Measuring Data Change using VersOn

BENNO LEE, Rensselaer Polytechnic Institute, USA
PETER FOX, Rensseleaer Polytechnic Institute, USA

Dot-decimal identifiers are traditionally used to label versions as well as indicate whether the current versions has a major, minor, or smaller different from the previous version. The method poses a challenge because differences between categories can be compared, but not within categories. The Versioning Ontology (VersOn) captures individual changes between versions in a versioning graph. The changes within the versioning graph can be enumerated, enabling a more precise change metric called change distance. We use the Global Change Master Directory (GCMD) Keywords to test the efficacy of change distance to dot-decimal identifier categories in assessing change. We found that change distance is able to more precisely measure changes between versions, identifying trends in behavior within and across dot-decimal identifier categories. Through analysis of the Global Change Master Directory (GCMD) Keywords Version 8.5, we found that VersOn enables data consumers to assess data change in ways relevant to the consumer and independent of the producer's assessment of change as indicated by the dot-decimal identifier assigned to the version.

CCS Concepts: • **Information systems → Semantic web description languages**; *Web log analysis*; *Resource Description Framework (RDF)*; *Web Ontology Language (OWL)*;

Additional Key Words and Phrases: Version, versioning, versioning ontology, data change, change metric, versioning graph

1 INTRODUCTION

The organization of data sets into versions is often determined by the data traditions of the data managers [2]. A common tradition in versioning practice is the use of dot-decimal identifiers to label versions [7]. In addition to serving as an identifier, the structure broadly categorizes the amount of change between versions as major, minor, or smaller. Depending on the total change, the associated category in the identifier of the previous version is incremented. Dot-decimal identifiers introduce a challenge in assessing data change because labels are applied by the data producer at the time of publication. As a result, the producer has sole authority on the context and amount of change introduced by a new version.

The Versioning Ontology (VersOn) captures the individual changes between versions using linked data [**?**]. VersOn also includes semantics to classify changes into additions, invalidations, and modifications. Because at least one change contributes to a new version, counting the changes provides a more precise metric than the broad categories. We used VersOn to develop a change distance which is more precise than the dot-decimal categorizations and enables users to contextualize change assessments.

2 PREVIOUS WORK

2.1 Global Change Master Directory

The Global Change Master Directory (GCMD) is a metadata repository used by the National Aeronautics and Space Administration (NASA) to store records of its available data sets [4]. GCMD employs a taxonomy of keywords to make NASA Earth Science data sets searchable. These words tag and label datasets into strictly defined categories [1]. The management team stored early versions of the keywords in Excel spreadsheets, later using a centralized distribution system, but data is not available prior to June 12, 2012. The Key Management Service (KMS) now serves the keywords directly in a variety of formats. Other than the initial version, named 'June122012', versions are given a dot-decimal identifier. According to the *Keyword Governance and Community*

Authors' addresses: Benno Lee, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA, leeb5@rpi.edu; Peter Fox, Rensseleaer Polytechnic Institute, Troy, NY, 12180, USA, pfox@cs.rpi.edu.

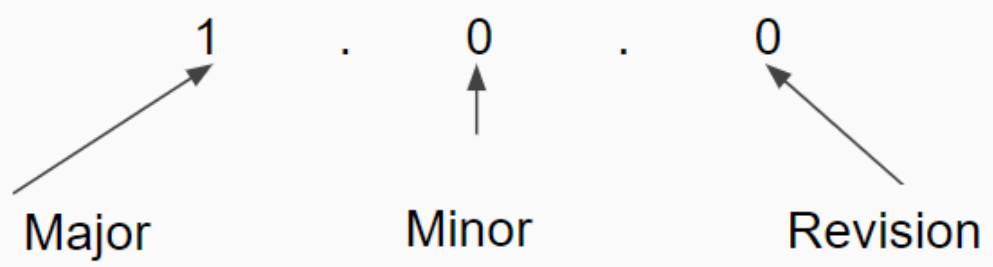

Fig. 1. A dot-decimal identifier is broken into a series of number separated by decimals. The left-most number indicates a grouping over major features. Intervals moving to the right decrease in significance. The third interval often has different names, but we have chosen to refer to the interval as the revision number.

Guide Document, "Full GCMD keywords list releases get a new major version number (e.g., 8.0). Incremental releases for updates to topics, terms, and variables get a new minor version number (e.g., 8.1)," [3].

Each keyword corresponds to a Universally Unique Identifier (UUID), and when combined with a web namespace, resolves to a data description of the keyword. Every identifier can be referred to per version by including the version's number at the web identifier's end, meaning that identifiers are consistent across versions. The namespace, UUID, and version form a versioned keyword's Uniform Resource Identifier. The taxonomy uses the concepts *skos:Broader* and *skos:Narrower*, where skos refers to the Simple Knowledge Organization System (SKOS) ontology name space, to form a tree hierarchy [5]. The tree's root is the keyword, "Science Keywords." The data set provides an interesting study case due its long sequence of versions and ready use of linked-data technology [6].

2.2 Dot-Decimal Identifiers

A dot-decimal identifier is a series of numbers joined together by a decimal point with the left-most number signifying a commonality across major features and the right-most number signifying a commonality across the most minor features, as seen in Figure 1. The amount of change between two versions is determined by finding the left-most interval at which the identifiers differ. Because the identifiers traditionally denote a version's location in a sequence, only one interval is changed at a time and is only incremented by one.

3 VERSIONING GRAPH

A **versioning graph** is a linked-data graph which captures the changes separating one version of a data object from another data object. Utilizing VersOn, the graph between two objects looks like a ladder with the rungs representing the changes affecting the versions as seen in Figure 2. The keywords comprise the VersOn Attributes making up the parts of the taxonomy actually changing, in Figure 2: 'Atlantic Multidecadal Oscillation', 'Isotope Ratios', and 'Cloud Asymmetry'. In order to detect the change and the kind of change, we use the unique keyword identifier assigned by the GCMD Keywords group to each keyword. Since the UUID for each keyword remains the same across versions, the unique keyword identifier can be used to align a mapping across versions. Additions and invalidations are detected by checking an identifier's presence within both versions. GCMD Keywords group provides a distribution encoded using SKOS [5]. The property *skos:Broader* is used to define the parent term of a keyword in the taxonomy, and we can determine a modification to the taxonomy occurs when a keyword's parent changes. A difference indicates that the word has been moved to a different place within the taxonomy since identifiers do not change across versions and a keyword only has one parent concept. The alignment assumes that there is no reason a keyword's preferred label would change, but still reports a value when it has new entries

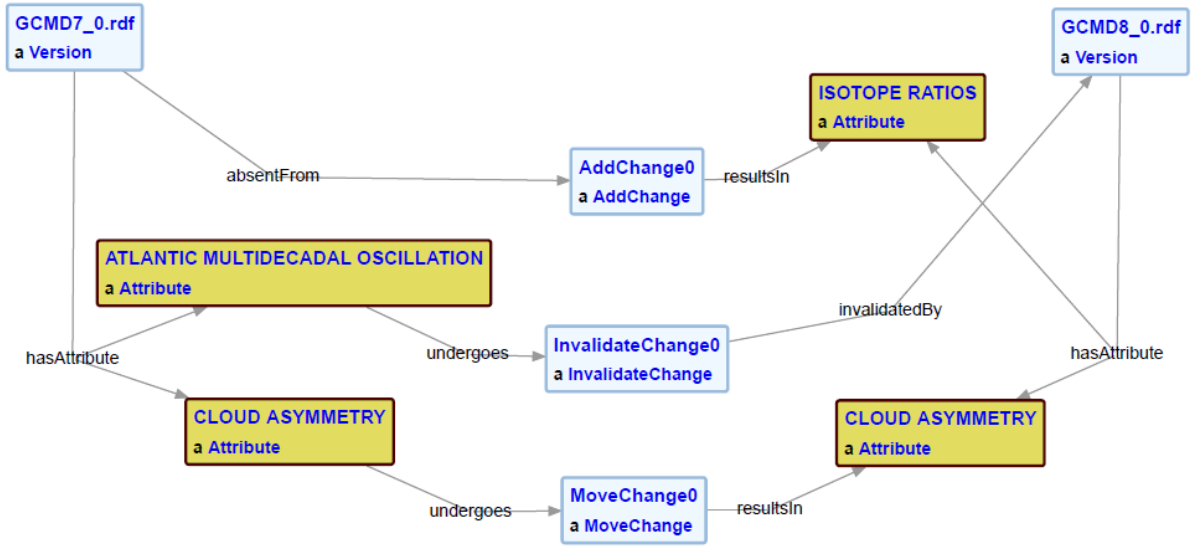

Fig. 2. A sample of the versioning graph between GCMD Versions 7.0 and 8.0, demonstrating an addition, invalidation, and modification.

Table 1. Global Change Master Directory Keyword change counts.

Transition	Add	Invalidate	Modify	Total
June 12, 2012 to 7.0	310	9	22	341
7.0 to 8.0	503	6	79	588
8.0 to 8.1	277	28	22	327
8.1 to 8.2	53	1	26	80
8.2 to 8.3	58	0	13	71
8.3 to 8.4	53	0	1	54
8.4 to 8.4.1	86	13	8	107

in the “notes" property. A change log was generated for each pair of consecutive versions in GCMD Keywords and embedded using JavaScript Object Notation for Linked Data (JSON-LD). Versioning graphs for each adjacent version were created by extracting JSON-LD from the corresponding change log, and entering the triples into a Fuseki triple store.

4 COMPUTING CHANGE COUNTS

We now use the rungs of the versioning graph to quantify the amount of change by counting the number of changes in each change type. Because VersOn has specific semantics for additions, invalidations, and modifications, the metric is also separated into three components. The addition, invalidation, and modification counts for each transition up to Version 8.4.1 are presented in Table 1. Modify changes are labeled as Moves to differentiate between types of modifications when discussing Version 8.5 in Section 4.1. Figure 3 plots the quantities and shows the individual change components as a bar chart and the total change as a line plot. The plot demonstrates that

dot-decimal identifiers show total summarized change as reflected by the total change. Using VersOn, the change metric provides insight into the nature and components of change in the data set, capturing version change with greater precision. From the bar chart, we can also see that additions dominate the kind of change in each interval.

A very clear discrepancy in the order of magnitude can be seen between major and minor version transitions except at the “8.0 to 8.1" transition. The publication of Version 7.0 and 8.0 both result in multi-hundred additions of concepts to the taxonomy. Versions 8.2 to 8.4.1 only produced between 50 and 90 additions, an order of magnitude lower than major version changes. There only exists a 33 addition difference separating changes induced by Version 8.1 and the changes created by Version 7.0. Assuming the dot-decimal identifier assignment is correct, the quantitative cutoff between full and incremental releases lies around 300 addition changes. The additions in “8.4 to 8.4.1" in Figure 3 numbers almost a hundred, providing evidence that the trend of decreasing order of magnitudes may not continue as the granularity of the version identifier increases.

4.1 GCMD Version 8.5

While applying the URI based mapping method described in Section 3 to Version 8.5, we collected a very alarming result. In Table 2, ‘URI Based’ shows the number of additions and invalidations at about the size of the entire GCMD Keywords data set. In addition, no modifications are revealed, and even the root node "Science Keywords" has been invalidated, meaning the GCMD Keywords group had removed and replaced the entire taxonomy.

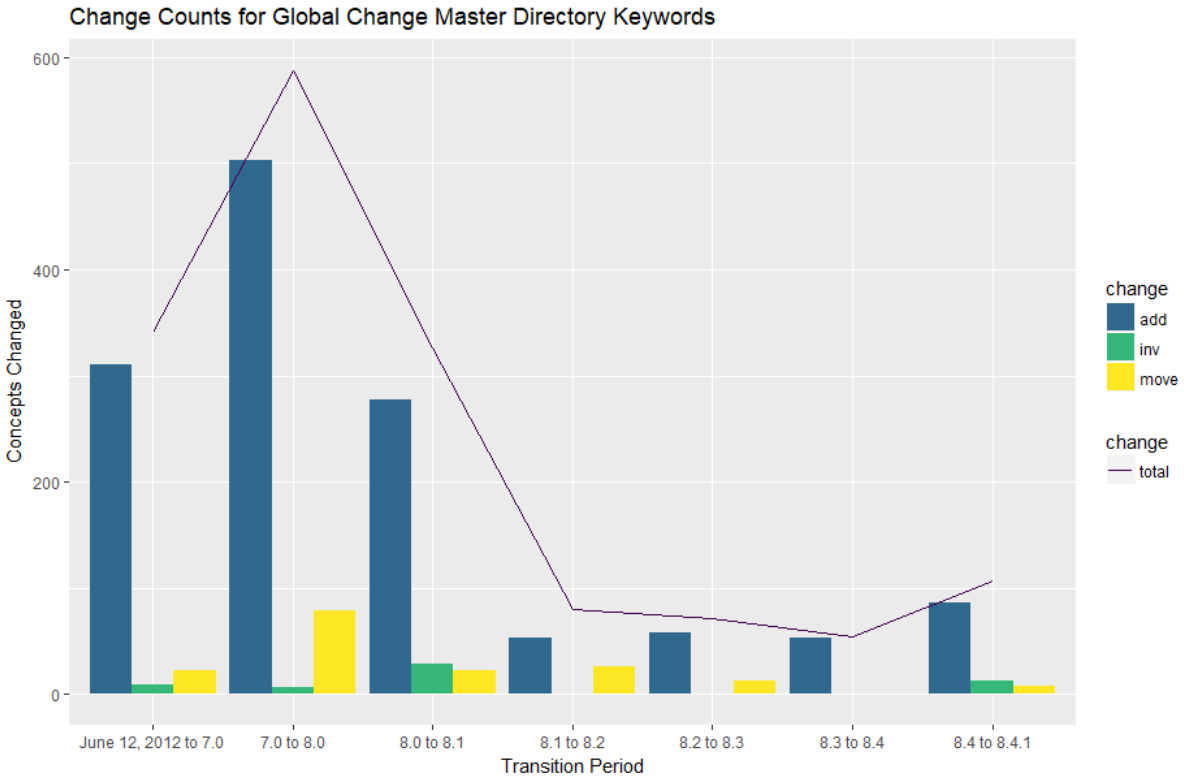

Fig. 3. Add, Invalidate, and Modify counts from the beginning of the Keyword Management System to Version 8.4.1.

Table 2. Difference in Version 8.5 mapping methods.

Mapping Method	Add	Invalidate	Move	Modify
URI Based	3097	3031	0	0
Bridged	68	2	22	3007
Silent	68	2	22	0

Further investigation of the root word reveals that the name space for the keywords has changed from `http` to `https`. To provide context, National Aeronautics and Space Administration (NASA) mandated a transition to secure protocols, and the group changed the name space to ensure the URIs remained resolvable. Since the URIs are lexically unique, the new name space means new URIs no longer refer to the same object after the protocol change. Because the keyword identifiers no longer match, the mapping approach results in the total invalidation of keywords from 8.4.1 and the addition of keywords from 8.5. The dot-decimal identifier for the transition from Version 8.4.1 to 8.5 does not match the number of changes in the versioning graph.

Changing the mapping method to account for the new namespace provides a pathway to compare the perceived change by the producer as evidenced by the version identifier with the change distance in the versioning graph. To do this, the mapping treats identifiers with `http` and `https` the same, revealing keywords actually added, invalidated, and moved. The code breaks apart the UUID from the URI and uses the prior namespace or current namespace, respectively. Differences in change distance become much clearer after controlling for the altered name space in Figure 4 under the label 'Bridged'. The Silent method ignores the change in namespace by considering concepts with matching UUIDs as unchanged. Only the Silent method follows the previous minor release trend of being addition dominated and having less than a hundred additions.

5 CHANGE METRIC ANALYSIS

The dot-decimal identifier scheme employed by GCMD Keywords defines two levels of change: major and minor. According to the *Keyword Governance and Community Guide Document* [3], "Full GCMD keywords list releases get a new major version number (e.g., 8.0). Incremental releases for updates to topics, terms, and variables get a new minor version number (e.g., 8.1)," [3]. The document does not explain the purpose or distinguishing qualities of versions with a third level identifier i.e. Version 8.4.1. VersOn improves the change evaluation between versions by increasing the precision used to distinguish versions from a categorical metric to a spectrum. Rather than a major or minor change, the metric accounts the number of changes between adjacent versions. Using VersOn additionally breaks down the total summary metric into component parts: addition, invalidation, and modification. The breakdown in Figure 3 illustrates that additions dominate the kind of change made to versions in GCMD, a conclusion unreachable looking at either total change or the dot-decimal identifier.

We originally pursued the hypothesis that the VersOn change metric would improve the ability to discern between major and minor GCMD Keyword versions on the assumption that an order of magnitude difference exists in the total changes between major and minor transitions. The results in Figure 3 show that '8.0 to 8.1' is not an order of magnitude different from major version publications and '8.4 to 8.4.1' is also not an order of magnitude separated from minor version releases. Looking at the specific values of 'June 12, 2012' to Version 8.1, there may exist a threshold around 330 total changes where data producers consider a new version to be a full keyword release. A similar threshold from minor to revision does not seem to exist, and the total number of changes in the revision is actually greater than most other minor releases in the data. The analysis does not claim that change distance should be the sole mechanism in determining version identifiers. Addition, invalidation, and

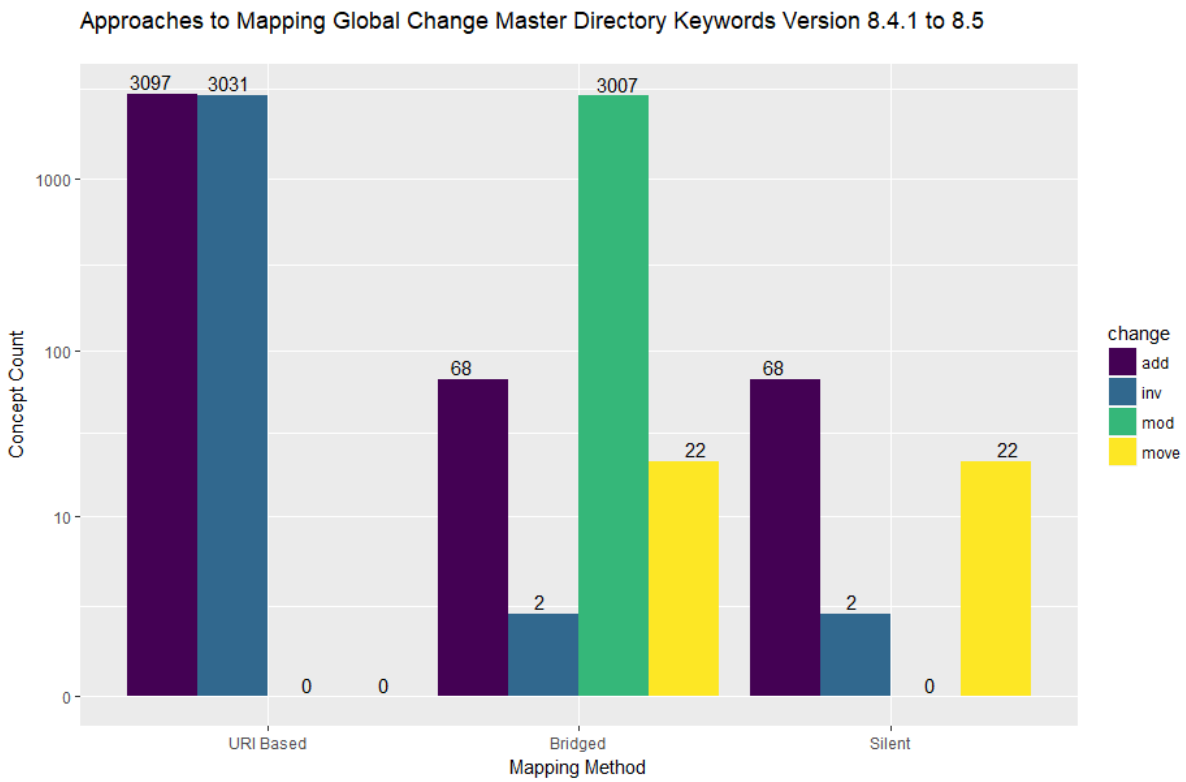

Fig. 4. Add, Invalidate, and Modify counts using different methods of mapping identifiers in Global Change Master Directory Keywords Version 8.4.1 to 8.5.

modification provides deeper insight into how a data set is changing, but some changes can be more impactful than others which this analysis does not capture.

The results of applying VersOn to Version 8.5 revealed multiple methods in assessing the amount of change in a data set. Comparing the 'URI Based' and 'Silent' mapping methods, we can see the discrepancy between a linked-data user's perceived change and the GCMD Keyword group's perceived change. We know that GCMD Keyword group uses the 'Silent' method since the method is the only one to produce a magnitude of change not on the order of the entire data set. The 'URI Based' and 'Bridged' methods require new major version numbers, but a minor version number was assigned. The metric also models informed GCMD Keywords Version 8.5 customers who were aware of the namespace change by manually mapping matching UUIDs as modifications. The results demonstrate the inflexibility of dot-decimal identifiers which must be applied at the time of publication entirely based on the data producers assessment of the total change to the data set. VersOn enables data users to contextualize the data by the data's utilization and compute an appropriate change metric post-publication. Making sure that data consumers have the ability to assess change in data sets when the requirements for change differs between producer and consumer must be addressed.

6 CONCLUSION

Dot-decimal identifiers are labels often used to indicate the categorical amount of change a new version introduces to a data set. The change distance computed using VersOn provides a more regular, precise method to evaluate change using standardized semantics by separating change into types and enumerating individual changes. The counts for GCMD Keywords were acquired by querying each transition's versioning graph. The change distance did not reveal an order of magnitude separation between major and minor versions, but suggested a threshold indicating a full new release. The change metric also showed that GCMD Keywords is a data set which generally adds data with each version release. The analysis of Version 8.5 highlights the dynamic between data producer and consumer roles. Using URI best-practice or GCMD Keywords consumer practice, the number of changes amounted to a new full keyword release. Because the data producer is the authority on version labeling, a minor version number was assigned based on different metrics. The VersOn based change distance enables consumer contextualized assessments based on utilization.

REFERENCES

[1] 2016. Keyword FAQ. https://wiki.earthdata.nasa.gov/display/CMR/Keyword+FAQ Accessed on: December 12, 2016.

[2] B.R. Barkstrom. 2014. *Earth Science Data Management Handbook: Users and User Access*. Vol. 1. CRC Press, Boca Raton, FL, USA. https://books.google.com/books?id=pI3rTgEACAAJ Accessed on: July 12, 2014.

[3] Global Change Master Directory (GCMD). 2016. *Global Change Master Directory (GCMD) Keyword Governance and Community Guide Document* (version 1.0 ed.). National Aeronautics and Space Administration (NASA). https://cdn.earthdata.nasa.gov/conduit/upload/5182/KeywordsCommunityGuide_Baseline_v1_SIGNED_FINAL.pdf Accessed: June 10, 2018.

[4] Zina Ben Miled, Srinivasan Sikkupparbathyam, Omran Bukhres, Kishan Nagendra, Eric Lynch, Marcelo Areal, Lola Olsen, Chris Gokey, David Kendig, Tom Northcutt, Rosy Cordova, Gene Major, and Nanine Savage. [n. d.]. Global Change Master Directory: Object-oriented Active Asynchronous Transaction Management in a Federated Environment Using Data Agents. In *Proc. of the 2001 ACM Symp on Applied Computing*. 207–214. https://doi.org/10.1145/372202.372324

[5] Alistair Miles and Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. W3C. http://www.w3.org/TR/2009/REC-skos-reference-20090818/ Accessed on: July 12, 2018.

[6] Tyler Stevens. 2016. NASA GCMD Keyword Version 8.4 Released. https://wiki.earthdata.nasa.gov/display/CMR/NASA+GCMD+Keywords+Version+8.4+Released Accessed on: February 10, 2017.

[7] Ben Tagger. 2005. A Literature Review for the Problem of Biological Data Versioning. Online. https://doi.org/10.1.1.104.2137