DATASET VERSIONING THROUGH CHANGELOG ANNOTATION

By

Benno Lee

Prepared for:

Peter Fox, Thesis Advisor

Jim Hendler, Advisor

Deborah MacGuiness, Member

Beth Plale, Member

Rensselaer Polytechnic Institute
Troy, New York

November 2016
(For Graduation December 2017)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

ABSTRACT

The growth of data driven technologies has "exploded" the need for reliable, high quality data. This data powers the science of major agencies like NASA and laboratories like CERN. The changes made to maintain its quality has serious implications to the quality of the data that system uses as they propagate down stream through a workflow. Provenance plays a large part in identifying factors that contribute to data change. Technologies like PROV and OPM have allowed researchers to instantiate and track the entities and activities which went into generating their data sets. However, these ontologies do not cover change information in detail, and arguably, this falls outside their purview. Provenance discloses the contributors to the generation of a data object, but versioning describes the relationship that object has with its previous or future iterations. Changes to an object's provenance undoubtedly creates a new version of that data, but measuring the magnitude of that change still remains difficult. Current methods to quantitatively determine the distance between versions of a data object often involve comparing provenance graphs, but these approaches lacks the detail to make meaningful comparisons. There is a gap in understanding the transition from an old data set to a new one, and developing a more detailed understanding of change information allows users to comprehend how a data set evolves over time.

Much work has already been accomplished towards filling in this gap and laying the foundations to further address issues in this area. The first part of this thesis presents a concept model developed as a foundation to capture changes across versions. It does so by capturing the relationships involved in the three core versioning operations: addition, invalidation, and modification. Two data sets stored in Excel spreadsheets were used to develop and test the versioning model with the goal of describing changes in a scalable way that can expand to other applications. In open source software, changelogs provide more detailed change information to users and developers as a documentation artifact. However, changelogs traditionally present its contents only as human readable formats text. This work adopts changelogs as

a documentation tool to communicate change, but also adopt semantic web technologies by encoding machine readable content into the log using RDFa. In such a way, the changelog can be provided as an openly available HTML page, required to use RDFa, which is also machine accessible.

Work needs to continue to completely demonstrate the utility of the versioning model and its applicability. Much of the model construction and distribution has already been demonstrated with spreadsheets, but to ensure flexibility and applicability, other contexts must be hooked into the model. Databases provide a unique challenge because transactions do not create new instances of the data object. The versioning of ontologies provides significant insight into the growth of vocabularies to a domain as well as expand the application of older data sets. GCMD keywords do not constitute an ontology, but changes to terms within the hierarchical vocabulary have significant impact on the searchability and discoverability of Earth Science data sets. Applying the process of versioning the vocabulary could provide greater freedom to evolve without sacrificing the ability to find data sets. As a final endeavor, a method needs to be developed to utilize the structure of the resulting versioning graph in performing a flux-like calculation and give a quantifiable distance measure for the amount of change between versions of data.

CHAPTER 1
INTRODUCTION

John C. Maxwell once said, "Change is inevitable. Growth is optional." While this inspirational quote refers to the human character, it also holds true for scientific datasets. With changing technology, data collected by researchers grew at an astounding rate. NASA's Atmospheric Science Data Center (ASDC) reported a growth from hosting around five million files to twenty million files between 2001 and 2004 as seen in Figure 1.1 [2]. The ATLAS project at CERN reports that it generates on the order of four thousand new datasets per day from experimental tests alone shown in Table 1.1 [1]. The sheer volume of data generated per year by each of these organizations easily demonstrates the futility of managing these data archives manually. The desire and ease in which data transparency can be provided to not only researchers but also the public lies behind the drive of expanding the availability of high quality data holdings. The key to meeting these demands is automation, not only in distribution, but also in data quality management. However, these two dynamics are at odds with each other. Many NASA datasets have required re-processing of their data, either to improve data quality or to correct for errors [12]. However, we can also see that the number of distinct users doubled over the course of three years at the ASDC while the amount of data distributed more than triples. As such, the strain of informing and providing updated data to this body of users grows tremendously. The solution, thus, lies in passing on the ability to verify data quality to the users. Data traceability now becomes particularly important to identify sources that contribute to improved data quality. It creates a need to understand not only that a data set has changed, but to also understand how much a data set has changed. Data versioning is the method of tracking the changes performed on a data set and determining the extent to which it has changed. In this document, data versioning is approached using technology provided by semantic technologies and applying them to artifacts currently generated by scientific data sets.

Metric	2001	2004
Data Volume of Holdings [TB]	340	1,250
Number of Files	5 M	20 M
Data Volume Distributed [TB]	34	114
No. of Distinct Users	6,000	12,000
Production Jobs Run per Day	2,000	5,000

Figure 1.1: Summary of ASDC Holdings for the years 2001 to 2004 from Barkstrom and Bates [2]

Metric	Magnitude
Data per day [TB]	1
Data per year [PB]	20
Datasets per acquisition	O(4000)
Simulation Datasets per day	O(1500)
Versions per Dataset	O(1)
Files per Dataset	O(100)

Table 1.1: Summary of ATLAS data set generation in 2008 from Branco et al.[1]

1.1 Data Set Proliferation

An interesting statistic to note in Table 1.1 is the order of versions per dataset. While it does not indicate that the data have marked volatility, it does communicate that the data has a tendency to change over time. This means that not only does an archive maintain a data set, but it may need to maintain multiple instances of that work over time. Therefore scientists need the ability to discern between instances of the data they use when comparing results from other researchers. With the web's development, libraries and library sciences provide a steady evolution in methods to identify data collections. The challenges and goals that face physical libraries remain valid even as data collection migrates to electronic alternatives [13]. Digital storage and the Internet has opened new opportunities and methods to administer book data by separating logical representations and physical representations [2]. Publications, for instance, may be sorted and rearranged in a digital view along a

wide array of characteristics to provide the most logical presentation for searching. Comparatively, users search a physical collection by conforming to a rigid organizational system adapted to the provider's needs, not the user. It has also added a plethora of new content types such as wikis, blogs, and other document formats which have never seen physical print. All these new documents need a form of data management [14]. However, the migration has not been without its problems. Early citations used stagnant Uniform Resource Locators (URL) to refer to online documents, but this would lead to a condition known as link rot where moving the document would invalidate the URL [15]. This eventually led to the development of Persistent URLs (PURL) which also succumbed to link rot, and this eventually led to the distributed Digital Object Identifier (DOI) system used to track documents today [3]. In the table taken from Duerr et al. [3], DOIs represent the most suitable identifier used for science. It's origins begin in the Handle system, which can be seen as a generalization of the DOI system. The DOI network provides a robust system to track documents, but when tracking data, it faces difficulty following the rate of change with some more volatile data sets. Distribution organizations assign a DOI whenever a new edition of a document becomes available, and due to the publication process, documents change very rarely so a new DOIs are rarely necessary. However, data sets are products and thus succumb to the iterative process of error correction and growth. Data collection often continues on after initial publication. DOI distributors treat new files like new sections to a paper and changes to files as edits so a new identifier must be issued to the data set. This behavior becomes entirely too slow as data providers begin to allow users to dynamically generate data products from existing data according to their needs [16].

With respect to Figure 1.2, no identification scheme fits the description of a scientific identifier. Duerr et al. define a use case to make the argument that scientifically unique identifiers are necessary, "to be able to tell that two data instances contain the same information even if the formats are different" [3]. This highlights the necessity of being able to discern between the logical content of a dataset and the physical form of that object. As identifiers often use physical characteristics of the data to make comparisons and determine similarity, a proper representation

Table 2 Suitable identifiers for each use case where solid green indicates high suitability, vertical yellow stripes indicates good to fair suitability; and orange diagonal stripes indicates low suitability

Identifier Type	Unique Identifier		Unique Locator		Citable Locator		Scientifically Unique Identifier	
	Dataset	Item	Dataset	Item	Dataset	Item	Dataset	Item
ARK								
DOI								
XRI								
Handle								
LSID								
OID								
PURL								
URL/URN/ URI								
UUID								

Figure 1.2: Table of predominant identifiers used in science. From Duerr et al. [3]

becomes difficult. However, a model recently released model by the Health Care and Life Sciences (HCLS) Interest Group may provide a solution when used in conjunction with other identifiers [4]. Their model, shown in Figure 1.3, separates the concept of a dataset into three parts. The highest level summarizes the data as an abstract work, perhaps better described as a topic or title. This data topic can have multiple versions as it changes over time. The version can then be instantiated into various distributions with different formats. Such a set of linked data would provide sufficient data to discern between two data instances of the same information. However, it is unclear as to whether a single identifier would be able to encompass such complete information.

The organization of the model has interesting parallels with Plato's theory of Forms where he proposes that there exists a perfect idea of an object which has imperfect realizations in the physical world. Likewise, the summary description of the data set represents the ideal which the data seeks to capture. The model then uses a series of versions represented using the Provenance, Authoring, and Versioning (PAV) ontology to document the physical change of the data as it approaches the ideal[17]. PAV produces an interesting entry point into explaining the data set's development as well as recognizing the imperfection inherent in data capture. As

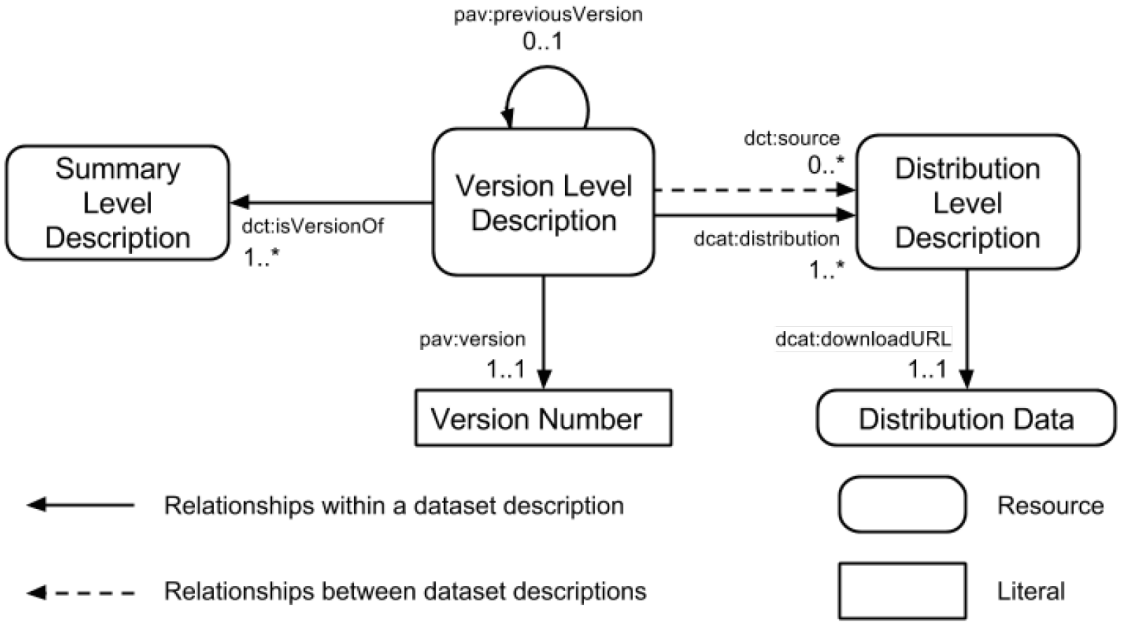

Figure 1.3: Data model from the Health Care and Life Sciences Interest Group separating data into three levels: works, versions, and instances. From Dummontier et al. [4]

a concept in the model, linked data can further extend the description to provide details necessary in identifying a particular version of data. However, PAV only has retrospective views of data versions since a version can only point back to previous instances.

For similar reasons, treating data as documents produces problems when applying technologies from software management [11][18]. Structure provides the most significant distinguisher between data and software since a data set with a removed file remains usable but a software project would break. The function of code comes from its content, but the function of data comes from its ability to store and organize data. This should not be confused with data formats which impose structure onto data in much the same way programming languages provides a medium to express actions. However, exporting data in different formats is currently easier than exporting code into different languages. Data sets do not represent a single object, unlike a software project[5]. They are compact representations of all possible subsets of the data set, which are also datasets. The Atmospheric Radiation Measurement (ARM) Program publishes data daily from sensors deployed across the globe, but a

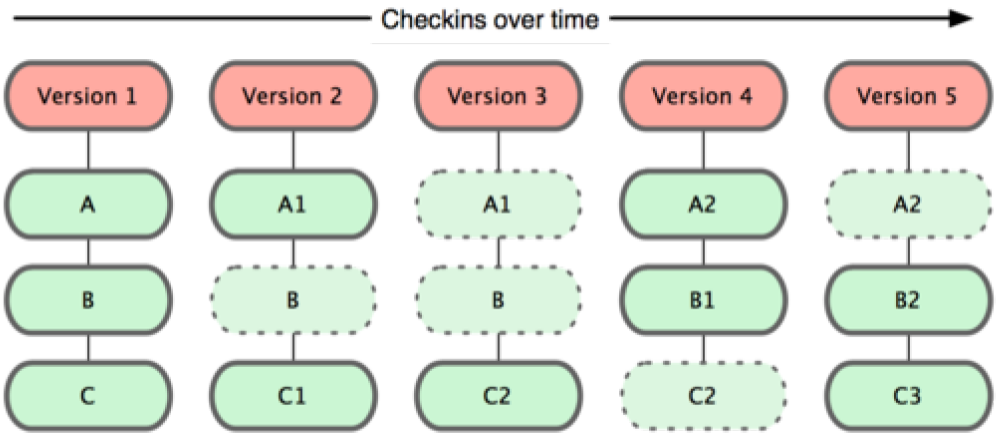

Figure 1.4: GIT stores changes in the repository as snapshots of individual files. Figure 1.5 from [5]

user may only desire files from a specific region, files from only the year 2012 in the Southern Great Plains region, or from only the month of February for the collection site's lifespan [19]. These can all be considered arbitrary subsets from the data sets generated by the data collection program, but a software project would not be able to sustain such arbitrary filtering. GIT stores the files comprising a repository as snapshots of each file in the store, and when developers make changes to a file, the store creates a new snapshot of that files as seen in Figure 1.4. A version then becomes defined by the complete collection of snapshots within the repository at the time, and the files cannot be subsetted due to the dependencies between the files in the version. As a result, when modeling the commit history, or the history of versions, the set of files can be compacted since all components are necessary to produce the object. Demonstrated in Figure 1.5, a developer creates a branch of the master line with commit C3 from C2. The entire repository at C2 gets copied into the branch, and when the work is done, it gets merged back into the master branch. When a user then orders the software generated by this repository, it results from the compilation of the entire commit whereas a data set does not require the same completeness to be used. For this reason, the structures of data sets and software becomes incompatible and software versioning technologies are insufficient to capture this nuance.

Furthermore, returning to Figure 1.4, GIT actually can be considered a re-

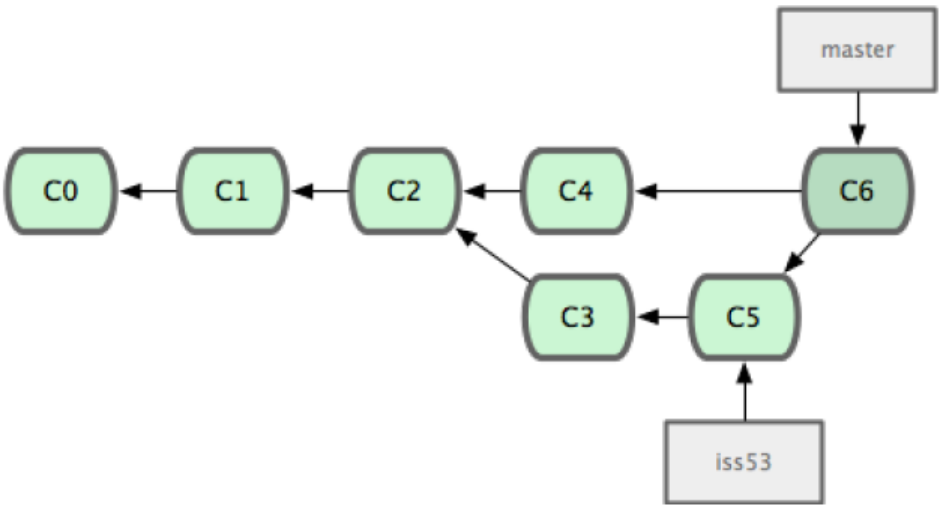

Figure 1.5: Example of a commit history with branching stored in GIT. Figure 3.17 from [5]

duction in terms of software versioning methodology. The constant dynamic in play with tracking versions, perhaps with Computer Science in general, is the trade off between space and time. Maintaining full snapshots of repositories proves extremely space consuming, but retrieving these snapshots occurs near instantaneously. In general, prior to this, the conventional wisdom was to instead store one snapshot and a set of deltas or differences between software documents and then use processing time to reconstruct a particular commit by applying the differences [11]. Deltas can further break down into forward or backward changes, reflecting the importance of keeping the snapshot of the oldest or newest instance. This produced very lean versioning systems, but after a long string of changes, applying deltas becomes very time consuming so periodical snapshots become necessary. With costs for storage space compared to the relative size of software files dropping, GIT prioritizes the ability to quickly swap between and develop different branches of the system. This interplay between space and computation time can also be seen within the data management space as keeping snapshots of large databases will be intensely costly but less so for spreadsheets.

The techniques employed by these technologies, however, can remain applicable to data sets and are often necessary when communicating change data to users. Version producers often refer to versions using numbers in the dot-decimal style [20].

While the values often signify the Major-minor numbers associated with the version, the names remain meaningless and can arbitrary assignment such as Ubuntu released numbered by Year-month values [21]. The arbitrary nature of the numbers often entails referring to versions by English nicknames instead. Such a regular method of naming release versions also means that determining the magnitude of change between two releases becomes impossible. Numbering the version this way, however, does allow computers and readers to quickly parse the version name and discern that a change has occurred, but little value exists beyond that [22]. The technique of distributed and federated employed by GIT does provide significant value to modern methods of versioning data [23]. As data workflows and data set dependencies grow, their volatility also expands, meaning that they become more likely to generate new versions. The federated approach available in the GIT environment allows developers to establish change dams that collect modifications and releasing the data at regular intervals,reducing the changes to a manageable flow.

1.1.1 Unified Systems

Working with data as documents leads to the shortfall of technologies, but working with the data of documents has led to significantly greater contributions. Many libraries often work in collaboration in order to provide a wide selection of texts over a limited number of physically available documents. The University of Virginia demonstrates the ability to achieve a unified library system using a combination of XML and web service technologies of their disparate assemblage of libraries [24]. The challenge involves providing a common landscape in order to compare the quality requirements imposed on the repositories. Versioning systems provide a notable mechanism to make this decision as quality determines when to generate new versions and what items belong to the same groupings of data. The comprehensiveness of XML and web technologies also allows this approach to apply to other systems and research areas as well. This becomes particularly relevant as innovations in computation technology generates small, volatile data sets to integrate into larger data managers [25]. In this application, the data food chain then becomes represented by smaller applications generated in situ and then

unified with other data sets as they move up the food chain to a large, unified data distribution center.

Unified libraries represent a part of a larger collection of systems that rely on the propagation of data through heterogeneous systems to produce rapid complex solutions. The grid provided a unique environment that had to handle a variety of inputs, and therefore, different input data could run on distinct sets of grid services. This meant that different versions of the same data could be generated by differing services on the same grid [26]. The CERN grid for the Compact Muon Solenoid experiment separates the physical and logical storage of files, allowing multiple users to refer to the same file without needing to copy the file across the grid [27]. While the structure and construction of the grid reduced the uncertainty introduced by varying hardware, it raised questions on data quality by abstracting the transparency to underlying services. Cloud services have recently replaced the grid due to its flexibility in the services available to its autonomous systems. As the scale and complexity of autonomous systems grow, it becomes more difficult for one system or organization to manage all the circles necessary to produce data deliverables. The ability to propagate relevant data change data across autonomous systems then assures valid quality in interactions between domains [28]. Not only does this ensure uniformity through system interaction, but it also ensures transparency with respect to the data and methods used to produce conclusions [29]. This often means that systems will need to negotiate a contract and establish a mutual interface to exchange data. Occasionally, this contract can be formal, but more ideally, the establishment of a standard lineage model or format would allow a greater variety of systems to interact with each other without needing lengthy contractual exchanges.

1.2 Defining Versions and Versioning

Researchers often use the words version and versioning without needing to explicitly defining them because they appear ubiquitously in data management systems. As a result, the definition of a version varies depending on the application, but they all try to capture a common idea. Barkstrom describes versions as homogeneous groupings used to control, "production volatility induced by changes in algorithms

and coefficients as result of validation and reprocessing," [6]. He separates groupings by the granularity at which observations are collected. This construction allows users to easily determine the similarity or dissimilarity of objects within the grouping. However, NASA employs a rigid data processing structure with well defined levels. This means that the resulting hierarchy is well balanced.

Another application defines a version as a "semantically meaningful snapshot of a design object," [29].

It is more constructive to define versions as a matter of purpose. When two objects are versions of each other, they are considered to still serve the same purpose. For data production services, this would translate to there they exist in the workflow. Two objects must, therefore, share provenance, and the more they share, the more likely versioning practices will provide meaningful results.

Versioning is the activity of exposing changes relating two or more versions of an object. Versioning differs from provenance because an activity does not establish an object as a version of another even though the activity may create it. The relationships exist as a matter of state-based properties which are exposed through versioning. These relationships have greater significance if the objects are versions of each other.

1.3 Data Quality/Provenance

The fundamental challenge to determining data quality, its subjectivity, needs clarification. Conceptually, a data set on desert climate likely has very poor data quality and relevancy to a study in whale biology. However, in a more quantitative sense, that same desert climate data can have excellent quality with respect to its correctness, expression, and traceability. With the hybridization of data sets from disparate agencies to provide big data solutions, collaborations plays an ever present role in achieving broad, valid findings. This requires good quality data and the ability to determine when data with better quality becomes available. [30]. The primary focus, generally, involves tracing the lineage of artifacts and activities that lead to the current data. This provides insight into possible sources of error as well as validating the assumptions made in generating a data set for future use.

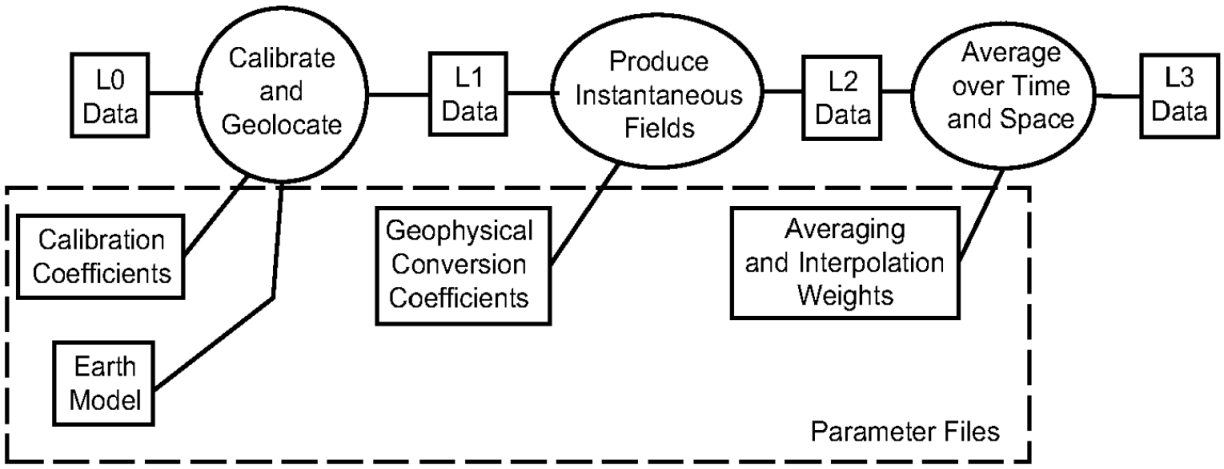

Figure 1.6: NASA organizes its data into three levels depending on the amount of aggregation and the distance the data is removed from the original sensor measurements. Figure 1 from [6]

There are several characteristics that can describe a data's quality, but the one most relevant to data versioning is provenance. It describes the sequence of events that lead to the construction of an object [31]. In art this describes the sequence of people who have had ownership of a piece of artwork. For data, provenance relates the history of inputs and operations that result in a data object such as a plot or data set. NASA defines three levels of data processing, seen in Figure 1.6, that encompass the stages required to turn a raw signal from satellite instruments into physical measurements into global aggregate summaries [6]. Each stage computes a dataset using a collection of input data, processing scripts, and calibration values. This collection forms the provenance for the resulting level of data, and it contributes to the production history of the higher level products it feeds into. More specifically, scripts and code which "Produce Instantaneous Fields" operates on "L1 Data" using "Geophysical Conversion Coefficients." The diagram clearly illustrates that should "Calibration Coefficients" used to generate "L1 Data" change, the resulting data would cause "Produce Instantaneous Fields" to yield different "L2 Data." As such, strong ties lie between identifying versions and locating differences between provenance changes.

As data sets grow, this process becomes even more confusing to coordinate so version control systems often manage provenance. Current research endeavors

to provide high quality data clearly becomes more formalized as data becomes concentrated in massive data warehouses [32]. The focus of a majority of versioning research focuses on lineage retrieval which becomes ever important as evidence grows that researchers generate data faster than they can reasonably track [33]. This poses a particularly difficult problem as provenance provides a potent means of data auditing. With provenance, data producers ensure the trustability of their data inputs, either ingested from external sources or integrated from internal data sets. Fairly reassuring results have been found when combining lineage management and error reporting systems [34]. The errors provide a context for the changes made to advance the lineage of the data set and the version manager demonstrates that a problem has been addressed and how it was corrected. This system becomes extremely important when considering that agency funding often depends on the ability to account for the value of a project's dataset [35]. The data analytics required to determine the value of data collected by a project also requires the provenance to ensure that the analysis is also reproducible. The basic provenance often collected by hand now needs to be collected automatically in order to facilitate collaboration, especially with projects that are farther away from the data.

Early attempts at encoding provenance data into semantic models include the development of the Proof Markup Language [36]. While this was originally developed to express inference reasoning through traceable graph relations, the model can also be used to express the provenance of products using the same transitions. The power is that it is able to use terms defined on the Semantic Web to construct inferences. This early demonstration of the ability for web based semantic technologies also expresses complex relations in a way that can be reasoned over and computed. It then allows for autonomous solutions to understanding change as data freshness begins playing a significant role in successful system function [37].

Not long after began the development of the Open Provenance Model (OPM) [38]. Driven by the uncertain needs and sometimes conflicting conventions of different scientific domains, the model sought to find a method to standardized the way in which provenance data is captured while also keeping the specification open to accommodate current data sets through the change. In an experimental case, the

model has been applied to sensor networks to automate and unify their provenance capture even as they grow [39]. To aid in the adoption of the OPM, the framework Karma2 was developed to assist integrating provenance capture into scientific workflows [40]. It reduces the amount of modifications required to adopt the OPM through web services and, more importantly, integrates into scientific workflows. With the magnitude of data collection endeavors, it is no longer feasible for scientists to stay close to the data and must take a more abstract view of their data collection activities. Scientific workflows provides this high level view of complex data collection, curation, and analysis [41]. The value then of integrating provenance capture and workflow design is that lineage planning can then take place at a high level of scientific work. This gives insight into how different parts of the workflow fit together and how new exploratory expansions may occur.

Following the OPM, PROV is a W3C recommendation that deliniates a method to express data provenance with semantic technologies that has been accepted as a World Wide Web Consortium (W3C) Recommendation [42] [43] [44]. The recommendation uses a conceptual model relating activities, agents, and entities together to describe data production lineage [45] [46] [47]. Intended as a high level abstraction, it describes data as entities that are generated by activities enacted by agents. This basic relationship is very powerful in its ability to describe data production activities. The expression of the conceptual model occurs through the PROV Ontology (PROV-O), which can be conveyed through various resource description languages [9] [48] [49]. The ontology is further formalized into a functional notation for easier human consumption [50] [51]. One particular strength that has contributed to the adoption of PROV is its ability to link into other ontlogies, making it easier for existing semantically enriched data sets to adopt PROV [52] [53]. Like the OPM, a framework has also been developed to alleviate workflow integration through Komadu [54]. The framework improves over its predecessor, Karma, by no longer utilizing global context identifiers that were no necessarily shared throughout the workflow.

The PROV Ontology provides four different concepts that begin to encapsulate the provenance relationship between data versions. The ontology defines a

prov:Generation as "the completion of production of a new entity by an activity," [9]. This means that the generation must result from a prov:Activity. In versioning, activities play a much less active role since changes become exposed from comparing like objects. It creates a relationship between an entity and an activity, but such a relationship may lead to an implication that a change in the activity resulted in changes in the resulting version. Changes could also result from modifications in the input data, leading to an entirely new generating activity rather than a modified one. Prov:Invalidation likewise makes a similar connection between activities and entities. This means that PROV-O does not have the direct means to communicate the addition and invalidation relationships which exist in a versioning context. Since versioning relationships result from state-based effects between version entities, a more contextually appropriate relationship would connect two entities together. The property prov:Derivation does relate together two entities and the ontology defines it as, "a transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity. " [9]. In the case of the MBVL dataset, none of these three assertions hold true. The four versions are being simultaneously considered so one is not being transformed into another as would a sequential set of versions. Additionally, since we do not know which version is the best, we cannot consider any version an update of the others. Finally, no entity pre-existed as the data sets resulted from an ongoing analysis and further steps have not been developed.

PROV has played a significant contribution in maintaining the quality and reproducibility of datasets and reporting in the National Climate Assessment (NCA) [55]. This implication signifies that there is an increased likelihood of adoption through other scientific fields as a result of this reporting. The Global Change Information System, which houses the data used to generate the NCA, uses PROV to meticulously track the generation of its artifacts and results as they are used in the report [56]. This means that not only does the data have a traceable lineage to verify quality, but the content of documents can have the same verifiability [57].

1.3.1 Changelogs

Changelogs, sometimes called patch notes, are artifacts resulting from the versioning process often found in major software projects. They document the changes made within the system and seek to explain, in human language, the motivations behind changes [58]. The logs provide significant utility to both users and producers as it can serve as both documentation and tutorials. Many users will often refer to the patch notes in order to decide how to adapt to changes made to the system they use, either data or software. Meanwhile, changelogs aid producers through team transitions by keeping a history of decisions made to improve the project. This is particularly evident in the realm of open-source projects as developers can contribute without having been part of the original development team. The need for documentation to bring new programmers up to speed for a project drives the ability to keep the project alive.

Open source projects have much more consistent adoption of changelogs than data sets, possibly resulting from complex code techniques emerging earlier than large data methods. These logs provide a great source of value to developers as they can be used to give insight to the health of a software project [59]. These projects have a tendency to die rather quickly after initial enthusiasm and with the rather low overhead cost to start new open-source projects, some automated methods of determining the progress of a project is needed. It would give insight into the maturity of a project's development team as well as the likelihood that team members will correct errors within the code. However, readability proves to remain a significant hurdle as current development change logs contain solely human readable text. While machines may still be significantly removed from the ability to comprehend the impact of changes made to a data set or software code, they are currently opaquely blocked from consuming any of the content within logs more than understanding they contain text. The transition between different versions of large datasets is then left largely up to the human user's ability to understand and process the modifications mentioned within the change log.

As mentioned previously, changelogs also allow developers to link bugs and errors with their corrections in new versions of the code [60]. This gives feedback to

the user community that corrections have been addressed as well as ensuring that modifications to the code base are driven by improving the project. It also has the added benefit of creating a system that can be used to link the introduction of new features with the emergence of new bugs [61]. The resulting discoveries help reveal patterns of development and prevent further occurrences of problematic code. Therefore, providing an machine consumable changelog would accelerate and assist in navigating through dataset changes and error corrections.

1.3.2 RDFa

In order address the human readability of data change logs, this project considers the use of the Resource Description Framework in Attributes (RDFa) framework [62]. Figure 1.7 illustrates the semantic difference between what web crawlers and what humans see when they consume web pages. People intuitively understand that certain strings represent meaningful information, and RDFa seeks to encode that understanding natively into the document in order to allow them the ability to consume the information more effectively. The framework leverages the existence of various established vocabularies in order to provide a standardized understanding of web documents, and it opens the ability for them to interact with the web pages more intelligently. The benefits of embedding RDFa into change logs is twofold. First, the change log would need to be marked up in HTML in order to accept RDFa. As a result, the log would also become available on-line and thus, more openly accessible to data users through the utilization of web based search engines. This would allow data users to better determine personally how a change applies to their specific application. Large companies such as Google have already begun making endeavors in equipping their web crawlers to consume structured data such as RDFa from web pages. Second, the simple application of RDFa attributes encodes the entries within the change log in a format consumable by machines [7]. RDFa has already had significant success in adoption across a variety of web publication platform and eases the search for their content [63]. In these applications, however, the developers use RDFa to describe the content on the page, to indicate a string is actually a name for example. This project endeavors to use RDFa to embed an RDF

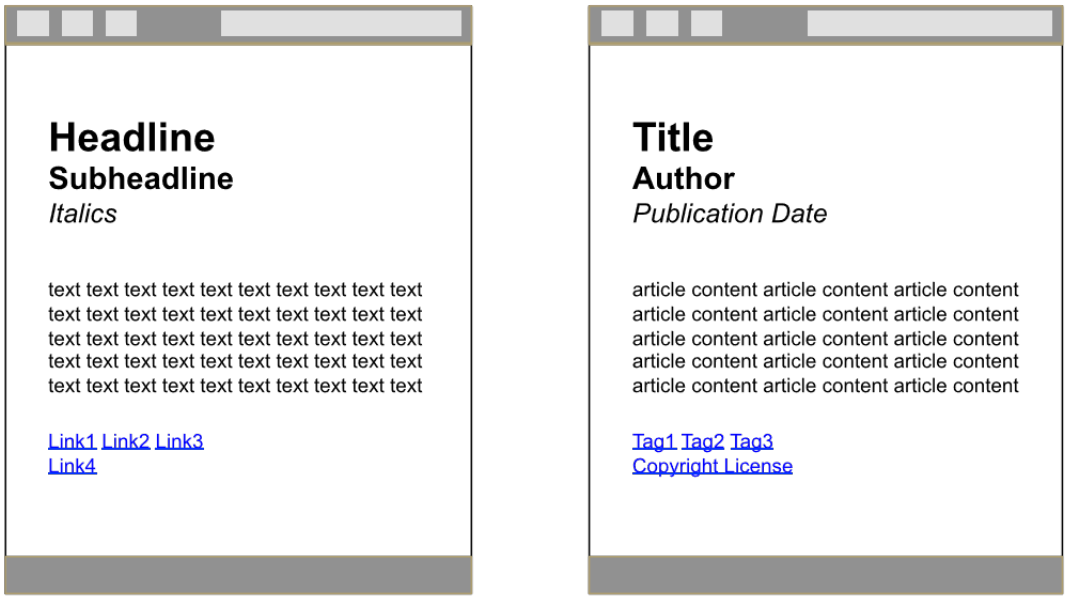

Figure 1.7: Illustration of the difference in what autonomous systems see when crawling a web page and what humans see when reading the same material. Figure 1 from [7]

graph into the web page instead, and therefore, the data becomes captured instead of described. The language does have the ability to transform into RDF, but the slight nuance between intended use means that a more complicated deployment of the attributes will be required. Using a previously established standard eases the adoption of encoding required to communication change information to autonomous systems.

1.4 Provenance Distance

Understanding provenance and workflows only provide only a portion of the view into a data set's lineage landscape. The workflow provides an understanding of how a data set fits into the bigger picture of data analysis and the provenance gives a method to reproduce the data set for data quality purposes. As such, workflow and provenance describe data sets in a very flat and static manner, allowing

for prospective reasoning as to how it may respond to changes made by the data producers. However, this places the burden of determining the magnitude of change in quality, as changes in provenance mean changes in quality, on the data producer. Consider again that data quality is subjective with respect to the data consumer's usage, and the difficulty in determining the significance of a new version becomes apparent. With increasing complexity, data workflows have developed in such a way that even subtle changes have serious implications for other parts of the workflow [10]. The responsibility of determining and communicating the magnitude of data alterations falls to versioning systems.

A very rudimentary way to communicate change distance uses the version number of the data set. Returning to discussing the dot-decimal notation often used to identify versions, version numbering follows a hierarchical method of systematically counting the releases made to a data set or system based on the perceived magnitude of the change. The problem lies with identifying the extent of perturbations performed upon the data set. The primary function of the number is to indicate compatibility not changes. Therefore, a release can result from five perturbations or fifty modifications. These cases become challenging since some users will experience larger perturbations resulting from a change than others. As a result, using fewer well-established categories avoids this problem, but it loses many details in resolution of the extent to which the data set may change. In addition, there is no standardization as to what each of the numerals used in a version number represents, and this significantly hinders interoperability between information systems. Data managers will often discuss whether data sets are qualitatively different enough to warrant incrementing one of the version numerals. While the dot-decimal method is easy to implement and use, its broad categorization severely impairs its ability to express version information beyond a basic functional extent.

Another approach is following the provenance of two data sets and identifying differences between the lineages of the two data sets. The total difference between the data is known as their provenance distance. This distance measure is very new as the availability of computable provenance has been developed fairly recent. One endeavor to compute over provenance has shown a marked ability to predict

disk usage based on the lineage of a data object [31]. Efforts have also been made to summarize provenance representations to improve consumption [64]. While the ability to compute over provenance data has been demonstrated, the comparison of two provenance graphs has yet to be widely studied.

Using PROV to represent provenance data in a semantic model produces an acyclic directed graph with labeled nodes. As a result, the provenance distance problem reduces to the similarity measurement problem. When measuring similarity, algorithms determine how far two graphs are from being isomorphic [65]. General graphs have similar complexity to determine similarity, but node labeling simplifies this process by providing a method to match nodes together. Other methods also exist to determine similarity under different conditions such as edits necessary to transform one graph into another [66]. Some methods focus primarily on edge changes [8]. In Figure 1.8, the left graph transforms through a move of edge 1 and a rotation of edge 4, resulting in an edit distance of two. Such changes in a provenance graph would demonstrate a change in dependencies between objects used to generate a final product of note. A difference in provenance distance would then provide context for differences between conclusions. Similar results from data with small provenance distance would ensure reproducibility, while large distances signals reinforcing confirmation of a result. However, differences can always be expected when considering different versions with more detailed comparisons required to determine the implications of changes made to a data set as a result of a provenance change. Meanwhile, when results disagree, a small provenance distance indicates that findings may not be reproducible. This kind of analysis resembles comparison measures employed in determining semantic similarity [67]. The main difference lies in semantic similarity comparing the distance between two concepts within a graph as opposed to the distances between the graphs themselves. However, it does reveal that using semantic graphs can have incredible impact in extracting implied relations between data they store.

There already exist methods which compare workflows based on quality criteria that leverages provenance to bound quality of service [68]. However, these procedures focus primarily on quick retrieval and efficient storage instead of lever-

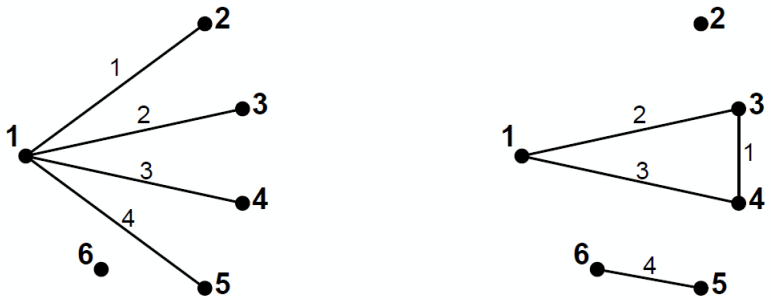

Figure 1.8: The labeled graph on the left transforms into the right graph under two edge edits. Figure 2 from [8]

aging the latent information accessed by reasoning across data set versions [69]. The distance measures previously mentioned rely solely on provenance graphs to compute results, but this is obviously insufficient. When considering the provenance of a data object, methods only consider the activities and entities that took an active role in the production of it. A new version of an object has a familial relationship with its previous versions, but in most cases, they do not take an active role in its generation. For this reason, detailed change information falls outside of provenance's scope and it can be seen in PROV using a single relation to link different versions of a data object. Without detailed change information, determining the difference between two data objects in a metric beyond broad strokes becomes difficult if not impossible.

This is not to say that provenance becomes useless in computing change distance, but it largely serves as an indicator than a measure. If there is any difference in provenance, then something must have changed. For example, if workflow uses a new script to generate a data set, changes can be expected in this new data set. However, what this script does differently than the old one requires a more detailed understanding and description than lineage can provide [33]. Additionally, if no changes were made to the script, but new data was produced, it likely indicates that some inputs have changed. The ramifications for the resulting data set will be difficult to determine without understanding how the original inputs have changed. Only knowing that they have changes is insufficient. Being able to understand the extent that modifications to data or workflows impact the results greatly improve a

producer's ability to generate high quality data.

1.5 Data Versioning Operations

Architecture has a principle that says form follows function, but, for data, form equals function. As a result, data has as many different forms as it has functions. Biological experiments often use data within cyclical data workflows where outputs are immediately fed back into new experiments [29]. Even though the goal of the experiment is the final data set, all the intermediary data sets provide significant value in reaching the goal. Libraries store data about their collections in large databases where both old and new versions of literature need to be maintained [30]. Some data exist in such a highly constrained environment that it must be managed at near the hardware level [70]. The struggle no longer becomes generating data, but instead, fitting the data into a format that users find useful and can consume.

The challenge of data versioning systems is to provide a unifying environment that can handle the plethora of forms and functions of its data. At its core, versioning systems only need to concern themselves with three operations: addition, deletion, and modification. Most literature surveys do not realize the significance of this commonality as this means that versioning methods can be described by delineating how each operation is approached by a system [29] [13]. Data addition generally constitutes the least complicated versioning operation because it interacts the least with pre-existing data. However, new data does share context with pre-existing data and provides a method of measuring data set growth. Since data sets no longer have to be used in its entirety and can be freely subsetted, a data set's complexity increases significantly with its growth. Every new file added to a data set doubles the number of available subsets.

Data deletion, however, has a more philosophical difference between systems. From the perspective of a versioning specialist, data should never be deleted since knowing why data was excluded is as important as knowing why data was included. The software versioning manager GIT uses a method of compressing older data to conserve space without deleting the data [5]. Pragmatically, this is not always possible due, generally, to the physical constraints of storage space. In high energy

physics, observational data often cannot be re-collected due to cost, and as a result, poor quality data cannot be re-processed or replaced [35]. The decision in this document is to use the term invalidation when referring to data removal operations as it implies that whether permanently deleted or not, there exists a more valid alternative.

Data modification encompasses the most involved data versioning operation. As a result, it often comprises a majority of the description of a data versioning service. In truth, data modification can be summarized as the invalidation of an instance of a data object - which can be a file, a record, or anything in a data set - followed by the addition of a new instance of that data object. However, this kind of operation is used so often to fix errors and update data sets that it is considered a unique operation. Modification owes its complexity to interacting with both pre-existing data from the invalidation stage and new data from the addition stage. However, this compound relationship fully contextualizes the relationship the operation has in relating the old data and the new data. In some cases, this only provides forward or backwards references between data versions, but having both gives users context for data's current state and update to new data [71].

Due to the ubiquity of the data addition, invalidation, and modification operations in versioning systems, the conceptual versioning model presented in Chapter 3 centers around capturing the relationships established by each of the operations. While other functions exist commonly in versioning systems such as object locking to prevent simultaneous conflicting changes, viewing to see the version an object belongs to, and branching to allow distributed modifications, these functions comprise the space of utility operations that support the three core processes.

1.5.1 Types of Change

The study of versioning operations further breaks down into categorization of change types data sets may undergo. While the meaning of operations are fairly easy to understand, not all changes have the same impact. As mentioned previously, version numbering separates perturbations into categories based on the impact the producer believes it has on the project. In this project, changes are categorized into

scientific, technical, and lexical changes. The granularity of the categorization does not consider the magnitude of change within the individual values stored by a data set as actual values vary depending on application and domain. Focusing on a more abstract representation of kinds of change allows for a better understanding of its impact while not being too precise to be domain specific.

Scientific changes comprise the family of changes which have the greatest impact on a project or data set. It indicates that modifications have been made to the most fundamental methods used to create a data set. These can include changes to algorithms used or sampling methods which may require a change in how users consume the new data. These changes have the largest implications for data consumers as it can have serious consequences for the soundness of their results. However, these kinds of changes is not always caught on the production end of data generation. While very large modifications can easily be determined to produce a scientific change, more subtle changes or interactions can also have larger ramifications, and data producers may initially view this as a technical change due to data quality's subjectivity. Technical impacts do not change the underlying science of the data, but impose a large enough change as to warrant notice. Structure alteration and unit conversions count as technical changes since the dataset now needs to be consumed differently but remains valid for use. In one of the data sets used by this projects, concentration units were originally reported in parts per million and then in cc isotope ratios. This would constitute a technical change since the data presents the same scientific measurements but in a different manner. Lexical changes belie the transformations that can best be described as corrections. Filling in previously missing values or fix erroneous values may be lexical changes. While they have the smallest impact on results and conclusions, these changes can allow computations to be performed when previously missing data discouraged such behavior.

The exact category that a particular change falls into can be controversial. The decision to change concentration units from parts per million to milligrams per milliliter poses a Technical change for a data producer. However, for a data consumer, the change may be viewed as a Scientific change as it invalidates the methods they had previously used. This conflict in view illustrates the data consumer-

producer dynamic. In general, data producers are in control of the methods of versioning, but data consumers determine the classification of a data change. Producers tend to use versioning systems to ensure data quality of service through audits and recovery tools [35]. Meanwhile, a consumer will analyze the historical changes and determine the impact this may have to their data use. As a result, this means that data versioning systems must communicate a dynamic view of the changes in a system contextualized by the user of that data.

1.6 Thesis Statement

The growth of innovative data capturing and storage technologies has led to new challenges in properly tracking the changes stored data sets undergo. Researchers store data in a variety of formats, from documents to databases, but since the growth in project scope, many have relied on versioning methods from software management technologies to track their data's evolution. However, these techniques fail to properly capture the changes data undergoes because they do not take into account the impact that a data's structure has on its function. In order to maintain data quality, producers use provenance meta-data capture the series of activities and agents involved in generating a data entity. Emergent technologies and frameworks have been developed to digitally capture this information including OPM and PROV. This data can be used to give insight into the magnitude of change a data entity has undergone by comparing the differences in provenance between the two entities, but this can only be done in broad strokes without more detailed change data.

By looking at the versioning transactions operating on a data set, data consumers can have a better idea as to the extent their data changes. This thesis document develops a concept model to formally characterize the activities and relationships among data objects when transitioning between versions. The model would improve current discussion on data versioning by providing a common understanding of terms and activities, changing versioning from a rule of thumb chore to a valuable research activity. It then demonstrates the model's utility by applying it to tabular spreadsheets, and then generalizes it to other contexts. In addition, this con-

tribution addresses in more detail the problem of measuring change distance, which using only provenance data could not accomplish, by utilizing the graph structure of linked data. Through this process, it creates and makes accessible machine readable change logs which allows for clearer deductions when comparing the extent of changes. This enriches the version documentation process without developing new artifacts, but now allows for the introduction of automation into part of the data auditing workflow. In addition, the method reinforces the need for dynamic, publicly available change information as data becomes more flexible to accommodate research in more diverse fields.

CHAPTER 2
PREVIOUS WORK

A number of instances in the literature mentioned previously uses the term versioning and provenance interchangeably. Mayernik et al. also notice a similar phenomenon although they use the term lineage instead of versioning [72]. The version model introduced by Barkstrom in Figure 1.6 to organize NASA's satellite data collection actually refers to a simplified workflow describing the provenance used to produce each level of data [6]. The diagram does not compare objects from the same level since changes to contributing components are only used as an indicator for version change. In actuality, objects have much more complicated development structures than the one dimensional lifespan indicated by the transition from Level 0 data to Level 3. The PROV Ontology model in Figure 2.1 outlines more explicit inter-relations between data objects, and it provides a new dimension with which to consider the interactions of data objects. More specifically, it outlines the explicit process of an agent performing an activity using an entity to produce a new entity. In the context of the level system, an agent (either a program or individual) performs a "Produce Instantaneous Fields" activity using "L1 Data" to produce "L2 Data." However, higher level data sets rarely use only one instance of lower level data. Calibration values may result from daily readings collected from another data set, but the generality of the ontology allows these relationships to be explicitly expressed. A more realistic provenance graph looks like the one in Figure 2 of an ozone indicator in which a Level 3 object results from the interrelation of multiple lower level products. An interesting observation of note is that Tilmes remarks in 2011 [10],

> Consider the relatively common case of the calibration table, which is an input to the L1B process, changing. Even though the version of the L2 or L3 software hasnt changed, the data files in the whole process have been affected by the change in the calibration.

which Barkstrom already observes in 2003 [6]

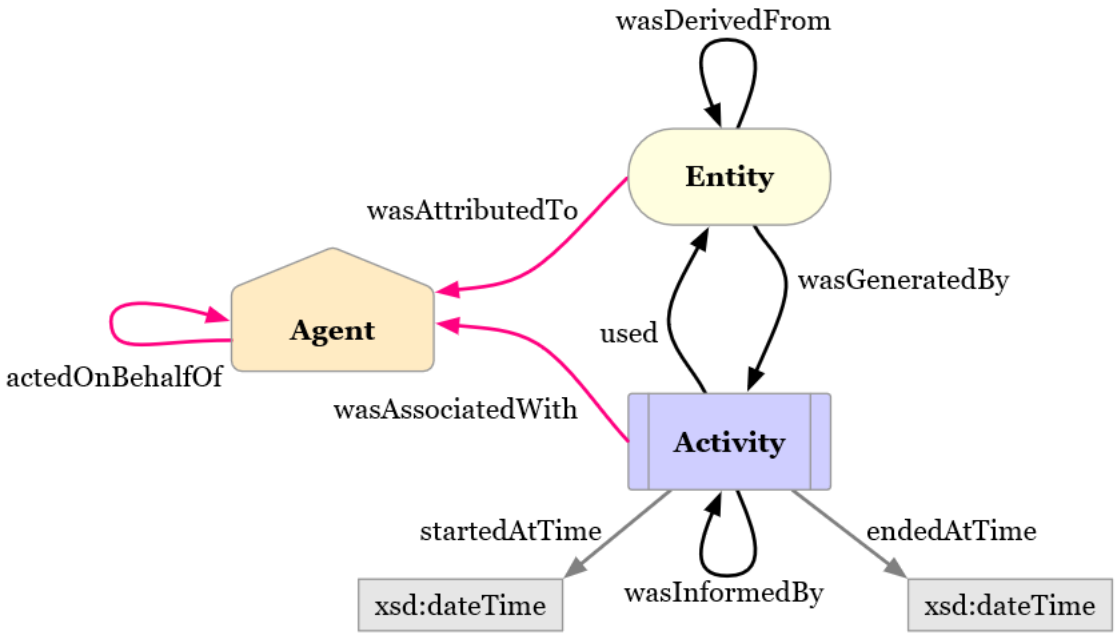

Figure 2.1: Diagram of the PROV Ontology. Figure 1 from [9]

> If scientific data production were easy, instruments would have stable calibrations and validation activities would discover no need for corrections that vary with time. Unfortunately, validation invariably shows that instrument calibrations drift and that algorithms need a better physical basis. Within a Data Set, we can think of a Data Set Version as a collection of files in a Data Set that have a homogeneous Data Production Strategy. Within a Data Set Version, we expect the code producing the files to be stable. We also expect that the algorithm input coefficients will be stable as well. The intent of data production is to produce data whose uncertainties are statistically similar under similar conditions of observation.

indicating a basic view that despite eight years in difference, the continuation of software focused versioning resulting in difficulties of data oriented collections.

If the level system provides a length and provenance indicates a breadth of a workflow, a version system can be considered to provide a height to a total workflow. Referring back to the HCLS data model terminology in Figure 1.3, as objects within a workflow, as in Figure 2, change versions, the structure of the workflow as well

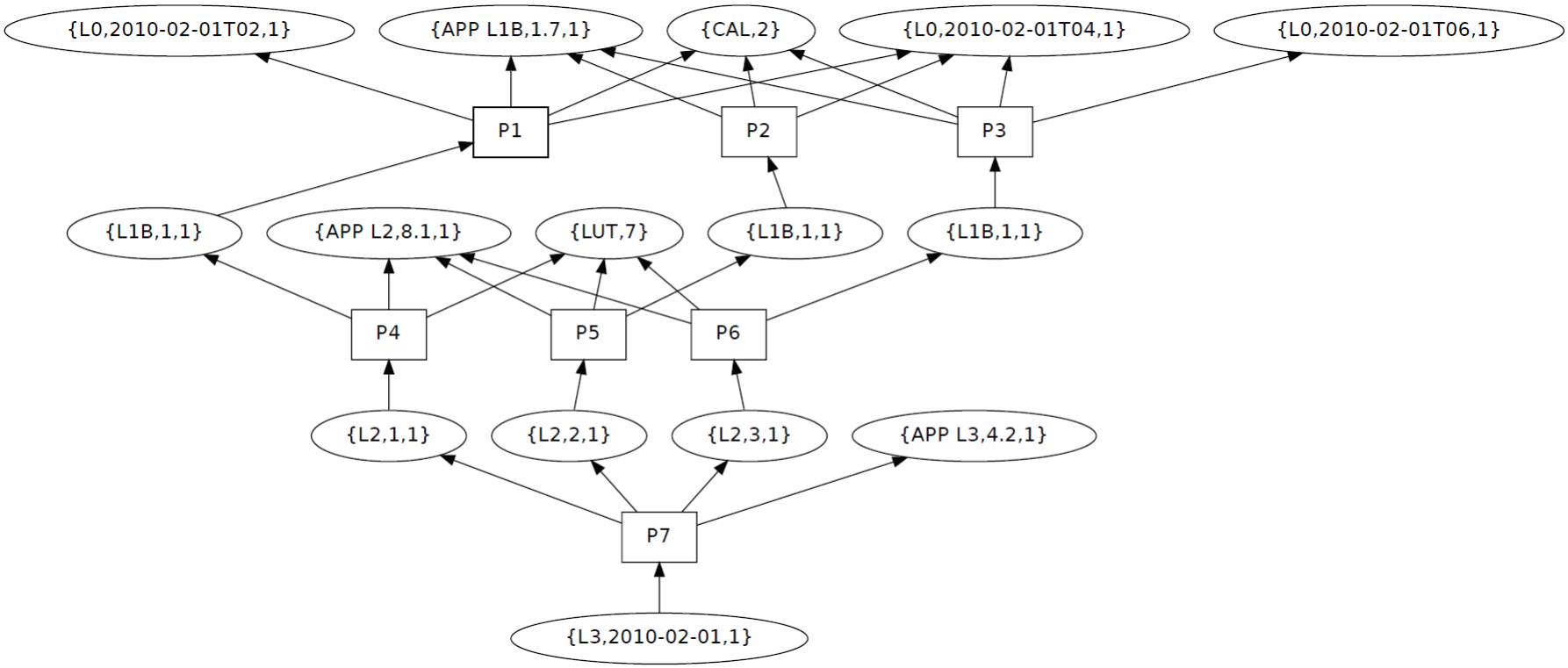

Figure 2.2: Provenance graph of a Level 3 data product, showing the inter-relations between different data products in generating the final product. Figure 2 from [10]

as the Summary Description of the final object, in this case the L3 Ozone product, remains the same. Instead, the new versions add layers like building blocks over the foundations of the original workflow structure. Version control systems then provide the mortar linking the blocks together to give the lineage capture procedure a solid structure. The PAV Ontology provides a means to track versioning information through linked data by introducing *pav:version* to cite versions and *pav:previousVersion* to link them together in order [17]. It does so in comparison to the Dublin Core concept *dc:isVersionOf* which records, "Changes in version imply substantive changes in content rather than differences in format" [73]. PAV argues that a new concept becomes necessary to cover cases where new versions do not have to be substantive but can still be alternate editions of the original object. Of note is the retrospective nature of the PAV ontology and PROV-O since it places primary emphasis on the most recent edition of an object. Figure 2.3, shows how RCS stores older versions as back deltas and branches as forward differences. The retrospective nature of back deltas results from development focus on the latest version, in this case 2.2. However, the forward differences provide a method to migrate from version 1.3 to the front of the branch. This characterizes the difference between a focus on data tracking, like that performed by provenance, to data migration, which users must undergo in order to consume the latest version. Mayernik et al. also find that, "Prospective records document a process that must be followed to generate a given class of products whereas retrospective records document a process that has already been executed" [72]. Retrospective provenance and versioning provides the ability to ensure data trustability and data quality among resources. However, researchers must follow a prospective versioning record in order to keep their research up to date.

To bring the discussion to an actual application, the GCMD released version 8.4 of their keywords, adding a slew of new values and modifying a select few [74]. At the time of release, many data repositories can be currently assumed to be using the previous or older versions of the keywords. As the taxonomy is not a class-based ontology, changes to the keywords have significant implications to the semantics of a data set described by those keywords. A data producer wishing to expose their

data sets using the new version of the GCMD Keywords must use a prospective method to translate their current descriptions to the new version. By comparison, the GCMD would use a retrospective measure to record the changes made to their keywords. PAV would not be able to address the prospective problem as a result of it is, "a lightweight vocabulary, for capturing just enough descriptions essential for web resources representing digitized knowledge" [17]. A detailed transition from GCMD Keywords 8.3 to 8.4 would significantly undermine the lightweight nature of the vocabulary. GCMD does provide a short summary of changes made in a version, but this would come in the form of text rather than structured data.

2.1 Spreadsheets

In this project, spreadsheets were chosen for study as they resemble text-like data objects while still maintaining a level of complexities. Though not as well encapsulated as other data format types such as the Hierarchical Data Format (HDF) or Network Common Data Form (NetCDF), spreadsheets provide many helpful tools that scientist favor for quick data storage and distribution over comma separated values (CSV). There also exists other document-like formats that are not discussed in this paper such as eXtensible Markup Language (XML). The initial work was done with the "Noble gas isotope abundances in terrestrial fluids" workbook (Noble Gas) [75]. The "Paragenetic Mode for Copper Minerals" workbook (Copper Data) was used to give better insight into data changes due to collaboration with the data

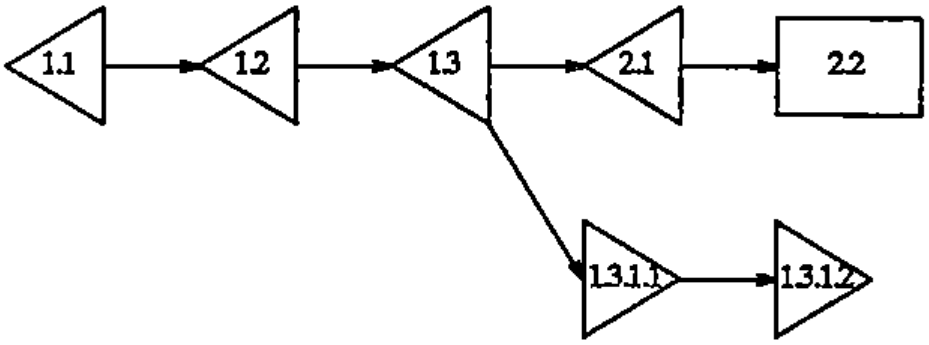

Figure 2.3: Commit history of an object in RCS with changes in the main line stored as back deltas and side branches stored as forward deltas. Figure 5 in [11]

set's author [76].

The Noble Gas data set was initially published on June 11, 2013 and then released a second version on March 8, 2015. Many significant changes were made to the data set between the two versions, which makes this data set particularly challenging to version. The physical structure of the data set changed from eight separate Excel spreadsheets to a single spreadsheet. The second version also trimmed 195 columns to 54 columns in the second release. In addition, many new locations were surveyed and added to the second release. Documentation accompanied the data set explaining different components of the spreadsheet and its usage, but it included no versioning information. This lack of versioning or transitioning information indicates a focus on data usage rather than data maturation, which is not a particularly bad approach. It makes logical sense to simply download the latest data set when it becomes available and not worry about the format of the invalidated data set. This approach convenient for new users of the data as the cost to consumer the new version of the data is the same cost they would have spent to acquire the data in the first place. However, users of the old data are disproportionately effected by the change in versions since old code and workflows may need to be updated to accommodate the changes in addition to the cost of consuming the architecture of the old data set. In this case, users would need to read the documentation to understand whether 182 from the June data set is still available in the March data and, if it is, in which column it resides in the March spreadsheet. This brings to light the additional concern for the Noble Gas data that the documentation is not easily machine consumable, meaning that all mapping activities will need to be performed manually. Not only is this approach time consuming, but it also does not scale well into larger data sets.

The Copper data set was acquired during the process of a workshop to generate new methods of visualizing mineralogy data, initially on June 8, 2016. The process entailed trying various orders and organization for the data and results in various new versions of the data that depend on varying filtering requirements, acquired on August 21, 2016. Unlike the Noble Gas data set, the Copper Data had no accompanying documentation, since the primary consumers of the data at the workshop

were also mineralogy experts. However, this data set had more stable characteristics including physical and logical structure. Only two columns were removed from the transition to the second version, but sixteen new columns were added to the data collection. It also demonstrates a change in orientation with respect to data usage since the previous data set was designed to be distributed for general usage and discovery. In this case, the structure and organization of the data within the set was driven for a specific purpose in the development of more expressive visualizations. As a result, versioning information is driven by developmental needs instead of the other way around with versioning information bridging the gap between software migrations.

The data files from both data sets can easily be tracked using standard version management services such as GIT or SVN. Likewise, there exist comparison tools like Spreadsheet Compare from Microsoft Corporation that can generate diff-like outputs for each of the data sets. In conjunction with commit logs, the comparison outputs provide a basic versioning methodology that describes the data set's evolution. However, these applications rely on human attention and interaction to operate, and with larger data sets, proper documentation becomes difficult to maintain. With the Copper Data, the demand for new versions of the spreadsheets exceeded the time necessary to document version history as a result of rapid product evolution during the workshop. In consequence, the process to manually commit and annotate changed data impairs the natural progression of scientific development.

2.2 Database Systems

Databases remain the most relied upon technology for storing and searching large quantities of data rapidly. While the dynamic combination of tables means that data bases remain flexible enough to represent complex objects, it also means that they represent a much more complicated case for attribution. Since tables may be combined in different ways to answer complex queries, indexes do not remain constant across requests to the database. The approaches to database versioning typically focus on ensuring the reproducibility of queries to the database. This can often be difficult as with spreadsheets since changes to the content or structure can

result in different solutions from the database for the same query even using time stamps. For example, consider the query to select all columns of row A from a database on March 1st, then the database undergoes a schema change to add a new column to the table on April 1st. A subsequent request for all columns of row A would include the new column which does not represent the response on March 1st. In addition, even if the data is timestamped, the time signature is associated with the row and not the schema, meaning that the query may still return row A with the new column with a NULL value, depending on the distribution. The query, not the data, would need to be modified to exclude the new column.

This presents as a challenge because unlike data files and spreadsheets, databases are generally not instanced. Databases often store massive quantities of data and replication of that data to archive snapshots or distribution frequently proves too costly to be feasible. Instead, interaction with the database occurs from a centralized source through transactions. Various methods have been studied to manage changes within these systems focusing primarily on schema versioning, emphasizing data's structural component [77]. This provides a method to enact a transactional rollback on the database to execute queries in an environment reminiscent of the original execution. The framework of the resulting database environment can become quite complicated as a result of the complexity of the tables representing intricate data objects [78]. This results from the need to manage the time instances of realization, storage, and validity. The datum becomes realized at collection, then stored upon entry into the database, and finally valid until the present or new data replaces it. More recently, new methods have been developed to adjust to the enormous quantities of data populating modern databases, focusing on query citation rather than data citation [79] [80]. Citation by query avoids the complexities involved with referencing data that can grow and move. However, this method relies on the existence of a versioning system for data. This method also recognizes that modifying queries to operate on the current state of the database may often be easier than rolling back transactions or schema to reproduce the results of a query [81]. As a result, to versioning a database system may be more feasible as data size increases by applying methods to the query results and not to the data.

The RRUFF Database is "an integrated database of Raman spectra, X-ray diffraction and chemistry data for minerals" [82]. It features a web accessible change log using the transactional log generated by the database software[1]. As the records in specific tables change, the log reports these changes, supplying persistent access to the modifications made to the RRUFF data. The approach to this alteration information highlights the always on-line approach to modern databases where changes to the data do not constitute a new database. The log demonstrates strong versioning characteristics with not only a breakdown of the change components, but also a commentary on the motivation for the difference. In addition, its HTML structure allows automated web crawlers to systematically consume the version information. With the integration of web ontologies, the change log would also be intelligible to automated agents.

2.3 Ontologies

On-line ontologies are a different way of storing data than relational databases that has found significant traction within Semantic Web applications. They form graphs, relating a vocabulary of terms and relationships together to model complex interactions within an application's domain. Since the ontology is represented as a graph, it has more expressiveness than relational databases. The objects no longer need to share uniform structure and fields when entered into the database. Ontologies improve interoperability between scientific data sets by allowing differing data to share a common vocabulary and be comparable. Like other data, ontologies change regularly as definitions and relationships update to better represent their source material [83]. As a connected graph, they easily lend themselves to providing mappings between changes and versions within the ontology. New transitions would be represented as a simple link between new and old concepts. This is particularly important on the Semantic Web since most reasoning and interactions are handled automatically by underlying services. Ontologies, thus, benefit the most when providing both forward and backward mapping as it allows more up to date systems to interact with entities that haven't migrated yet [71]. Incomplete mappings, where

[1] `http://rruff.info/index.php/r=rruff_log_display`

transitions exclude either forward or exclude backward mappings, retain value as backward mappings inform traceability and forward mappings communicate advances in the domain. However, the uncertain landscape of web services means that full ontology mappings prove invaluable to making data inter-operable. Advances in ontology change detection have made tools which automatically generate mappings between versions of an ontology available [84]. However, in this project, the focus remains on improving the description of these mappings to provide not only descriptions but also explanations for the transition.

The Semantic Sea Ice Interoperability Initiative (SSIII) is a project combining the efforts of the National Snow and Ice Data Center (NSIDC) and the Rensselaer Polytechnic Institution (RPI) Tetherless World Constellation [85]. In their initiative, they developed an ontology to describe a core set of sea ice terms [86]. They have made available the second version of their ontology. The next edition of the ontology follows some adjusted best practices to improve modularity and coverage of the sea ice vocabulary. The new formulation would require the movement and addition of new concepts as well as the modification of other ontological entities. As of this writing, the ontology does not have a method to provide mappings between versions of the ontology even though the concepts are managed through GitHub. Currently, Ruth Duerr has developed a color coded visualization of the new ontology's layout using a CMap seen in Figure 2.4. The maps use a five color system with green representing no change, purple showing a moved concept, yellow coding an unmoved entity with changes to its name or definition, red showing a change to the ontology or nomenclature, and blue representing an entirely new concept.

The Global Change Master Directory (GCMD) is a metadata repository used by NASA to story records of its available data sets [87]. It employs a keyword ontology to search for Earth science data in NASA data sets. These keywords tag and label datasets into strictly defined categories in order to make them more discoverable [88]. Version 1.0.0 of the GCMD Keywords was published on April 24, 1995, and as of the time of writing, the most recent version of the keywords is 8.4. As can be seen, the naming scheme of the versions changed since the first publication of the keywords. In the initial scheme, each part of the decimal system

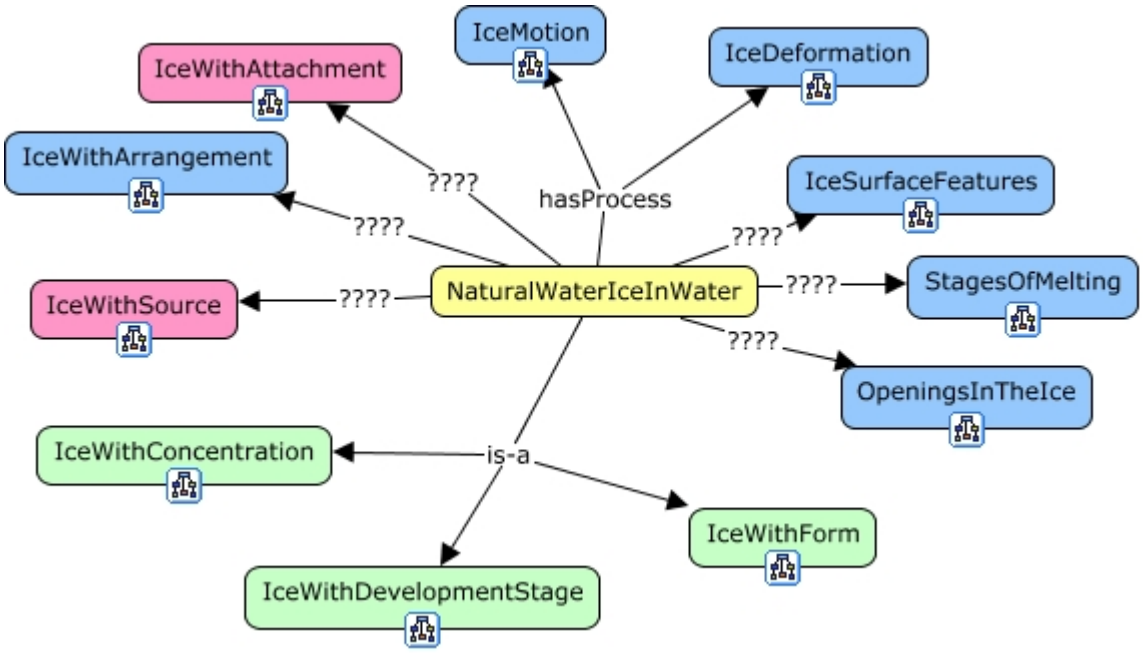

Figure 2.4: Concept map created by Ruth Duerr to organize the Sea Ice Ontology's Version 3 development.

represented a different level of the GCMD Keyword hierarchy: category, topic, and term, respectively. Incrementing a number in the version name indicates a change occurring in that level of the ontology. However, this gave very little information on compatibility between versions, and the ontology currently employs a more standard Major.minor release naming scheme.

The data set provides a very interesting case to study because of its history of medium storage. The GCMD originally distributed the keywords in a spreadsheet format, but later migrated onto database services as the scope and demand grew. The data may currently be accessed through a dynamic web service that can provide results in a variety of linked data formats. As a result, it leverages the endeavors made towards the environments mentioned in the previous two sections. The keywords have an accompanying change log, but due to the variety of mediums, the early logs are difficult to interpret. Since they attempt to use web technologies, the keywords each have unique identifiers that can be dereferenced using a Universal Resource Identifier (URI). Attribution, therefore, has mostly addressed by the source material. This is to be expected as a result of curated application of linked data

principles. Due to the shift from spreadsheet to databases, there exists a disconnect between the early versions of the ontology and modern editions. The work done in this project will be able to link them and provide a road map through the evolution and migration of the vocabulary as well as guide future evolutions of the keywords.

CHAPTER 3
CONCEPTUAL MODEL

The goal of dataset versioning is to expose the relationships between versions of a dataset. To do this, the concept model relates three kinds of objects, versions, attributes, and changes, with three kinds of changes, addition, invalidation, and modification. To do this, we create a mapping between an original set and a new dataset. As mentioned previously, the operations conducted by data versioning systems boil down to primarily three operations: addition, invalidation, and modification. Since these operations are so prevalent, we use these three procedures to characterize the relationships between versions. A modification is straight forward to model because it maps together two attributes of the version that exist, but addition and invalidation are a little different. Because the item doesn't exist in one version or the other for addition and invalidation, it forms a '0 to 1' relationship between the attributes. This causes a problem conceptually because without a concept on one end, there is nothing to connect on one end. The chosen solution was to use the version concept as the anchor in place of the non-existant attribute. This observation leads to the structure of the conceptual model used in this dissertation. The construction of the relationship decides the kind of change that is occurring. It then becomes easier to identify which change is occurring based off of whether attributes exist or not in which version. In addition, while the figures in this chapter only show the attribute relationships as 0 to 1, 1 to 0, and 1 to 1, it is more valid to consider the relationships as 0 to X, X to 0, and X to Y in cardinality. A modification may change a single location attribute into two separate latitude and longitude entries, for example.

From the discussion above, three kinds of objects appear: versions, attributes, and change. That is to say we cannot properly represent versioning with versions alone. The changes which we are interested in result from comparing the parts or attributes of the versions. The Dublin Core Term hasPart provides a sufficient property to relate versions and their attributes together. An observation that will

not be further explored in this research is that attributes can also be versions. For example, when comparing two revisions of a dataset, the attributes would be data files. However, these files will be versions of each other, and their attributes will be, if tabular data, the rows of each file. This nesting speaks towards the granularity by which an individual desires to perform version, but also demonstrates the challenge of using major and minor numbers to capture version change with the current dot-decimal identifier scheme.

An obvious concern about using this method of mapping, of finding attributes that are common and uncommon, between two or more objects is not unique two versioning. In order to ensure that the relationships being exposed by the mapping is valid, we must go back to the definition of versions. By requiring that the objects to have common provenance, we establish that performing a comparison to make a mapping compares related objects and information. The second requirements, that the objects share the same workflow step, establishes that when we do make a mapping, that the attributes actually are the same. This also addresses the possibility that we are comparing objects that have different purposes at separate points in a workflow, but share provenance as a result.

3.1 ADDITION

When a change adds a new attribute to a version, it only needs to refer to version two and its corresponding attribute. The reasoning should be fairly obvious as the attribute never existed in version one, and therefore, there is nothing to refer to and no need to form a relationship between the change and version one. However, by linking the addition change to version one, we address a difficulty with comparing provenance graphs. When two data objects have identical structures, it is difficult to determine what time the objects were added to the dataset and which version they belong to. As a result, determining the compatability of the two objects becomes difficult. The change contributions to the dataset evolution appears naturally using this construction. The resulting model can be seen in Figure 3.1. Some relationships are specifically left out, such as that between Change A and Version 2, to not confuse identification of other types of changes. The relationship between Change A and

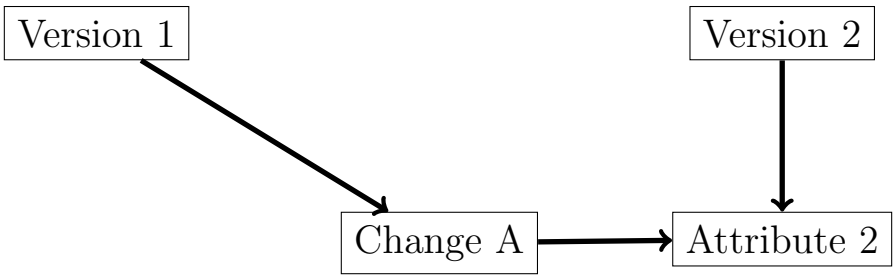

Figure 3.1: Model of the relationships between Versions 1 and 2 when adding an Attribute 2 to Version 2 as a result of Change A

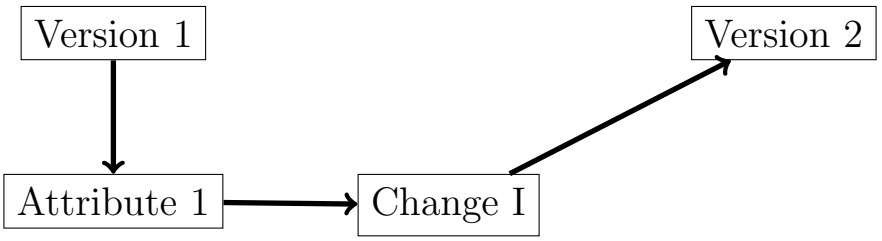

Figure 3.2: Model of the relationships between Versions 1 and 2 when invalidating Attribute 1 from Version 1 as a result of Change I

Version 2 can still be implied from Attribute 2.

3.2 INVALIDATION

The Invalidation operation corresponds to the delete concept found in other applications. The choice of invalidation over delete results from the policy that, in versioning, data should never be deleted. In practicality, this may not be particularly feasible due to space limitations and relative validity. In either case, the change invalidates an attribute in version one, resulting in version two. Unlike the Addition operation, Invalidation forms a clear relationship between both versions, which can be seen in Figure 3.2. Notice again that since Attribute 1 no longer exists in Version 2, there is no corresponding Attribute 2 to refer to.

From Figure 3.1, we can see the confusion that could result from requiring explicit relationships between versions and changes in both the Addition and Invalidation operations. Linking Change A to Version 2 would create a duplicate connection and provides a mechanism to identify when items specifically enter or leave a version.

3.3 MODIFICATION

The final operation is Modification, and it maps a change from one attribute from version one to its corresponding attribute in version two. The particular type of change in this case is purposely left out in order to allow data producers to subclass and customize the resulting graph to properly reflect the versioning that they desire.

3.4 MULTIPLE LINKED VERSIONS

Using the construction outlined in the previous three sections, many changes can be compiled together into a graph in a changelog. After all additions, invalidations, and modifications have been compiled into a single graph, a complete mapping from version one to version two may be developed. The orientation of the relationships in the graph allows a flow to be created from attributes in version one to corresponding attributes in version two. Taking version two and performing the same graph construction to a version three results in not only a flow from version two to version three, but also from version one to version three. As a result, the flow can be used to construct a mapping from version one to version three or any future version.

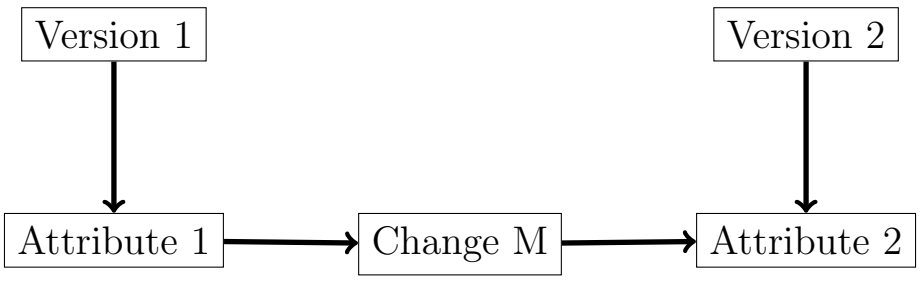

Figure 3.3: Model of the relationships between Versions 1 and 2 when modifying Attribute 1 from Version 1 as a result of Change M, resulting in Attribute 2 from Version 2

CHAPTER 4
VERSIONING TABULAR DATA

The initial goal of the research sought to develop the method of calculating provenance distance between two data objects. The value in this measure lies in determining whether similarity of the activities responsible for producing the objects provides context for reproducibility and result comparisons. The "Noble gas isotopes in hydrocarbon gases, oils and related ground waters" database has the desirable qualities for this comparison of varied sizable provenance and multiple versions to provide comparable changes. However, gaps appeared to hinder the approach with using provenance data to measure change distance. To begin, each of the spreadsheet database's rows were considered to be a separate data object, as opposed to the individual file since this structure changes in the subsequent version, as explained later. Each row contained an entry indicating the reference used to compile the readings stored, and this entry was used as the data entity to produce a provenance graph with the PROV model as seen in Figure 4.1. An important challenge to note when creating these graphs is that in version 1, the references were stored in a very human readable fashion. The entry could be stored as a string or numeral even though all values were numbers. In addition, the values were both comma separated and ranges indicated by a dash. Version 2 of the database corrects many of these problems with consistent content type and presentation. The documentation which accompanies the data set does not detail any changes to the compilation procedure that would indicate this improvement. The conclusion then follows that even though the two data objects have essentially the same provenance graph, it does not capture the operational change which has occurred within the data.

Version 2 also imposes many new changes that improve the data's readability. This can be seen with the unification from files per region to a single file, reduction of columns, and better standardization of value format within a column. The accompanying documentation includes instructions on how to read each column within the spreadsheet, but makes no mention as to the changes made to the

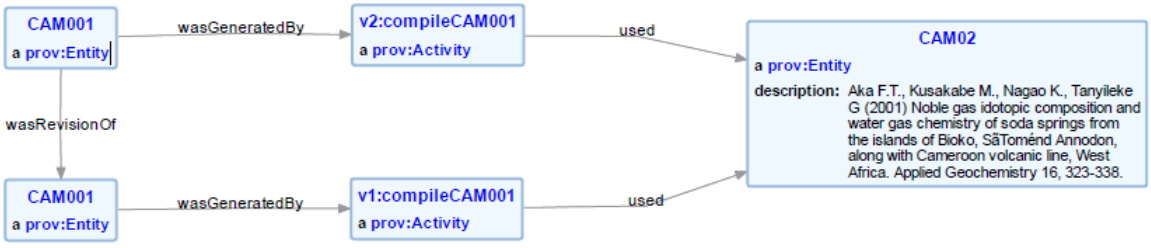

Figure 4.1: Provenance graph for the entry CAM001 entry of the Noble Gas Database. Other than the labels, the structure of each of the data objects is very much the same.

original version to produce the current release. In software, this would take the form of a change log, but they also provide the developer a chance to explain his or her motivation for making those changes. In this case, an example would be the two versions reporting concentration in different units. As a result, the first goal to quantify the amount of modification between the first and second versions needed a change log document to codify the differences. For small applications, a listing of modifications sufficiently explains the transition to a new data set, but for larger applications a machine-readable change log demonstrates potential for significant value as previously mentioned.

Lack of familiarity with the data set and it's authors immediately posed a challenge to verifying the resulting change log's validity. The Paragenetic Mode for Copper Minerals database did not have these same constraints and also featured a more limited set of change. With the process's validity now verifiable, the versioning model could now apply to the resulting change log, but at that point, the model still included capturing the actual values in the data object that changed. Including the actual data into the change log gives concrete details as to how the object behaves when it changes, and is common practice. However, when modeling the version, data within the object provides a level of granularity that does not transport well from one information system to the next. In addition, the resulting linked data graph stores double the amount of data than the actual change, once for the linked data and again for the values. As a result, the model leaves out including the data. This allows the model to remain open and adapt to more complex versioning procedures.

The RDFa implementation in the HTML change log makes a trade-off that

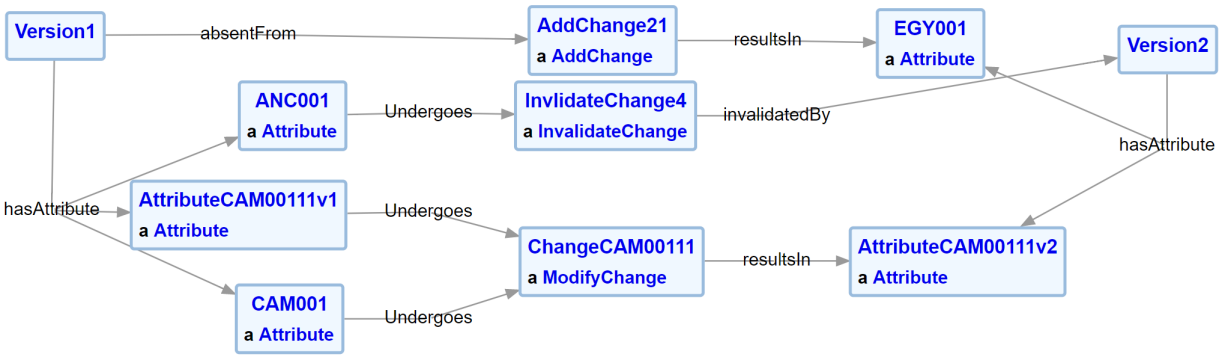

Figure 4.2: Some initial entries from versions 1 and 2 of the Noble Gas dataset

bends the original intent of the framework, but leverages its ability to translate into RDF. RDFa natively adds context to describe specific text instances, such as a string of text being a name or another constituting a phone number, by encoding it in the format Subject Predicate Object. The text being described appears as either the Subject or Object, and the remainder becomes implied from previous entries. In the change log, no text string directly denotes a modification so it must be explicitly injected through the document source. In addition, similar text entries appear close together, such as pairing column numbers with each other and keeping values side-by-side. However, this does not follow the order in which objects appear to encode them into the model, meaning that relationships must often be explicitly defined. The resulting source thus directly defines the entire relationship of the entries and objects into the graph without using any of the human readable parts of the log. However, this means that RDFa parsers can directly extract the full linked data graph similar to the one in Figure 4.

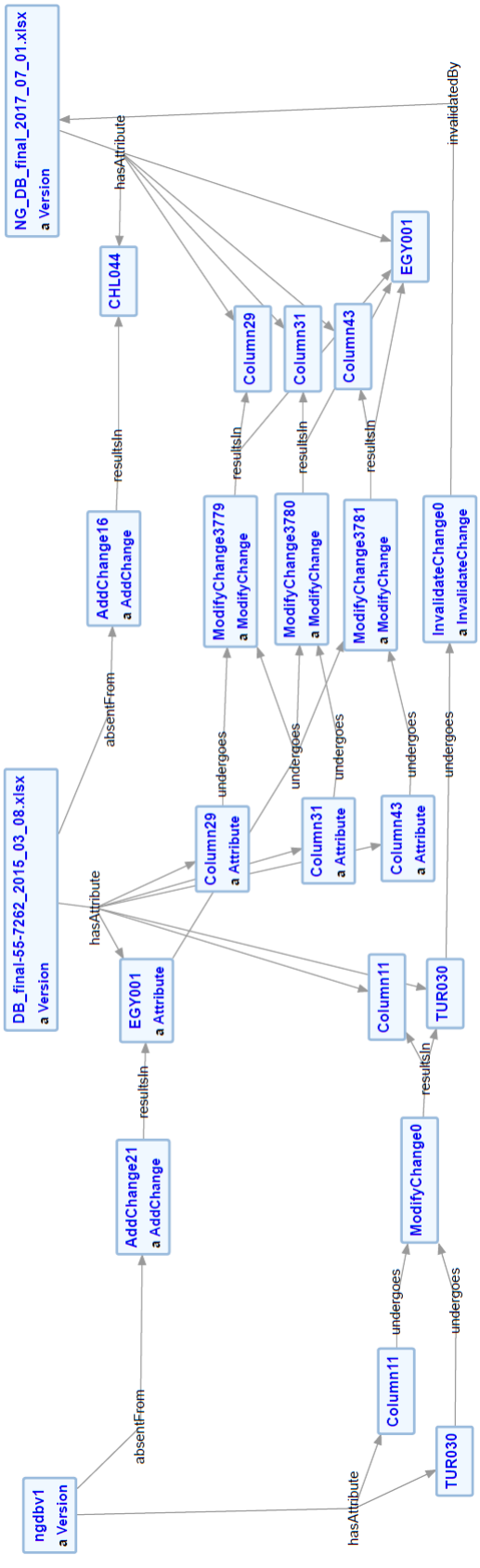

Figure 4.3: Versioning Graph representing the linked data graph with selected entries of additions, invalidations, and modifications after the publication of the third version.

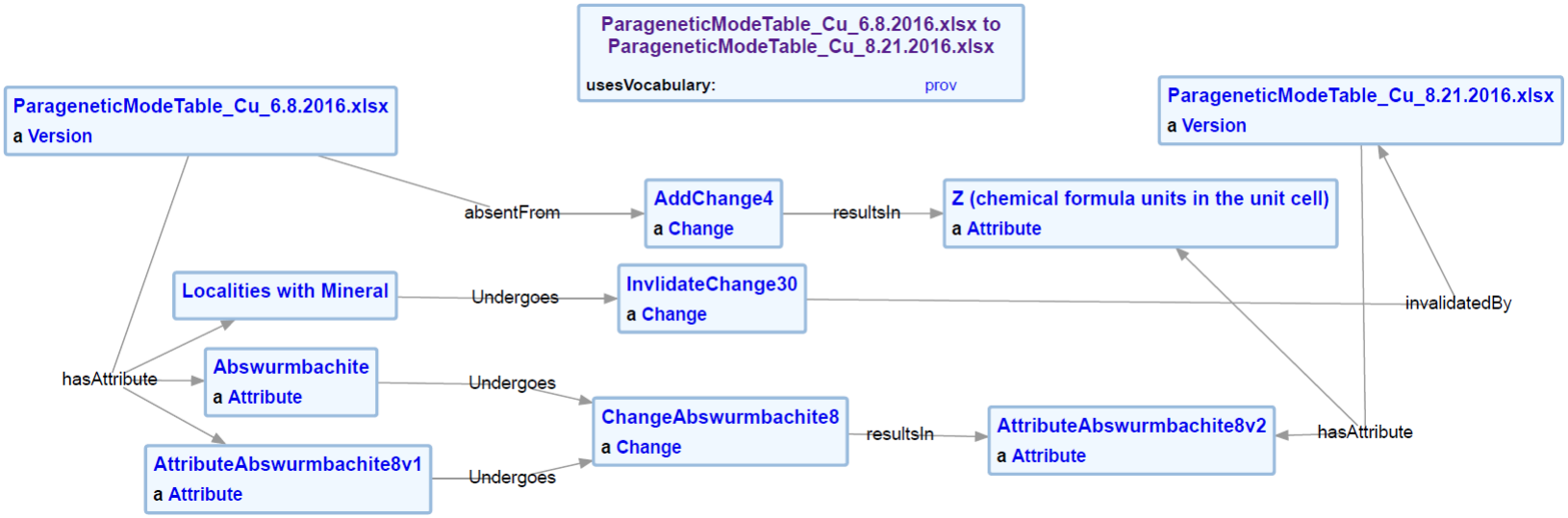

Figure 4.4: Versioning Graph representing the linked data graph with selected entries of additions, invalidations, and modifications.

CHAPTER 5
DATABASE VERSIONING

Databases already maintain a part of versioning history with a transactional log. However, they pose an interesting change in context compared to spreadsheets. Often version comparisons occur between instances of spreadsheet files, but with databases, modifying transactions do not generate a new instance of the database. Identifying a version would then need to adapt and link transactions to versions. This can be done through query based citations as described by Proll and Rauber [79]. The transaction log also more specifically states the attribute involved, making detection of new attributes into the database more straightforward. This addresses a particular concern in spreadsheet rows since their attributes have a tendency to be less consistent. Since RRUFF already possesses an automatically, web accessible change log, the work in this area focuses primarily on deconstructing their code to hook in the concept model with RDFa.

CHAPTER 6
ONTOLOGY VERSIONING

6.1 Sea Ice Ontology

A great many annotations were made to Version 2 of the Sea Ice Ontology. Maintaining those notes after migration to the new version would provide great value to the project. From the color code in Table 6.1, yellow and blue changes correspond directly to Modification and Addition transitions, respectively. In theory, the green concepts would also qualify for a Modify categorization since they would likely have a new URI, meaning that an attribute exists between two versions and something has changed. However, this would result in a product that is at least the size of the union of the two versions, greatly hindering the scalability of the approach. The obvious solution would be to leave the attributes unlinked, as in the approach with the spreadsheet application where no change was detected. The Invalidation change would cover concepts which do not have a mapping into Version 3. Therefore, the remaining transition types, purple and red require more specific attention than a usual Modify.

6.2 GCMD Keywords

GCMD Keywords do not qualify as a standard web ontology since it does not constitute a class hierarchy. As a result, the addition, removal, or modification of any term within the keywords has a significant impact on the semantics of using

Color	Description
Purple	Moved since the previous version of the ontology.
Green	Still the same.
Yellow	In the same place; but perhaps the name or definition have changed.
Red	Suggest changes to be made to both the ontologies and the nomenclature.
Blue	New concepts to be added to the ontologies.

Table 6.1: Color code used by the concept maps made by Ruth Duerr during the planning phase of the Sea Ice Ontology's development.

a keyword to describe a data set. More recent editions of the keywords, available through the Key Management Service (KMS), distributes they keywords in RDF format, and this allows them to be referenced through a unique web identifier. The identifier drastically simplifies the attribution of changes between versions of the keyword list. However, older editions of were stored and distributed using Excel spreadsheets. This provides an interesting juxtaposition between versioning the spreadsheet as mentioned in a previous chapter and tracking the keyword changes. Results from this activity enables data sets described with different versions of the GCMD Keywords to remain discoverable within the same system.

Better quality data for the early versions of the keywords needs to be acquired before scripts can make mappings to newer releases. While easy to reference, the identifiers used to store the newer keywords are not immediately interpretable so further work needs to be done in order to form a mapping. Once the versions have been mapped, the workflow for publishing the change data follows the same process as in the previous chapter.

6.2.1 Creating the Versioning Graph

The keywords can be acquired in a number of different formats, but the one chosen was RDF. In the file, each concept uses a unique URI as an identifier and uses the concepts skos:Broader and skos:Narrower to establish the hierarchy. When the GCMD group release a new version, they reuse the same URI for reused concepts. To create a mapping, new and old concepts can be easily filtered out by finding URIs which are also skos:Concepts but only exist in one version or the other. The remaining concepts must therefore be modification changes or unchanged. Since, the primary data the hierarchy contains is its parent and children nodes, modifications occur when the concept moves around in the hierarchy. To determine this, the values of a concept's skos:Broader relationship is checked.

Since the URIs are reused, the history of a concept can be tracked through the versioned lifespan of the concept.

In Figure 6.1, we can see that changing from version 8.4.1 to 8.5, there is a large spike in changes. Note that the number of each adds and invalidates number

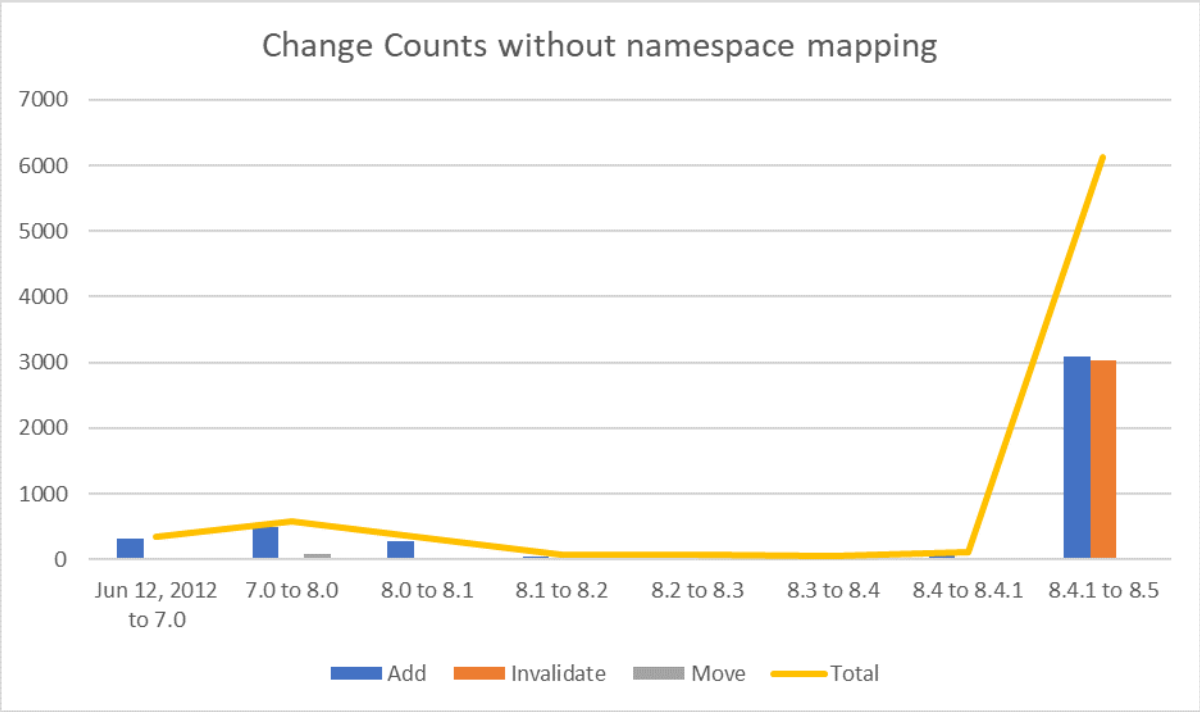

Figure 6.1: Add, Invalidate, and Modify counts in Version 8.5. The counts show change magnitudes and indicate that major and minor changes differ by orders of magnitude.

to approximately the size of the entire keyword list. This is a result of a NASA policy change requiring web resources to use the https protocol. As a result, all the URIs of the concepts changed, and the entire list was essentially removed and then re-added. Without the breakdown of the magnitude of changes into the three subtypes, the total reported change would be both difficult to interpret and exceed the size of the dataset significantly.

In Figure 6.2, we can see an accounting of the changes with the URI namespace. After controlling for the namespace change, we can see that the dataset is dominated by add changes. In addition, in migrations across major numbers, the magnitude of adds range in the hundreds. This is an order of magnitude different than the numbers of add changes while stepping across minor number differences. From the identifier scheme and the change counts, it is clear that the keyword management team expected only minor changes and names the version as such even though changes to namespace constitute significant differences according to linked data

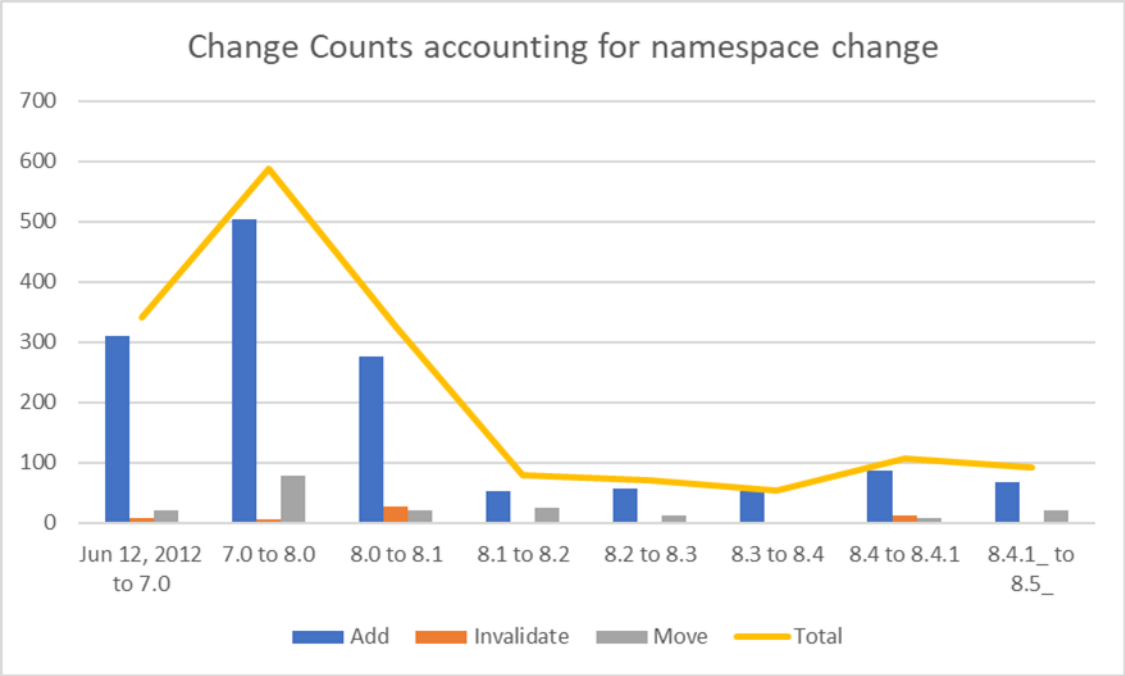

Figure 6.2: Add, Invalidate, and Modify counts ignoring the namespace changes in Version 8.5. The counts show change magnitudes appropriate for the identifier.

definitions.

CHAPTER 7
FUTURE WORK

While much has been expounded upon to develop the framework to conduct version distance calculations, nothing has been said as to how such a calculation would work. The concept model specifically relates objects in a single direction from attributes in older versions to the corresponding one in newer editions. Once values are assigned to each of the change types, a flux-like calculation can be performed to characterize the change moving from one side to another. However, the resulting calculation may need to be more complex since the length of change logs is not standardized. For example, a long list of small changes could over-shadow a few significant modifications. Possible solutions could include sub-classing the change types to give a wider range of weights or normalizing the values across a range to give comparable results.

Generating results for database and keyword versioning fall largely on applying attributes to its data since the remaining workflow remains the same across contexts.

REFERENCES

[1] M. Branco, D. Cameron, B. Gaidioz, V. Garonne, B. Koblitz, M. Lassnig, R. Rocha, P. Salgado, and T. Wenaus, "Managing atlas data on a petabyte-scale with dq2," *Journal of Physics: Conference Series*, vol. 119, no. 6, p. 062017, 2008. [Online]. Available: http://stacks.iop.org/1742-6596/119/i=6/a=062017

[2] B. R. Barkstrom and J. J. Bates, "Digital library issues arising from earth science data," 2006.

[3] R. E. Duerr, R. R. Downs, C. Tilmes, B. Barkstrom, W. C. Lenhardt, J. Glassy, L. E. Bermudez, and P. Slaughter, "On the utility of identification schemes for digital earth science data: an assessment and recommendations," *Earth Science Informatics*, vol. 4, no. 3, p. 139, 2011. [Online]. Available: http://dx.doi.org/10.1007/s12145-011-0083-6

[4] M. Dummontier, A. J. G. Gray, and M. S. Marshall, "The hcls community profile: Describing datadata, vversion, and distributions," in *Smart Descriptions & Smarter Vocabularies*, 2016. [Online]. Available: https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_3

[5] S. Chacon, *Pro Git*, 1st ed. Berkely, CA, USA: Apress, 2009.

[6] B. R. Barkstrom, *Data Product Configuration Management and Versioning in Large-Scale Production of Satellite Scientific Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 118–133. [Online]. Available: http://dx.doi.org/10.1007/3-540-39195-9_9

[7] "Rdfa 1.1 primer - third edition: Rich structured data markup for web documents," March 2015, accessed: December 24, 2016. [Online]. Available: http://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/

[8] W. Goddard and H. C. Swart, "Distances between graphs under edge operations," *Discrete Math.*, vol. 161, no. 1-3, pp. 121–132, Dec. 1996. [Online]. Available: http://dx.doi.org/10.1016/0012-365X(95)00073-6

[9] T. Lebo, S. Sahoo, and D. McGuinness, "Prov-o: The prov ontology," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/REC-prov-o-20130430/

[10] C. Tilmes, Y. Yesha, and M. Halem, "Distinguishing provenance equivalence of earth science data," *Procedia Computer Science*, vol. 4, pp. 548 – 557, 2011.

[Online]. Available:
http://www.sciencedirect.com/science/article/pii/S1877050911001153

[11] W. F. Tichy, "Rcsa system for version control," *Software: Practice and Experience*, vol. 15, no. 7, pp. 637–654, 1985.

[12] B. Barkstrom, *Earth Science Data Management Handbook: Users and User Access*. CRC Press, April 2014, vol. 1. [Online]. Available:
https://books.google.com/books?id=pI3rTgEACAAJ

[13] S. Burrows, "A review of electronic journal acquisition, management, and use in health sciences libraries," *Journal of the Medical Library Association*, vol. 94, no. 1, pp. 67–74, 01 2006, copyright - Copyright Medical Library Association Jan 2006; Document feature - Graphs; Tables; ; Last updated - 2016-11-09. [Online]. Available:
http://search.proquest.com/docview/203517273?accountid=28525

[14] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum, "A time machine for text search," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '07. New York, NY, USA: ACM, 2007, pp. 519–526. [Online]. Available: http://doi.acm.org/10.1145/1277741.1277831

[15] S. Lyons, "Persistent identification of electronic documents and the future of footnotes," *Law Library Journal*, vol. 97, pp. 681–694, 2005.

[16] B. R. Barkstrom, T. H. Hinke, S. Gavali, W. Smith, W. J. Seufzer, C. Hu, and D. E. Cordner, "Distributed generation of nasa earth science data products," *Journal of Grid Computing*, vol. 1, no. 2, pp. 101–116, 2003. [Online]. Available: http://dx.doi.org/10.1023/B:GRID.0000024069.33399.ee

[17] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark, "Pav ontology: provenance, authoring and versioning," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 37, 2013. [Online]. Available:
http://dx.doi.org/10.1186/2041-1480-4-37

[18] S.-Y. Chien, V. J. Tsotras, and C. Zaniolo, "Version management of xml documents," in *Selected Papers from the Third International Workshop WebDB 2000 on The World Wide Web and Databases*. London, UK, UK: Springer-Verlag, 2001, pp. 184–200. [Online]. Available:
http://dl.acm.org/citation.cfm?id=646544.696357

[19] M. Macduff, B. Lee, and S. Beus, "Versioning complex data," in *2014 IEEE International Congress on Big Data*, June 2014, pp. 788–791.

[20] A. Stuckenholz, "Component evolution and versioning state of the art," *SIGSOFT Softw. Eng. Notes*, vol. 30, no. 1, pp. 7–, Jan. 2005. [Online]. Available: http://doi.acm.org/10.1145/1039174.1039197

[21] "Common questions: Ubuntu release and version numbers," Canonical Ltd., accessed: December 12, 2016. [Online]. Available: https://help.ubuntu.com/community/CommonQuestions##Ubuntu%20Releases%20and%20Version%20Numbers

[22] J. Dijkstra, *On complex objects and versioning in complex environments*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 13–23. [Online]. Available: http://dx.doi.org/10.1007/BFb0024353

[23] P. Cederqvist, R. Pesch *et al.*, *Version management with CVS*. Network Theory Ltd., 2002.

[24] S. Payette and T. Staples, *The Mellon Fedora Project Digital Library Architecture Meets XML and Web Services*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 406–421. [Online]. Available: http://dx.doi.org/10.1007/3-540-45747-X_30

[25] K. S. Baker and L. Yarmey, "Data stewardship: Environmental data curation and a web-of-repositories," *The International Journal of Data Curation*, vol. 4, no. 2, pp. 12–27, 2009.

[26] J. Kovse and T. Härder, "V-grid-a versioning services framework for the grid," in *Berliner XML Tage*, 2003.

[27] K. Holtman, "CMS Data Grid System Overview and Requirements," CERN, Geneva, Tech. Rep. CMS-NOTE-2001-037, Jul 2001. [Online]. Available: http://cds.cern.ch/record/687353

[28] R. Rantzau, C. Constantinescu, U. Heinkel, and H. Meinecke, "Champagne: Data change propagation for heterogeneous information systems," in *In: Proceedings of the International Conference on Very Large Databases (VLDB), Demonstration Paper, Hong Kong*, 2002.

[29] B. Tagger, "A literature review for the problem of biological data versioning," Online, July 2005. [Online]. Available: http://www0.cs.ucl.ac.uk/staff/btagger/LitReview.pdf

[30] U. K. Wiil and D. L. Hicks, "Requirements for development of hypermedia technology for a digital library supporting scholarly work," in *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 2*, ser. SAC '00. New York, NY, USA: ACM, 2000, pp. 607–609. [Online]. Available: http://doi.acm.org/10.1145/338407.338517

[31] D. Dai, Y. Chen, D. Kimpe, and R. Ross, "Provenance-based object storage prediction scheme for scientific big data applications," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 271–280.

[32] P. Vassiliadis, M. Bouzeghoub, and C. Quix, *Towards Quality-Oriented Data Warehouse Usage and Evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 164–179. [Online]. Available: http://dx.doi.org/10.1007/3-540-48738-7_13

[33] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: A survey," *ACM Comput. Surv.*, vol. 37, no. 1, pp. 1–28, Mar. 2005. [Online]. Available: http://doi.acm.org/10.1145/1057977.1057978

[34] M. Fischer, M. Pinzger, and H. Gall, "Populating a release history database from version control and bug tracking systems," in *Proceedings of the International Conference on Software Maintenance*, ser. ICSM '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 23–32. [Online]. Available: http://dl.acm.org/citation.cfm?id=942800.943568

[35] R. Cavanaugh, G. Graham, and M. Wilde, "Satisfying the tax collector: Using data provenance as a way to audit data analyses in high energy physics," in *Workshop on Data Lineage and Provenance*, Oct. 2002.

[36] P. P. da Silva, D. L. McGuinness, and R. Fikes, "A proof markup language for semantic web services," *Information Systems*, vol. 31, no. 45, pp. 381 – 395, 2006, the Semantic Web and Web Services. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306437905000281

[37] M. Bouzeghoub and V. Peralta, "A framework for analysis of data freshness," in *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, ser. IQIS '04. New York, NY, USA: ACM, 2004, pp. 59–67. [Online]. Available: http://doi.acm.org/10.1145/1012453.1012464

[38] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, "The open provenance model: An overview," in *International Provenance and Annotation Workshop*. Springer, 2008, pp. 323–326.

[39] Y. Liu, J. Futrelle, J. Myers, A. Rodriguez, and R. Kooper, "A provenance-aware virtual sensor system using the open provenance model," in *2010 International Symposium on Collaborative Technologies and Systems*, May 2010, pp. 330–339.

[40] Y. L. Simmhan, B. Plale, and D. Gannon, "Karma2: Provenance management for data-driven workflows," *Web Services Research for Emerging Applications: Discoveries and Trends: Discoveries and Trends*, p. 317, 2010.

[41] F. Casati, S. Ceri, B. Pernici, and G. Pozzi, *Workflow evolution*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 438–455. [Online]. Available: http://dx.doi.org/10.1007/BFb0019939

[42] Y. Gil and S. Miles, *PROV Model Primer*, W3C Working Group, Apr. 2013, 30. [Online]. Available: https://www.w3.org/TR/prov-primer

[43] ——, "Prov model primer," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

[44] P. Groth and L. Moreau, "Prov-overview," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

[45] L. Moreau and P. Missier, "Prov-dm: The prov data model," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/REC-prov-dm-20130430/

[46] T. D. Nies, "Constraints of the prov data model," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/REC-prov-constraints-20130430/

[47] T. D. Nies and S. Coppens, "Prov-dictionary: Modeling provenance for dictionary data structures," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/

[48] H. Hua, C. Tilmes, and S. Zednik, "Prov-xml: The prov xml schema," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-xml-20130430/

[49] G. Klyne and P. Groth, "Prov-aq: Provenance access and query," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-aq-20130430/

[50] L. Moreau and P. Missier, "Prov-n: The provenance notation," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/REC-prov-n-20130430/

[51] "Ssemantic of the prov data model," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-sem-20130430/

[52] S. Miles, C. M. Trim, and M. Panzer, "Dublin core to prov mapping," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-dc-20130430/

[53] "Linking across provenance bundles," April 2013, accessed: December 17, 2016. [Online]. Available: https://www.w3.org/TR/2013/NOTE-prov-links-20130430/

[54] I. Suriarachchi, Q. G. Zhou, and B. Plale, "Komadu: A capture and visualization system for scientific data provenance," *Journal of Open Research Software*, vol. 3, no. 1, mar 2015. [Online]. Available: http://dx.doi.org/10.5334/jors.bq

[55] X. Ma, J. G. Zheng, J. C. Goldstein, S. Zednik, L. Fu, B. Duggan, S. M. Aulenbach, P. West, C. Tilmes, and P. Fox, "Ontology engineering in provenance enablement for the national climate assessment," *Environmental Modelling & Software*, vol. 61, pp. 191 – 205, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364815214002254

[56] C. Tilmes, P. Fox, X. Ma, D. L. McGuinness, A. P. Privette, A. Smith, A. Waple, S. Zednik, and J. G. Zheng, *Provenance Representation in the Global Change Information System (GCIS)*, ser. Lecture Notes in Computer Science. Santa Barbara, CA, USA: Springer Berlin Heidelberg, June 2012, vol. 7525, ch. Provenance and Annotation of Data and Processes, pp. 246–248. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34222-6_28

[57] X. Ma, P. Fox, C. Tilmes, K. Jacobs, and A. Waple, "Capturing provenance of global change information," *Nature Clim. Change*, vol. 4, no. 6, pp. 409–413, Jun 2014, commentary. [Online]. Available: http://dx.doi.org/10.1038/nclimate2141

[58] A. Capiluppi, P. Lago, and M. Morisio, "Evidences in the evolution of os projects through changelog analyses," in *Taking Stock of the Bazaar: Proceedings of the 3rd Workshop on Open Source Software Engineering*, J. Feller, B. Fitzgerald, S. Hissam, and K. Lakhani, Eds., May 2003, citation: Capiluppi, A., Lago, P., Morisio, M. (2003). ?Evidences in the evolution of OS projects through Changelog Analyses.? in Feller, P., Fitzgerald, B., Hissam, B. Lakhani, K. (eds.) Taking Stock of the Bazaar: Proceedings of the 3rd Workshop on Open Source Software Engineering ICSE?03 International Conference on Software Engineering Portland, Oregon May 3-11, 2003. pp.19-24.. [Online]. Available: http://roar.uel.ac.uk/1037/

[59] D. German, "Automating the measurement of open source projects," in *In Proceedings of the 3rd Workshop on Open Source Software Engineering*, 2003, pp. 63–67.

[60] K. Chen, S. R. Schach, L. Yu, J. Offutt, and G. Z. Heller, "Open-source change logs," *Empirical Softw. Engg.*, vol. 9, no. 3, pp. 197–210, Sep. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:EMSE.0000027779.70556.d0

[61] K. Herzig and A. Zeller, "Mining cause-effect-chains from version histories," in *2011 IEEE 22nd International Symposium on Software Reliability Engineering*, Nov 2011, pp. 60–69.

[62] "Rdfa core 1.1 - third edition: Syntax and processing rules for embedding rdf through attributes," March 2015, accessed: December 24, 2016. [Online]. Available: http://www.w3.org/TR/2015/REC-rdfa-core-20150317/

[63] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, and J. Völker, *Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 17–32. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-41338-4_2

[64] E. Ainy, P. Bourhis, S. B. Davidson, D. Deutch, and T. Milo, "Approximated summarization of data provenance," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '15. New York, NY, USA: ACM, 2015, pp. 483–492. [Online]. Available: http://doi.acm.org/10.1145/2806416.2806429

[65] B. Cao, Y. Li, and J. Yin, "Measuring similarity between gaphs based on the levenshtein distance," *Applied Mathematics & Information Sciences*, vol. 7, no. 1L, pp. 169–175, 2013.

[66] X. Gao, B. Xiao, D. Tao, and X. Li, "A survey of graph edit distance," *Pattern Analysis and Applications*, vol. 13, no. 1, pp. 113–129, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10044-008-0141-y

[67] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," in *Intern. Journal on Semantic Web and Information Systems (IJSWIS), 3(3):5573, July/Sept. 2006. Special Issue of Multimedia Semantics*, 2006.

[68] Y. Ma, M. Shi, and J. Wei, "Cost and accuracy aware scientific workflow retrieval based on distance measure," *Information Sciences*, vol. 314, no. C, pp. 1–13, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2015.03.055

[69] W. C. Tan, "Research problems in data provenance." *IEEE Data Eng. Bull.*, vol. 27, no. 4, pp. 45–52, 2004.

[70] M. D. Flouris, "Clotho: Transparent data versioning at the block i/o level," in *In Proceedings of the 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST 2004)*, 2004, pp. 315–328.

[71] M. Klein and D. Fensel, "Ontology versioning on the semantic web," in *Stanford University*, 2001, pp. 75–91.

[72] M. S. Mayernik, T. DiLauro, R. Duerr, E. Metsger, A. E. Thessen, and G. S. Choudhury, "Data conservancy provenance, context, and lineage services: Key components for data preservation and curation," *Data Science Journal*, vol. 12, pp. 158–171, 2013.

[73] (2012, Jun.) Dcmi metadata terms. DCMI Usage Board. Accessed: February 8, 2017. [Online]. Available: http://dublincore.org/documents/2012/06/14/dcmi-terms/

[74] T. Stevens, "Nasa gcmd kkeyword version 8.4 released," Aug. 2016, accessed: February 10, 2017. [Online]. Available: https://wiki.earthdata.nasa.gov/display/CMR/NASA+GCMD+Keywords+Version+8.4+Released

[75] B. Polyak, E. Prasolov, I. Tolstikhin, L. Yakovlev, A. Ioffe, O. Kikvadze, O. Vereina, and M. Vetrina, "Noble gas isotope abundances in terrestrial fluids," 2015. [Online]. Available: https://info.deepcarbon.net/vivo/display/n6225

[76] S. Morrison, R. Downs, J. Golden, A. Pires, P. Fox, X. Ma, S. Zednik, A. Eleish, A. Prabhu, D. Hummer, C. Liu, M. Meyer, J. Ralph, G. Hystad, and R. Hazen, "Exploiting mineral data: applications to the diversity, distribution, and social networks of copper mineral," in *AGU Fall Meeting*, 2016.

[77] J. F. Roddick, "A model for schema versioning in temporal database systems," *Australian Computer Science Communications*, vol. 18, pp. 446–452, 1996.

[78] P. Klahold, G. Schlageter, and W. Wilkes, "A general model for version management in databases," in *Proceedings of the 12th International Conference on Very Large Data Bases*, ser. VLDB '86. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1986, pp. 319–327. [Online]. Available: http://dl.acm.org/citation.cfm?id=645913.671314

[79] S. Pröll and A. Rauber, "Citable by design - A model for making data in dynamic environments citable," in *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications, Reykjavík, Iceland, 29 - 31 July, 2013*, 2013, pp. 206–210. [Online]. Available: http://dx.doi.org/10.5220/0004589102060210

[80] M. Helfert, C. Francalanci, and J. Filipe, Eds., *DATA 2013 - Proceedings of the 2nd International Conference on Data Technologies and Applications, Reykjavík, Iceland, 29 - 31 July, 2013*. SciTePress, 2013.

[81] S. Proell and A. Rauber, "Scalable data citation in dynamic large databases: Model and reference implementation," in *IEEE International Conference on Big Data 2013 (IEEE BigData 2013)*, 10 2013.

[82] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, "1. the power of databases: The RRUFF project," in *Highlights in Mineralogical Crystallography*, T. Armbruster and R. M. Danisi, Eds. Walter de Gruyter GmbH, 2015, pp. 1–30. [Online]. Available: http://dx.doi.org/10.1515/9783110417104-003

[83] C. Ochs, Y. Perl, J. Geller, M. Haendel, M. Brush, S. Arabandi, and S. Tu, "Summarizing and visualizing structural changes during the evolution of biomedical ontologies using a diff abstraction network," *J. of Biomedical Informatics*, vol. 56, no. C, pp. 127–144, Aug. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2015.05.018

[84] M. Hartung, A. Gro, and E. Rahm, "Contodiff: generation of complex evolution mappings for life science ontologies," *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 15 – 32, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046412000627

[85] "Overview," accessed: February 14, 2017. [Online]. Available: https://nsidc.org/ssiii/

[86] Ssiii ontologies. National Snow & Ice Data Center. Accessed: February 14, 2017. [Online]. Available: https://nsidc.org/ssiii/ontology.html

[87] Z. B. Miled, S. Sikkupparbathyam, O. Bukhres, K. Nagendra, E. Lynch, M. Areal, L. Olsen, C. Gokey, D. Kendig, T. Northcutt, R. Cordova, G. Major, and N. Savage, "Global change master directory: Object-oriented active asynchronous transaction management in a federated environment using data agents," in *Proceedings of the 2001 ACM Symposium on Applied Computing*, ser. SAC '01. New York, NY, USA: ACM, 2001, pp. 207–214. [Online]. Available: http://doi.acm.org/10.1145/372202.372324

[88] "Keyword faq," Earthdata, 2016, accessed: December 12, 2016. [Online]. Available: https://wiki.earthdata.nasa.gov/display/CMR/Keyword+FAQ