Some guidelines for Conceptual modeling to help the data repurposability

Sanaz Nabavian[1] [0000-0003-1025-6870]

[1] Memorial University of Newfoundland and Labrador, NL, Canada

snabavian@mun.ca

Abstract. A quintillion bytes of data are created every day. Reusing the collected data for different purposes is a better option in many cases than gathering new data. However, preparing existing data to match the requirements of new uses can be difficult. This research aims to give some guidelines for designing a dataset which is more repurposable. As conceptual modeling is the heart of designing an Information System, I will focus on how defining the self-defining concepts could help datasets to be reused in other concepts and to improve the datasets connect ability to other datasets.

Keywords: data repurposing, conceptual model.

1 Introduction

With the growing importance of data-driven decision-making, organizations are thirsty for data. This has led to widespread repurposing data - taking data collected with one set of uses in mind and adapting it to answer questions not anticipated when the requirements and conceptual model for the original uses were determined.

Repurposing data is an important practice because of the sheer volume and diversity of data that exists and that can be adapted to new uses. Repurposing data saves the time, effort, and money that would otherwise be required to collect data from scratch.

Repurposability of data is its amenability to repurposing based on the features it currently has. These features are the result of design decisions made that guide different aspects of the data collection process, and the data architecture (e.g., type of database like relational, NoSQL, Graph). For example, suppose a data set has a very common structure for representing an address: Country, province, and city. Some cities located in Armenia in October 2020 became a part of Azerbaijan from November 2020, according to the 2020 Nagorno-Karabakh ceasefire agreement. This happens for provinces in the same countries as well. This example illustrates how we can rethink the role of conceptual modeling in supporting (or impeding) data repurposing.

Conceptual modeling is a vital step in determining data requirements for an application. In this step, designers identify the entities and their relationships relevant to some intended uses by examining the (domain)'s instances, rules, and environment. The domain here is the part of the world in which the information system is performing. A conceptual schema (model) should describe all relevant general static and dynamic aspects, i.e. all rules, laws, etc., of its defined domain (Olive, 2007). According to Olive, a conceptual model would not be

considered complete and correct if it does not represent all relevant domain knowledge. As you can see, the compliance of the conceptual model with the domain knowledge is very high in projects that have a well-defined set of uses. However, in developing a repurposable dataset, it is necessary to consider extending the domain of the dataset life cycle for more than one project or one organization (another domain).

The concept of the data could be changed when the domain is changing. A person can be a student, a professional worker, and a parent, and each concept has its significant characteristics in each domain. The concept of the data could also be changed when we add or join other domains to the main domain. For example, if a school system wants to include their professional information as well.

On the other hand, the term concept itself is a tricky and interesting phenomenon. According to Murphy, concepts are fuzzy, and it is sometimes difficult to distinguish between members and nonmembers of a concept, like a chair and a stool (2002). Although, some of the instances are clearly chairs due to the number of legs, their size, and presence of back and arms. However, for a bigger stool with a back, categorization could be more challenging. Also, in some cases observing

more instances can change some of the features of a concept. For example, swans in most areas are white, but black swans are common in southwestern and eastern Australia. Conceptual modeling could be strictly matched to the current situation and observed instances or more general and open to potential new instances. More rigid conceptual modeling makes it hard or impractical to connect the obtained data structure to other datasets because it might not cover all the future instances (e.g., black swans or stools with back), or the new features of a thing in another context (e.g., a student who is a parent as well).

Is it possible to predict all the features of an entity in different domains? If this means defining a universal concept that fits every domains, probably it is not possible or it is not the concern of this paper because of previous unsuccesful attempts to find universal taxonomies. For instance, Leibniz aimed to find a universal language of science by defining the alphabet of human thoughts and representing knowledge with that basic alphabet. His ambition has not fared well onwards (Eco, 2000). Past attempts in the natural sciences turned out to messier than we thought. Cladistis hoped that their system of nomenclature would

yield a taxonomic system without junk categories like the Linnean system has. That project failed much to biologists' surprise.

This research will provide ways to improve repurposability during conceptual modeling so that by connecting to other datasets, the final data would be enriched. It will focus on how conceptual modeling can enhance data repurposability and give insights into how current datasets can be evaluated to assess their repurposability. The research will focus on two solutions: 1) data independence, and 2) adding global features. These solutions help by giving more transferability and connectivity potential to a dataset.

Data independence refers to defining self-determined data that do not contain any types of agreements. For instance, the geographical point has all the information you might need to find the related city, province. Saving self-determined data helps repurposability because it would not need any extra analysis or transaction to prepare the data. Instead, you could join the dataset with other updated datasets or historical datasets.

Another way to ensure that you can reuse the data is by adding global features. Numbers or characteristics that do not have only local/limited meaning in the defined domain. For instance, student id might be the best data for the school domain but not the best identifier for connectivity and reusability. According to the new technologies, we might be able to generate better identifiers for the dataset.

These suggestions might not seem very complicated and unachievable. However, I want to check some of the datasets to evaluate how the rigidity of the conceptual model changes the repurposability of the data.

2 Methodology

For this research, 14 papers have been reviewed so far. The list of the papers is provided in Appendix 1. In each piece, they explained data preparation steps to reuse that data in another context. I categorized some of the issues that were routed in conceptual modeling.

It would be beneficial to provide evidence that the proposed suggestions will improve the repurposability. To test that, comparing two datasets with and without proposed features might be helpful. The symposium could help me in designing and improving the methodology.

3 References

1. Eco, U. (2000). Kant and the platypus: Essays on language and Cognition. Harcourt Brace.
2. Murphy, G. L. (2004). *The big book of concepts*. MIT Press.
3. Olivé A. (2007). *Conceptual modeling of information systems*. Springer.

Appendix 1

Article	Year	data base
Towards Safe Cities: A Mobile and Social Networking Approach	2012	Yelp Data, Crime Data, Census data sets
Twitter-derived neighborhood characteristics associated with obesity and diabetes	2017	Twitter
Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City, 2012–2013	2012	Yelp
On the brink: Predicting business failure with mobile location-based check-ins	2015	Foursquare
Individualism during Crises: Big Data Analytics of Collective Actions and Policy Compliance amid COVID-19	2020	Google Community Mobility Report, GoFundMe
Identifying Soccer Players on Facebook Through Predictive Analytics	2017	Facebook
Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube	2012	YouTube
Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia	2016	Wikipedia
Understanding User-Generated Content and Customer Engagement on Facebook Business Pages	2020	Facebook
Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews	2014	epinions.com
Online Product Opinions: Incidence, Evaluation, and Evolution	2012	Bazaar voice
Content Contributor Management and Network Effects in a UGC Environment	2012	
Towards the Improvement of Topic Priority Assignment Using Various Topic Detection Methods for E-reputation Monitoring on Twitter	2014	Twitter