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

APPENDICES

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

ABSTRACT

This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text.

This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text.

This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text.

This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text. This is a sentence used to take up space and look like text.

CHAPTER 1
INTRODUCTION

Dataset versioning is the process by which changes to data are tracked and documented. Version numbers are commonly seen in relation to new releases of software such as MATLAB or R, but current version naming schemes borrow from software contexts. While it may have been possible to manually manage the changes to data when datasets were relatively small, research centers must now supervise on the order of tens of millions of files with changes being made at rates of thousands of processing jobs per day.

Versioning information is widespread across devices and software in the modern world. From the numbering of the latest smart phone to the patch number of the newest release of MATLAB, scientists must deal with a range of labels and formats when performing their research. Science data, likewise, also has a tendancy to change. Datasets are subjected to data audits and error corrections regularly to maintain a level of data quality.

Agencies and research groups have collected new data at an incredible rate. The amount of data housed by NASA quadrupled from 2001 to 2004 [4] and high energy physics labs can generate on the order of 4000 new datasets every day [5].

1.1 Provenance

In a number of papers, authors describe models or systems that track changes in the workflow and how such changes would be reflected in new datasets. Barkstrom outlines the three tiered system employed by NASA in their remote sensing data workflows.This would be data provenance and semantic tools already exist to maintain this information. The PROV Data Model allows systems to encode provenance information into RDF. However, PROV expresses new versions with the wasRevisionOf or alternateOf property. should not be conflated with data versioning. While understanding the lineage of a dataset may reveal how changes propagate down

Provenance is the data used to describe the origin of an object (Merrium

Figure 1.1: This is the Caption for Figure 1 make it long to illustrate how it looks when wrapped around to the next line

Webster). While data provenance and data versioning are related, the priorities of one should not be confused with the other. Understanding the origins of a dataset and the activities leading to its creation may inform why two data objects are different, the changes made do not take priority. Many different views of versioning has appeared in the literature. When Barkstrom and various other experts refer to data versioning, they are actually referring to data provenance.

1.2 Changelogs

1.3 RDFa

1.4 Why Semantic Technologies?

This is a sentence to take up space [1]. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

Please refer to Figure 1.1.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

Table 1.1: This is the Caption for Table 1[1]

Here's	an	example
of	a	table
floated	with	the
`table`	environment	command.

1.5 This is a Section Heading

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

1.5.1 This is a Subsection Heading

This is a sentence to take up space and look like text. This is a sentence to take up space [2]. This is a sentence to take up space and look like text.

1.5.1.1 This is a Subsubsection Heading

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. Text before the footnote.[1] Text after the footnote. This is a sentence to take up space and look like text.

[1] Here's the text of the footnote.

CHAPTER 2
PREVIOUS WORK

PROV is a W3C recommendation that deliniates a method to express data provenance with semantic technologies. Using the model of relating activities, agents, and entities, data managers can express the origins of their datasets. However, when an entity is revised, the PROV data model can only express the relationship as a revision or that the new dataset was derived from the original. This leaves

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is shown in table 2.1.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

Figure 2.1: This is the Caption for the First Figure in Chapter 2. It is a long, long caption; we do not want to put the whole thing in the List of Figures. A Shorter Caption can go in the square brackets.

Table 2.1: This is the Caption for Table 2

Here's	another	example
of	a	table
floated	with	the
`table`	environment	command.

2.1 This is a Section Heading

This is a sentence to take up space and look like text. This is a sentence to take up space [3]. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

2.1.1 This is a Subsection Heading

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. Text before a footnote.[2] Text after the footnote.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. Text before another footnote.[3] Text after the footnote. This is a sentence to take up space and look like text.

[2] Here's the text of the footnote.

[3] Here's the text of the footnote.

CHAPTER 3
CONCEPTUAL MODEL

The conceptual model used within this thesis is built around the expression of three core versioning operations: addition, invalidation, and modification. These three activities can be represented by interacting with three types of concepts: versions, attributes, and changes. Versions represent the data entities being compared. These could be two different editions of a book or versions of software. It is important to understand that a version is an abstraction as it can be represented by multiple physical files. In the sections that follow, operations will only consider the interaction between two versions and will be explained later in the chapter. Versions then contain attributes representing a quantity being modified. Specifically for tabular data, attributes would correspond to an identifier that refers to particular rows or columns within the data. Attributes of the two versions are then connected by a change. This link functions as a very general concept which can be subclassed into more informative types such as unit changes, improving the expressivity of the model beyond PROV's revisionOf concept.

3.1 ADDITION

A difficulty with comparing provenance graphs is that two data objects can have identical structures, but be added to the dataset

3.2 INVALIDATION

3.3 MODIFICATION

REFERENCES

[1] This is the first item in the Bibliography. Let's make it very long so it takes more than one line. Let's make it very long so it takes more than one line. Let's make it very long so it takes more than one line. Let's make it very long so it takes more than one line.

[2] The second item in the Bibliography.

[3] Another item in the Bibliography.

[4]

[5]

APPENDIX A
THIS IS AN APPENDIX

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

A.1 A Section Heading

This is how equations are numbered in an appendix:

$$x^2 + y^2 = z^2 \tag{A.1}$$

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.

APPENDIX B
THIS IS ANOTHER APPENDIX

This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text. This is a sentence to take up space and look like text.