratory to data processing and interpretation can only be effective if the data model is appropriate. Traditional database design focuses on data models, declarative query language use, high throughput mechanisms and reliability (Cui *et al*., 2014) More recently, designing databases with the aim of assisting users through data-centric decision making has been given a central role in development. This is likely due to the sheer volume of information that has become available. When databases grow in terms of size and count, four areas of concern become apparent: data storage and scope,

processing capacity, level and method of digital curation, and data exchange facilitation (Zou *et al*., 2015). As mentioned in the motivation for this review, there are now more than 1500 human-related databases. So rather than attempt to provide a complete description of all the categories and sub-categories of biological databases, it would be more prudent to discuss projects that aims at indexing available databases. They can be conceptualized as databases of databases and will be referred to as *upper databases* for the rest of this review. The Nucleic Acid Research (NAR) upper database is called the Molecular Biology Database Collection (http://www.oxfordjournals.org/our_journals/nar/database/c/). This database had 1512 databases with fourteen main categories and 41 sub-categories indexed as at 2013 (Yu *et al*., 2015). On 26 August 2015 one additional main category could be viewed on the NAR upper database. The NAR publication has a Thomson Reuters impact factor rating of 9.112, a five-year impact factor rating of 8.867, and has been rated by the Special Libraries Association (SLA) as within the top 100 most influential journals. The 1512 databases indexed in the NAR upper database can therefore be assumed to be a conservative estimation. Some authors have even created indexes of upper databases. Bolser *et al*. (2015) tabulated ten upper databases as being similar in scope to their upper database called MetaBase.

2.6.4. Examples of small-scale experimental biological databases

Tangible User Interface (TUI) design has been put forward as a possible solution to help users cope with big data. This approach attempts to bridge the gap between the three-dimensional world and digital representations as a way to make user interaction more intuitive by *rich representations*. Eugenie++ is a TUI system that provides users with physical objects that can be manipulated on a horizontal multi-touch surface (Grote et al., 2015). The objects represent various navigational and logical methods that can be used to traverse the Massachusetts Institute of Technology (MIT) registry of biological parts and Pubmed. Genetalk is an example of a web application that serves as a communication platform by providing users with the opportunity to annotate genomic positions with a diverse range of data in a conversation-thread like manner (Kamphans and Krawitz, 2012). In essence, each annotated variant becomes a blog and the scientific community then comments on each variant to exchange clinical and experimental findings with other individuals in the community. This could include users providing links to the literature, reporting annotation errors, or creating help request tickets. One drawback is that there is a heavy reliance on user participation for the curation and continued viability of the application. Moreover, no reference is made by Kamphans and Krawitz as to how the level of user participation was quantified. Determining user participation levels is important because previous studies (Ives and Olson, 1984) pointed out that

there is not necessarily a positive correlation between user participation and information satisfaction.

2.6.5. Database schemata and data types

When a scientific problem has been formulated from observation and the manner in which data collection will be carried out has been established through experimental design, the next step for a scientist is to think about how data will be stored and presented for subsequent analysis. The database selection process should take into consideration the structure of the data that needs to be stored, but also the needs of the intended users of the data. The first step might be to define data types that will be stored together with their associated fields. This process is sometimes referred to as designing a database schema. The range of required data types will be influenced by the nature of the data to be collected and the manner in which the data will be accessed. Taking the project that this literature review motivates as an example, if a single alpha-numeric term is to be used during querying of a database, then a key-value, hierarchical database will be more appropriated than a relational database. Similarly, if all data in a database can be constrained as type string, a database management system (DBMS) could be prudently replaced with a native database where type declaration and conversion is implemented by the designer using scripting languages. The main disadvantage of using a non-standardized database over a DBMS is that reliability is sacrificed for efficiency (Stonebraker, 2010a), that is, the following four components (Figure 2) of ACID-compliance is compromised to an extent:

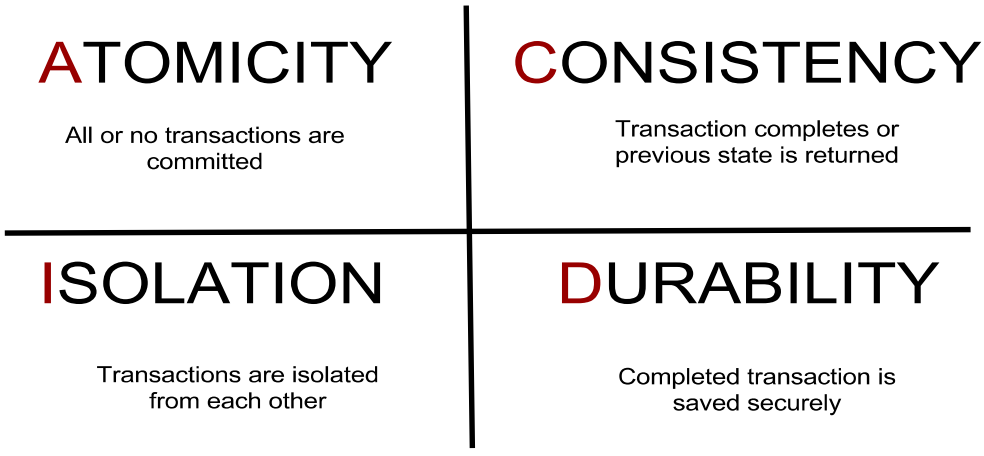

Figure 2: ACID compliance ideogram

There is a significant overlap between ontology design and database schema design, specifically

during construction of Entity Relationship diagrams (Kesh, 1995) and in a broader sense during construction of Unified Modeling Language diagrams (Alkoshman, 2015). As mentioned earlier, ontologies are considered to be more generic than database schemata and it is for this reason that ontologies can supplement database design. Database schemata are intended to assist administrators and to some extent users in interpreting large databases not only during design, but also during maintenance as well as updating and querying of such databases (Di Battista *et al*., 2002). Graphic representation of database schemata can assist in decreasing the complexity of lines upon lines of code, but can be a time-consuming process. Di Battista *et al*. (2002) proposed an automated graphic representation framework that would mediate XML data file submissions to a graph drawing application hosted on an off-site server with API functionally. Following processing on the external server, the same mediator would then return graphs of the submitted schema to the system hosting the schema. Graphviz (Ganser and North 2000) is an example of a more comprehensive graph drawing suite that can be customized according to the needs of the developer. The Graphviz application is coincidently also offered as a plugin to the Protégé ontology editor that was discussed in the biological ontologies section. Database schemata can be subdivided into at least three major categories: relational, hierarchical and graph based schemata (LinkedDataTools.com, 2015), with Leavitt (2010) including object-orientated schemata as a fourth type. Figure 3 illustrates the difference between the major schemata.

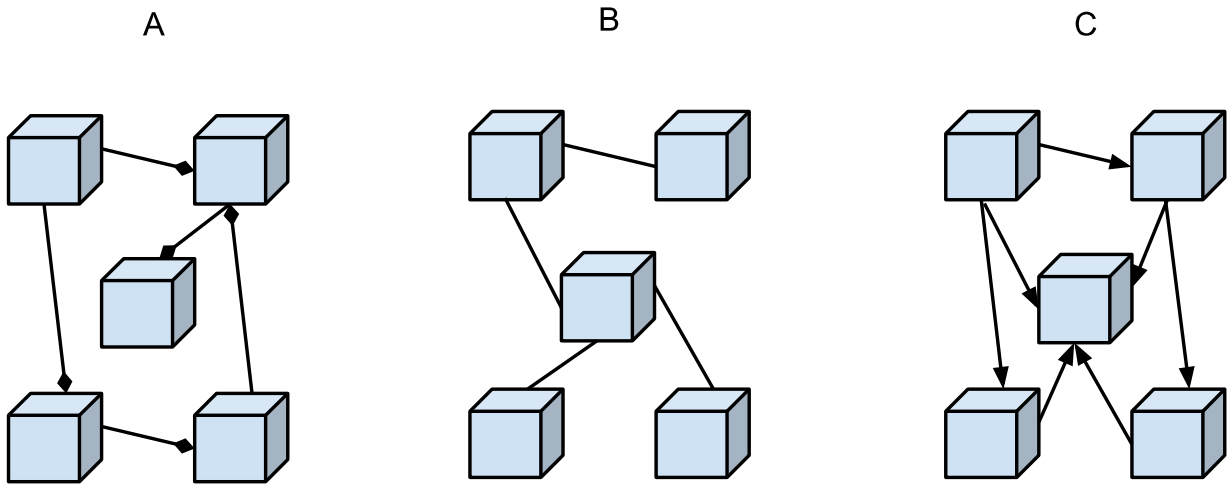

Figure 3: Graphical representation of the major database schemata: (A) relational (B) hierarchical (C) graph-based

In the above illustration the cubes represent classes and the connecting lines represent relationships between the classes. The black diamonds in Figure 1A indicate that the class to which it is nearest is in a compositional (“has-a”) relationship with the class connected to the other end of the line. It is

this latter class that is the parent class. The relationships depicted in Figure 3A are typical of those found in relational databases where the classes represent the contents of tables and the relationships indicate foreign keys references. Figure 3B in contrast depicts a hierarchical structure where parent nodes have more intrinsic value and the structure of the illustrations should adhere to the principal of orthogonality (Smith, 2008) that prohibits multiple inheritance and seeks to minimize the amount of classes (Ross *et al.*, 2005). Graph based schema as illustrated in Figure 3C is characterized by arbitrary relationships that does not make distinctions based on intrinsic importance. Furthermore, graph based schemata are the choice schema for semantic web integration planning (LinkedDataTools.com, 2015).

2.6.6. Database management systems

Relational databases are best suited for storing structured data (Leavitt, 2010) and as such are queried with Structured Query Language (SQL). In contrast, non-relational databases (colloquially referred to as NoSQL databases) are designed to handle unstructured data more efficiently than relational databases. A second characteristic of NoSQL databases is that they can be employed in distributed environments more easily than relational databases, thereby reducing the financial cost associated with isometric scaling. The three major categories of NoSQL databases according to Leavitt (2010) are: key-value, column-based, and document-orientated. Some authors make additional distinction by categorizing graph-based databases as a fourth type of NoSQL database (Tweed and James, 2010; Ponzanni, 2013). By comparing schemata categories as defined by Leavitt (2010) and LinkedDataTools.com (2015) in the previous section with NoSQL categories as defined in this section, it could be argued that there seems to be a consensus that at least key-value and column-based NoSQL databases are most appropriately conceptualized with hierarchical schemata and should therefore necessarily adhere to the principal of orthogonality and avoidance of multiple inheritance. Key-value databases are typified by nested key-value pairs, that is, each value acts as a column that can in turn be composed of a key-value pair. Conversely, column-based NoSQL databases typically have a single column that contains closely related key elements, while document-based stores do not put a constraint on the amount of columns.

NoSQL databases share the main disadvantage that non-standardized databases exhibit in that ACID-compliance is compromised. However, this shortcoming only becomes a concern when a database is distributed (Ponzanni, 2013). Brewer's theorem (more recently referred to as the CAP theorem) is an attempt to substitute the lost benefits of ACID-compliance (Stonebraker, 2010b;

Pokorny 2013), but adherence to this theorem will not be necessary for the purposes of the current project since the proposed system will not be distributed.

2.6.7. Managing software dependencies

Large scale, distributed software applications require detailed planning of software deployment related activities that follow the end of a project's development phase (Dearle, 2007). These activities can be subdivided into an installation phase (release, configuration and activation of the software) and a post-installation phase (monitoring, updating, deactivation and redeployment). Smaller projects require less stringent planning of the post-development phase, but managing change remains important because responding to change is considered more important in the world of software development than sticking to a predefined plan (Manifesto for Agile Software Development, 2001). Software management can be complex where the purpose of changes made is extended to include managing temporal and system properties, the various aspects of the object of change, and facilitating change events in itself (Buckley *et al*., 2005). Or, it can be limited by practical considerations to deciding on a versioning number scheme such as the Preston-Werner (Figure 4) semantic versioning scheme (Raemakers *et al*., 2014).

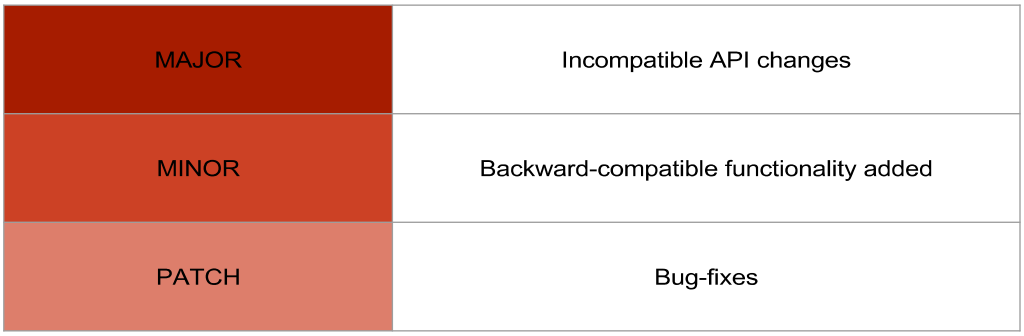

Figure 4: Preston-Werner semantic versioning scheme

The Preston-Werner scheme illustrated above follows the MAJOR.MINOR.PATCH naming convention where the release API is assigned the identifier 1.0.0 and each of the three digits will cycle up by one if the condition to the right of the respective categories has been met. To ensure interoperability in a system that employs software from multiple contributors, dependency registers and automated package managers provide the ability to preserve the state of a set of software components (Abate *et al.*, 2011) in time and handle many of the post-development phase activities

on behalf of the developer. Three exemplary package managers (Bevacqua, 2015) that provide functionality in a Javascript environment, such as would be the environment for this project, are the general purpose Node Package Manager (NPM), and the front-end specializing package managers Bower and Component.

2.6.8. User participation

The software development approach of this project is *mashup*-orientated rather than service-orientated or component-orientated. The main driver behind mashup development that distinguishes it from the other more conventional development paradigms is that it places an emphasis on user innovation (Capiello *et al*., 2011) promoted by the availability of *open services* such as Application Programming Interfaces (APIs) provided by, for example, Google and Twitter. In other words, during mashup development, the developer is also a user to some extent. This concept, of creating cycles of data re-use from heterogeneous resources including user generated content, is a cornerstone of what is popularly described as *Web 2.0* (O'Reilly, 2007). The mashup that will flow from this project will act as proof of concept of this development approach for a specific combination of publicly available resources viz. the Pubmed Central open access initiative and an open source JSON API. Capiello *et al.* (2011) further defines the characteristics of a mashup as: having a domain specific focus, providing an abstraction from detail, and providing immediate visual feedback on user actions. The envisaged product of the current project will meet all these characteristics.

User participation has long been considered essential to the success of software systems (Ives and Olsen, 1984), however, these authors also observed that proponents of user involvement rarely put forward empirically derived arguments based on strong theory. A three part series of papers discussing qualitative methodologies in the health care sciences included a well argued paper advancing the use of ethnography (Reeves *et al*., 2008). In their paper that acts as a guideline for determining behavioral, temporal and spatial dynamics in small subsets of populations, the authors suggest using nine observational dimensions during ethnographic research. Following these guidelines to design a questionnaire to determine the needs of biological data users at the South African National Bioinformatics Institute (SANBI) could be a prudent way to mitigate the possibility of inherent methodological flaws in user involvement quantification as pointed out by Ives and Olsen (1984).

2.7. Semantic web integration

2.7.1. Communicating science

Biological scientists more often than not use prior knowledge to make inferences about the function of unknown entities rather than use axiomatic rules (such as contained in formulas and equations) to gain further information (Stevens *et al*., 2000). This is why databases are important. In addition to the need for prior knowledge, scientists also need communication knowledge; they need to have reliable and efficient access to databases to aid in data comparison as well as in defining and constraining the data at their disposal. Ideally all of the available biological information should be integrated so that all of humanity's collected knowledge about biology can contribute to future biological research, but integrating all available biological knowledge is difficult because designing an all encompassing database inevitably leads to a series of compromises that sees a loss of information due to scientific and political ideals of different databases coming into conflict (Stein, 2003). A practical difficulty in establishing links between databases include putting in place standardized naming conventions for biological objects and concepts. That is, the same biological object is sometimes given different names and different biological concepts are sometimes given the same name.

2.7.2. Data virtualization

An alternative to physical integration of databases is called data virtualization (also know as data federation). The central idea of this technology is that heterogeneous resources could be utilized in combination by only keeping track of the location of data (Machado *et al*., 2015), rather than housing the actual data . The inverse of data federation is query federation, where single or multiple queries are combined and sent to a predefined set of databases. In addition to this alternative to data integration, data interoperability can be promoted by technologies that allow the underlying data in databases to be connected as discussed in the section on ontologies. An example of how tools could be employed to facilitate knowledge discovery by enhanced data integration and interoperability is by embedding HTML content with RDFa (Goble and Stevens, 2008) to enrich web content with semantic metadata tags from standardized vocabularies such as Dublin Core (Weibel *et al*., 1998) and FOAF (Brickley and Miller, 2012). Semantic web schemata overlap with ontologies in a similar manner to how database schemata overlap with ontologies. The unifying similarity is the absence of actual data populations. Indeed, ontologies such as the Web Ontology Language (counter intuitively

abbreviated as OWL) and OWL-S (Martin *et al*., 2004) have been specifically designed to encourage data integration using ontologies, and in the case of OWL-S where the "S" refers to "service", to promote automated discovery, invocation and interoperability of such web services.

2.7.3. Linked data

Over the last thirty years, linking data over the Internet, and before that over ARPANET, transitioned from flat files being served using File Transfer Procotol (FTP) to the current day exploitation of Hyper Text Transfer Protocol (HTTP) – the traditional language of the world wide web – to create *Linked Data* communication structures (Sheridan and Tennison, 2010). Linked Data (LD) is characterized by four defining principals as put forward by Tim Berners-Lee (who is credited as the inventor of the world wide web), namely: that Uniform Resource Identifiers (URIs) should be used to identify real world objects, that these URIs are in HTTP notation to ease dereferencing, that the data should be enriched using RDF type standards, and that reference should be made to other objects within the data. The idea that modern data repositories should conform to the concept of LD has gained so much traction that a five-star rating system, devised by Berners-Lee (Janowicz *et al*., 2014), is frequently being used to gauge LD vocabulary compliance, where stars are successively awarded for meeting compliance criteria as follows:

- ⭐ The data is accompanied by human readable dereferencable data.
- ⭐ The data is accompanied by machine readable dereferencable data.
- ⭐ The applied vocabulary links to other vocabularies.
- ⭐ Metadata about the vocabulary is available.
- ⭐ Other vocabularies link to the applied vocabulary (contrast with third star).

2.7.4. Representational State Transfer (REST)

Since Linked Data (LD) is inherently RESTful (Sheridan and Tennison, 2010), providing access to such data with servers that comply to REST constraints would, at least conceptually, decrease the complexity of such a service. These constraints of REST based implementations (that could also be

referred to as RESTful implementations) was discussed in detail by Pautasso (2014) in terms of differing levels of maturity and with reference to the main concepts of addressability, stateless interactions, uniform interface, self-describing messages, and hypermedia. The latter concept supports the interpretation that LD is inherently RESTful, but extends it to distinguish web services (or APIs) that comply with this constraint by assigning them the term Hypermedia APIs. The service that will be provided by the current project will place an emphasize on making content discoverable and could therefore be classified as an Hypermedia API. A third group of researchers advocated (Meng *et al.*, 2009) the use of RESTful web services over traditional web servers after empirically researching their respective suitability in distributed data integration. RESTful web services seem to not have any particularly drawbacks that need to be factored in during development besides from the fact that it relies heavily on the HTTP protocol which might or might not be replaced in the future.

2.8. SNPs: the determinant of genetic variation

2.8.1. Defining SNP based research

Genetics is the study of biological inheritance and subsequent variation among and within individuals, while genomics is more specific in that it concerns itself with the structure, function and mapping of genomes. A consortium led by Craig Venter revealed in 2001 that the human genome contains approximately three billion base pairs (Venter *et al*., 2001), with the bases being one of four naturally occurring nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). There is a fifth abundant monomer of nucleic acid called uracil (U), but this monomer is a constituent of RNA, not DNA. Uracil, which is derived from the deamination of cytosine (Brown, 2007), is here referred to as a nucleic acid monomer entity because the nucleotide uridine is thermodynamically unstable (Peña, 2015). It is important to note that uracil is not a polymorphism of thymine in the context of the term Single Nucleotide Polymorphism (SNP). Instead, substitution of A, C, G and T with each other in a strand of DNA represent SNPs. Together with Restriction Fragment Length Polymorphisms (RFLPs) and Simple Sequence Length Polymorphisms (SSLPs), SNPs are one of three types of DNA markers that are especially useful. These markers characteristically include at least two alleles (Brown, 2007). The quantity of SNPs vastly outnumber that of other DNA markers because of its experimental through-put being higher than that of RFLPs (low though-put) and SSLPs (medium-throughput) (Scarano, 2014). SNPs arise as a results of DNA mutations, which in turn is caused by errors in DNA replication or as a result of mutagens acting on

DNA (Brown, 2007). However, a second important note to make is that less than 1% of mutations that fall in the SNP category are functional (Venter *et al*., 2001), that is, they lead to a change in the function (gain or loss) of a protein that is synthesized by ribosomes using RNA that has been transcribed from the encoding DNA carrying the mutation.

2.8.2. SNP related literary resources

Conventions used to represent elucidated genetic data differs from one resource to another. The International Cancer Genome Consortium (ICGC) (https://icgc.org/icgc), lists data categories for variation together with their respective entry availabilities as it applies to sample generated information from 12979 donors. Of these, the database has Simple Somatic Mutation (SSM) category data for 8038 individuals, making the ICGC seems like the ideal place to search for literature on a given SNP using its identifier (e.g. rs1799950 – an SNP in the *BRCA1* breast cancer related gene), but a search for this rs identifier returns no results. When searching for this term using a more generic database such as dbSNP (that forms part of NCBI and therefore the Pubmed and Pubmed Central resource repositories), relevant results are returned for a plethora of information after which data in related NCBI databases can then be retrieved by selecting a database from a drop-down list. Selecting Pubmed from the drop-down list provides the user with a subsequent list of articles that relate to the SNP, with the option to navigate to related articles, however, no indication is given as to the relevance of the article to the identifier if more than one article is associated with the identifier. Instead, literature is sorted with a preference for more recently publicized articles (as discussed earlier in this review). Searching for the SNP by its identifier in non-institutionalized databases such as SNPedia (http://www.snpedia.com) returns results that point the user to literature relevant to the SNP, with an indication of whether this literature is provided under Open Access agreements, but the order in which the results are listed does not give the user an indication of how relevance was determined. In addition, the literature that SNPedia refers the reader to does not give an indication of the date of publication. Interestingly, SNPedia does link the identifier with keywords and provides quantification by mentioning other SNPs in context (e.g. “This SNP, a variant in the BRCA1 gene, is 1 of 25 SNPs reported to represent independently minor, but cumulatively significant, increased risk for breast cancer.”). However, specific literature ordered by quantified relevance is still out of reach, and more importantly, bulk querying of SNP identifiers is not available as part of the user interface.

2.8.3. VCF files

According to Venter *et al*. (2001) the human genome contains approximately four million polymorphisms (DNA bases with at least two alleles). This is much lower than the three billion base pairs present in the human genome, and a specialized file format containing only data on the occurrence of variation would therefore remove 99.8% of data generated during sequencing and found in, for example, FASTQ files (Cock *et al*., 2010), all of which would be redundant in the context of variation. Towards this end, Variant Call Format (VCF) files were developed to cater for variant analysis, not only among individuals, but also across multiple samples (Danecek *et al*., 2011). This file format was specifically developed for the 1000 Genomes Project, where the need for removing redundant raw sequencing data became necessary to efficiently compare results from the sequenced genomes of 1000 individuals. The following chart (Figure 5) illustrates the importance of reducing file size by plotting 1000 genomes compressed VCF file sizes as a function of human chromosome numbers:

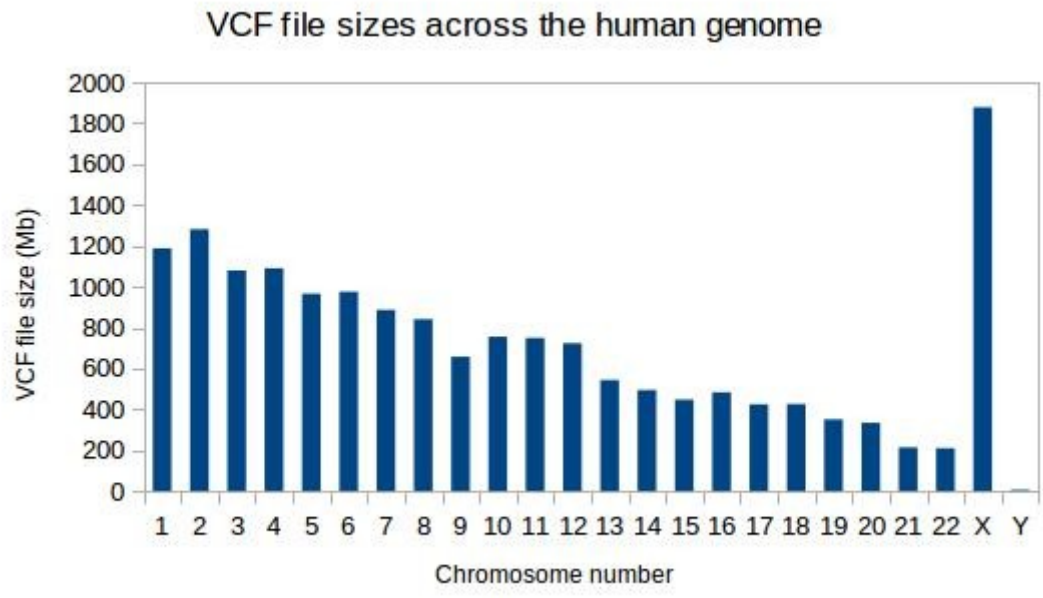

Figure 5: VCF file sizes for each of the 46 (23 x 2) human chromosomes

The above chart was compiled using data from the 1000 Genomes Project repository (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). The resulting plot depicts an overall decrease in data volumes across a sequential range of the human karyotype, that correlates to a decrease in chromosomal length (not depicted), with the X-chromosome seemingly overrepresented and the Y-Chromosome seemingly underrepresented in terms of data volumes. The cause for the sex chromosome related data bias is not clear, but the overall size of VCF files indicate the need for

removal of non-variant data, not only for the 1000 Genomes Project, but also for any other projects exceeding 1000 individuals. Although VCF files are more succinct than raw sequencing data in the context of variation, further parsing could be of benefit when, for example, only the ID column of a VCF file is of interest to the researcher. Software that allows for further parsing could be in the form of a standalone package, such as PyVCF (http://pyvcf.readthedocs.org/en/latest/index.html), or integrated as a tool in online bioinformatics platforms such as Galaxy (https://galaxyproject.org/).

2.9. Open access initiatives: literary sources of secondary biological data

2.9.1. Defining and motivating free articles

Open Access (OA) was defined by the Budapest Open Access Initiative (OAI) convention of 2001 as follow:

> *The literature that should be freely accessible online is that which scholars give to the world without expectation of payment...*
> (BOAI, 2002)

Publishing models based on this believe is becoming increasingly popular. Official figures show that OA articles are increasing at a rate of about 10% per annum and an estimated 71% of biomedical research articles published between 2011 and 2013 are currently available through OAIs (Archambault *et al*., 2014). In their report on OA proportions to the European Commission, these authors ascribed growth in the amount of OA papers to four drivers: increased interest in OA leads to more new papers being published under OA licenses, more paid papers becoming OA licensed for the same reason, expiration of embargo periods during which access to scientific literature is restricted, and an increase in the amount of overall published scientific papers per annum. The increase in freely available articles represent a new source of information for scientists that can benefit research in many ways. This is because paid-access models often only give free access to the abstract of an article, while in open access initiatives the reader has free access to the entire article. Shah *et* al. (2003) concluded that when creating a subset of literature for a database, access to the full text of an article is preferable to just having free access to the abstract. The following illustration (Figure 6) highlights the crux of their finding by incrementally saturating the dispersion of gene names per article subsection based on the likelihood that a gene name will appear in a given subsection:

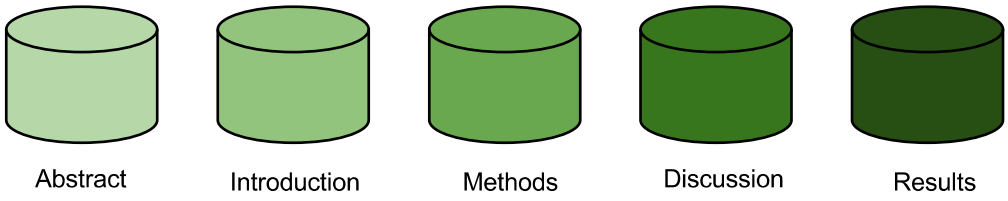

Figure 6: Gene name occurrence frequency depicted with incremental saturation (after Shah *et al*., 2003)

If cost is a limiting factor for a reader who is interested in determining the relevance of thousands of articles based on the occurrence of a gene of interest, as was the case in the study that the data for the above figure was drawn from, availability of free access to to all subsections of an article will be a primary determinant of the feasibility of a project.

2.9.2. Open access categories and licensing

The quality of free publications can vary widely and therefore OA material is often placed into the following categories: gold (e.g. listed in the Directory of Open Access Journals), green (listed in directories such as OpenDOAR that caters for self-archived material), others (papers available through large databases such as CiteSeerX and Pubmed Central) and rogue material (papers published without consent (see *US vs. Aaron Swartz* (2013))). The legality of OA is ensured by obtaining a copyright owner's consent with a license, such as the Creative Commons license, that may allow users to read, copy, download, link, crawl, and search articles in OA repositories (Suber, 2004). Persuading researchers to provide their written works under such licenses is not considered a hurdle to overcome since there is a long standing tradition in the scientific community to write articles for impact rather than profit. This tradition dates back to the creation of the first scientific journals in 1665 and benefits science by increasing the scope of publication through reduced publication cost. However, the current percentage of green open access journals only amount to between 10% and 20% of all published articles (Harnad *et al.*, 2008), with green OA papers constituting 90% of all OA material. Employers and funders who commit to mandating self-archiving could have a significant impact on the availability of green OA. Funding of OAIs sometimes rely on open source software and software technicians donating small amounts of their time. Suber (2004) mentioned that OAIs can be financially maintained by being granted space on university servers. Hosting of OA material in this way is not considered to be completely altruistic since the hosting institution gains by increasing its research output and visibility.

2.9.3. Pubmed open access corpus

The PubMed Central Open Access Initiative (PMC-OAI) is an example of an OAI repository using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It contains partial or complete access to articles published in a list of journals available at: http://www.ncbi.nlm.nih.gov/pmc/journals/. A default HTTP GET request to http://www.ncbi.nlm.nih.gov/pmc/utils/oa/oa.fcgi on 14 July 2015 revealed that the PMC-OAI subset contained 1075047 articles with the latest papers having been added earlier on the same day.

2.9.4. Proponents of open access

Open Access (OA) literature has been shown to have a greater research impact than literature based on reader-pays models (Antelman, 2004). According to the Australian Open Access Support Group (AOASG) (2015), there are eight properties of OA that makes it a viable publishing model: OA papers get higher citation rates, which leads to researchers gaining increased exposure (Harnad *et al.*, 2008; Mazloumian *et al*., 2011), access is made available to readers in developing countries (Chan *et al*., 2005), taxpayers get better value for their money (Harnad *et al.*, 2008, Phelps *et al*., 2012), research to practical application is better served (Willinsky, 2005), the public becomes better informed about scientific activities (Arzberger *et al.*, 2004, Willinsky, 2005), OA research is compliant with grant rules (Harnad *et al.*, 2008), and OA can influence public policy (Willinsky, 2005; Mazloumian *et al*., 2011).

2.9.5. Critics of open access

Concern is often raised about OA by suggesting that subscription services will be deprecated by free publications (Aronson, 2005), that OA will encourage or allow bypassing peer-review (Hunter, 2005; Haug, 2013), that it deprives authors of royalties (Goodman, 2004; Hunter, 2005), that it invites plagiarism (Goodman, 2004), and that it will become the refuge of second-rate or rejected material (Goodman, 2004; Hunter, 2005 , Haug, 2013) where conflicts of interested and copyright infractions are commonplace (Salem and Boumil, 2013). The argument that OA material in general is of inferior quality could be due to an unwillingness to recognize the distinction between OA categories and how peer-review processes are handled in each of these categories (Suber, 2009). In other words: the gold, green, other and rogue categories of open access material often seem to be erroneously equated with each other.

2.10. Ethical considerations relevant to biological data

2.10.1. Data privacy, ownership and consent

Even though abuse of personal records created during medical research on health and disease has not been documented (Gulcher *et al*., 2000) and is speculated to occur at most very rarely, it is still necessary to consider the impact of information management on privacy due to the legal implications that accompany violation of a person's right to privacy. This is especially relevant in the field of human genetics where governments, insurers or employers could theoretically discriminate against an individual based on that individual's genetic predisposition (Rindfleisch, 1997). To reduce the possibility of research data being used unscrupulously, various oversight committees have enacted best practices guidelines. The Icelandic data protection commission has for example outsourced encryption of all data gathered during disease based gene discovery to a third party (Gulcher *et al*., 2000). Their system ensures that data generated in the laboratory is de-identified before publication. Research have also been conducted into minimizing information loss during the process of de-identification and has led to proposed solutions to the k-anonymity and l-diversity computational problems encountered during the process of automated de-identification (Ghinita *et al*., 2007). In addition, watermarking algorithms traditionally used for copyright protection of media files has been integrated with binning algorithms to extend copyright protection to medical records (Bertino *et al*., 2005). Increasing demand for secondary use of medical data necessitates the need for protecting data owners while simultaneously still making it possible to determine the provenance of medical data. This complexity is compounded by the infeasible of obtaining specific consent for secondary use of medical data as highlighted by O'Neill (2003) in his examination on the limitations of informed consent.

2.10.2. Access to information and equitable treatment

Promoting public access to data is necessary because “...people have a critical stake in how experimental results affect their health, personal economy, and quality of life...” (McInerney *et al*., 2004). In the United States, patients have a right to review their medical records and this has led to the creation of online portals such as the Patient Clinical Information System (PatCIS) that makes it possible for patients to view their medical information by using the internet (Cimino *et al*., 2002). The creators of the PatCIS system reported that no adverse effects were observed by extending the right of patients to view their medical information from hard copy access to access through the

world wide web. Giving people the opportunity to access their medical records online could solve an observed inequality (Aday and Andersen, 1974) in access to healthcare between urban and rural individuals. Moreover, it would be easier to quantify the behavior of individuals who review their records online and that in turn would show support for the access-quantification-concept put forward by Donabedian (1972) that states that use of a service rather than the mere existence of a service is proof of access.

2.11. Conclusion

The motivation for this study arose from a general observation, noted as far back as the 1980's by some authors, that management of biological data rather than generation thereof has become the central area of concern for biologists and indeed for researchers and practitioners in other fields. Within the regional context that this project will be carried out, a suggestion was made that the expected increase in raw sequencing data from South African scientific activities, and specifically from the South African Human Genome Programme, should be supported by investigation into how subsequent data will be made available to the scientific community. The literature confirmed that SNPs are by far the most abundant DNA marker and that SNP data is often represented within VCF files, the latter being standardized by the well known 1000 Genomes Project as the medium of choice for representing biological variation extracted from raw sequencing data. SNP rs identifiers in the ID column of VCF files were recognized as the key terms of reference for SNP based variation, and as the probable query term when a scientist might conduct further research based on SNPs.

Having determined a specific focus for the project on SNP data within VCF files, further analysis of the literature revealed that there is a need to increase research efficiency when using unstructured data. An opportunity was recognized in the increased availability of open access literature that could serve as a source of semi-structured data that could in turn be integrated with query types that are initiated using rs identifiers. An argument was made that for an integrated service to be effectively deployed, it has to be aimed at complying with integrative technologies such as *Web 2.0,* semantic web technology, and linked data principals. The overarching influence that ontology has on each of these technologies was demonstrated by contrasting ontologies with database schemata and pointing out similarities with linked data principals. Moreover there was an overwhelming consensus in the literature that RESTful based web services architecture should be applied when designing *mashup* artifacts with an emphasize on semantic data integration.

The novel benefit that this project seeks to bring to the scientific community is to make it possible for researchers to (i) obtain the most relevant literature based on the frequency that a rs identifier has occurred with in a given article, (ii) make this facility available for bulk querying via a graphical user interface in addition to an application programming interface, and (iii) provide an upload facility for automated extraction of rs identifiers from of a VCF file. This combination of services has not been observed elsewhere in the extent of the literature covered during the review.

Examination of the current state of the literature has confirmed that there exists a gap in post variant call research that could feasibly be addressed by implementing a software system that integrates structured data with open access literature using free and open source software.

2.12. References

Abate, P., DiCosmo, R., Treinen, R., & Zacchiroli, S. (2011, June). MPM: a modular package manager. In *Proceedings of the 14th international ACM Sigsoft symposium on Component based software engineering* (pp. 179-188). ACM.

Aday, L. A., & Andersen, R. (1974). A framework for the study of access to medical care. *Health services research*, *9*(3), 208.

Alkoshman, M. M. (2015). Unified Modeling Language and Enhanced Entity Relationship: An Empirical Study. *International Journal of Database Theory and Application*, *8*(3), 215-227.

Antelman, K. (2004). Do open-access articles have a greater research impact?. *College & research libraries*, *65*(5), 372-382.

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, R., & Roberge, G. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels 1996–2013. *Science-Metrix-European Commission*.

Aronson, J. K. (2005). Commentary: Open access publishing: too much oxygen?. BMJ, 330(7494), 759.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., ... & Wouters, P. (2004). Promoting access to public research data for scientific, economic, and social development. Data Science Journal, 3, 135-152.

Attwood, T. K., Gisel, A., Bongcam-Rudloff, E., & Eriksson, N. E. (2011). *Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective*. INTECH Open Access Publisher.

Australian Open Access Support Group (AOASG). (2015). Welcome to the AOASG. [Web log post] Retrieved from http://aoasg.org.au/

Bertino, E., Ooi, B. C., Yang, Y., & Deng, R. H. (2005, April). Privacy and ownership preserving of

outsourced medical data. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 521-532). IEEE.

Bevacqua, N. (2015). *JavaScript Application Design*. Manning Publishing

Blumberg, R., & Atre, S. (2003). The problem with unstructured data. *DM REVIEW*, *13*(42-49), 62.

Bolser, D. M., Chibon, P. Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V. D., ... & Bhak, J. (2012). MetaBase—the wiki-database of biological databases. *Nucleic acids research*, *40*(D1), D1250-D1254.

Brickley, D., & Miller, L. (2012). FOAF vocabulary specification 0.98. *Namespace document*, *9*.

Brown, T.A. (2007). *Genomes 3.* New York, NY : Garland Science.

Buckley, J., Mens, T., Zenger, M., Rashid, A., & Kniesel, G. (2005). Towards a taxonomy of software change. *Journal of Software Maintenance and Evolution: Research and Practice*, *17*(5), 309-332.

Budapest Open Access Initiative (BOAI). (2002). Budapest open access initiative. Retrieved from http://www.budapestopenaccessinitiative.org/read

Bush, V. (1945). As we may think. *Atlantic Monthly*, 101-108.

Cannataro, M., Guzzi, P. H., Tradigo, G., & Veltri, P. (2014). Biological Databases. Springer Handbook of Bio-/Neuroinformatics, 431-440.

Cappiello, C., Daniel, F., Matera, M., Picozzi, M., & Weiss, M. (2011). Enabling end user development through mashups: requirements, abstractions and innovation toolkits. In *End-User Development* (pp. 9-24). Springer Berlin Heidelberg.

Chan, L., Kirsop, B., & Arunachalam, S. (2005). Open access archiving: the fast track to building research capacity in developing countries.

Cimino, J. J., Patel, V. L., & Kushniruk, A. W. (2002). The patient clinical information system (PatCIS): technical solutions for and experience with giving patients access to their electronic medical records. *International journal of medical informatics*, *68*(1), 113-127.

Cisco. (2014). Big Data: Not Just Big, But Different - Part 2 [Web log post] Retrieved from http://www.cisco.com/web/about/ciscoitatwork/enterprise-networks/docs/i-bd-04212014-not-just-big-different.pdf

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, *38*(6), 1767-1771.

Cui, B., Mei, H., & Ooi, B. C. (2014). Big data: the driver for innovation in databases. National Science Review, 1(1), 27-30.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158.

Dearle, A. (2007). Software deployment, past, present and future. *2007 Future of Software Engineering*, 269-284.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., ... & Finney, A. (2010). The BioPAX community standard for pathway data sharing. Nature biotechnology, 28(9), 935-942.

Di Battista, G., Didimo, W., Patrignani, M., & Pizzonia, M. (2002). Drawing database schemas. *Software: Practice and Experience*, *32*(11), 1065-1098.

Donabedian, A. (1972). Models for organizing the delivery of personal health services and criteria for evaluating them. *The Milbank Memorial Fund Quarterly*, 103-154.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal*,

22(2), 338-342.

Gansner, E. R., & North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, *30*(11), 1203-1233.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, *295*(1), 90-93.

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007, September). Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases* (pp. 758-769). VLDB Endowment.

Gingeras, T. R., & Roberts, R. J. (1980). Steps toward computer analysis of nucleotide sequences. *Science*, *209*(4463), 1322-1328.

Goble, C., & Stevens, R. (2008). State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, *41*(5), 687-693.

Goodman, D. (2004). The criteria for open access. *Serials review*, *30*(4), 258-270.

Gulcher, J. R., Kristjansson, K., Gudbjartsson, H., & Stefansson, K. (2000). Protection of privacy by third-party encryption in genetic research in Iceland. *European journal of human genetics: EJHG*, *8*(10), 739-742.

Grote, C., Segreto, E., Okerlund, J., Kincaid, R., & Shaer, O. (2015, January). Eugenie: Multi-Touch and Tangible Interaction for Bio-Design. In Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (pp. 217-224). ACM.

Handke, J. (2015). Micro-Lectures - Semantics. Retrieved from https://www.youtube.com/playlist?list=PLRIMXVU7SGRJnARkQhb6LO8KOi5rQf2Qn

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., ... & Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. Serials review, 34(1), 36-40.

Haug, C. (2013). The downside of open-access publishing. New England Journal of Medicine, 368(9), 791-793.

Hunter, K. (2005). Critical issues in the development of STM journal publishing. Learned publishing, 18(1), 51-55.

IBM. (2012). What is big data? Retrieved from http://www-01.ibm.com on May 3, 2015.

Ives, B., & Olson, M. H. (1984). User involvement and MIS success: a review of research. Management science, 30(5), 586-603.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D., & Vardeman, C. (2014). Five stars of Linked Data vocabulary use. *Semantic Web*, *5*(3), 173-176.

Kamphans, T., & Krawitz, P. M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics, 28(19), 2515-2516.

Kesh, S. (1995). Evaluating the quality of entity relationship models. *Information and Software Technology*, *37*(12), 681-689.

Leavitt, N. (2010). Will NoSQL databases live up to their promise?. *Computer*, *43*(2), 12-14.

LinkedDataTools.com (2015). Introducing Linked Data And The Semantic Web [Web log post] Retrieved from http://www.linkeddatatools.com/semantic-web-basics

Machado, C. M., Rebholz-Schuhmann, D., Freitas, A. T., & Couto, F. M. (2015). The semantic web in translational medicine: current applications and future directions. *Briefings in bioinformatics*, *16*(1), 89-103.

Manifesto for Agile Software Development. 2001. [Web log post] Retrieved from http://www.agilemanifesto.org/

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., ... & Sycara, K. (2004). OWL-S: Semantic markup for web services. *W3C member submission*, *22*, 2007-04.

Martone, M. (2012). Where do we go from here? (databases and ontologies). Retrieved from https://www.youtube.com/watch?v=fnN9fYeKiSo

Mazloumian, A., Eom, Y. H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. PloS one, 6(5), e18975.

McInerney, C., Bird, N., & Nucci, M. (2004). The flow of scientific knowledge from lab to the lay public the case of genetically modified food. Science Communication, 26(1), 44-74.

Meng, J., Mei, S., & Yan, Z. (2009, December). Restful web services: A solution for distributed data integration. In *Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on* (pp. 1-4). IEEE.

O'Neill, O. (2003). Some limits of informed consent. *Journal of Medical Ethics*, *29*(1), 4-7.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1), 17.

Ortega, J. L., & Aguillo, I. F. (2014). Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, *65*(6), 1149-1156.

Pautasso, C. (2014). RESTful web services: principles, patterns, emerging technologies. In *Web Services Foundations* (pp. 31-51). Springer New York.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830.

Peña, I., Cabezas, C., & Alonso, J. L. (2015). The Nucleoside Uridine Isolated in the Gas Phase. *Angewandte Chemie*, *127*(10), 3034-3037.

Phelps, L., Fox, B. A., & Marincola, F. M. (2012). Supporting the advancement of science: Open

access publishing and the role of mandates. Journal of translational medicine, 10(1), 13.

Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, *9*(1), 69-82.

Poli, R., Healy, M., & Kameas, A. (Eds.). (2010). *Theory and Applications of Ontology: Computer Applications* (pp. 1-26). Dordrecht: Springer.

Ponzanni, G. (2013). Introduction to NoSQL. [PDF slides]. Retrieved from http://profs.sci.univr.it/~pozzani/Materiale/nosql/01 - introduction.pdf

Price, W. N. (2015). Describing Black-Box Medicine. *Boston University journal of Science and Technology Law, Forthcoming*.

Raemaekers, S., Van Deursen, A., & Visser, J. (2014, September). Semantic versioning versus breaking changes: A study of the maven repository. In *Source Code Analysis and Manipulation (SCAM), 2014 IEEE 14th International Working Conference on* (pp. 215-224). IEEE.

Reeves, S., Kuper, A., & Hodges, B. D. (2008). Qualitative research methodologies: ethnography. *Bmj*, *337*.

Rindfleisch, T. C. (1997). Privacy, information technology, and health care. *Communications of the ACM*, *40*(8), 92-100.

Robbins, R. J. (1996). Bioinformatics: Essential Infrastructure for Global Biology1. *Journal of Computational Biology*, *3*(3), 465-478.

Ross, K. A., Janevski, A., & Stoyanovich, J. (2005, August). A faceted query engine applied to archaeology. In *Proceedings of the 31st international conference on Very large data bases* (pp. 1334-1337). VLDB Endowment.

Salem, D. N., & Boumil, M. M. (2013). Conflict of interest in open-access publishing. New England Journal of Medicine, 369(5), 491-491.

Scarano, D & Rao, R. (2014). DNA markers for food products authentication. Diversity, 6(3), 579–596.

Schulz, S., & Jansen, L. (2013). Formal ontologies in biomedical knowledge representation. *Yearb Med Inform*, *8*(1), 132-46.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords?. *BMC bioinformatics*, *4*(1), 20.

Sheridan, J., & Tennison, J. (2010). Linking UK Government Data. *LDOW*.

Smith, B. (2008, July). Ontology (Science). In *FOIS* (pp. 21-35).

Soldatova, L. N., & King, R. D. (2005). Are the current ontologies in biology good ontologies?. *Nature biotechnology*, *23*(9), 1095-1098.

Staab, S., & Studer, R. (Eds.). (2013). *Handbook on ontologies*. Springer Science & Business Media.

Stein, L. D. (2003). Integrating biological databases. Nature Reviews Genetics, 4(5), 337-345.

Stevens, R., Goble, C. A., & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. Briefings in bioinformatics, 1(4), 398-414.

Stonebraker, M. (2010a). SQL databases v. NoSQL databases. *Communications of the ACM*, *53*(4), 10-11.

Stonebraker, M. (2010b). Errors in database systems, eventual consistency, and the cap theorem. *Communications of the ACM, BLOG@ ACM*.

Suber, P. (2004). *Open access overview*. Retrieved from http://www.planta.cn/forum/files_planta/what_is_open_accessan_overview_2004_162.pdf

Suber, P. (2009). *A field guide to misunderstandings about open access*. Retrieved from http://legacy.earlham.edu/~peters/fos/newsletter/04-02-09.htm

Szeredi, P., Lukácsy, G., & Benkő, T. (2014). *The Semantic Web Explained: The Technology and Mathematics Behind Web 3.0*. Cambridge University Press.

The World Wide Web Consortium. (2015). How the Semantic Web Works [Web log post] Retrieved from http://www.w3.org/2002/03/semweb/

Tsarkov, D., & Horrocks, I. (2006). FaCT++ description logic reasoner: System description. *Automated reasoning*, 292-297.

Tweed, R., & James, G. (2010). A universal nosql engine, using a tried and tested technology. *White Paper, Creative Commons Attribution CC-BY*.

US v. Swartz, 945 F. Supp. 2d 216 (D. Mass. 2013).

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Beasley, E. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *portal: Libraries and the Academy*, *9*(1), 5-24.

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). *Dublin core metadata for resource discovery* (No. RFC 2413).

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. First Monday, 10(8).

Wirth, N. (1986). *Algorithms and data structures*. Upper Saddle River, New Jersey: Prentic Hall.

Yu, Q., Ding, Y., Song, M., Song, S., Liu, J., & Zhang, B. (2015). Tracing database usage: Detecting main paths in database link networks. *Journal of Informetrics*, *9*(1), 1-15.

Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological Databases for Human Research. Genomics, proteomics & bioinformatics, 13(1), 55-63.