UNIVERSITY OF AMSTERDAM

INDE lab

INtelligent Data Engineering

indelab.org | @INDE_LAB_AMS

Intro - Dutch-Belgian Database Day December 7, 2021

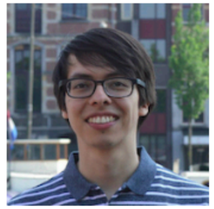
The INDE lab Team


Prof. Paul Groth


Dr. Frank Nack


Dr. Jacobijn Sandberg


Thiviyan Thanapalasingam


Daniel Daza


Madelon Hulsebos

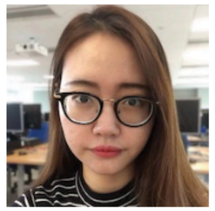

Dr. Sebastian Schelter


Dr. Sara Magliacane


Fan Feng


Corey Harper


Melika Ayoughi


Effy Xue Li


Stian Soiland-Reyes


James Nevin


Stefan Grafberger


Dr. Peter Bloem


Dr. Hartmut Koenitz


Dr. Stefan Schlobach


Valentina Carriero

INDELab started Nov. 5, 2018

Olivier Sprangers

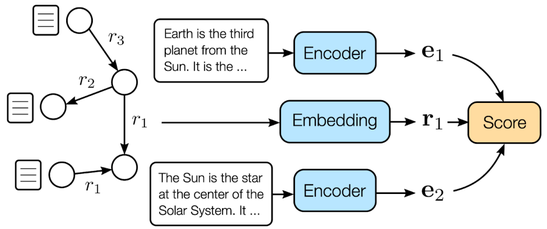
Shubha Guha

Barrie Kersbergen

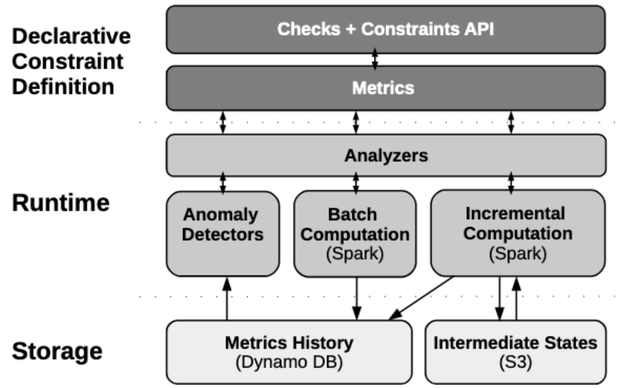
Research Topics at INDE lab

- Design systems to support people in working with data from diverse sources
- Address problems related to the preparation, management, and integration of data

- **Automated Knowledge Graph Construction**
  (e.g., predicting and adding new links in datasets such as Wikidata based on text)
- **Data Search & Reuse**
  (e.g., studies on GitHub hosted data; algorithms for making data FAIR)
- **Data Management for Machine Learning**
  (e.g., scalable concept drift detection for ML training data,
  integrated in AWS SageMaker Model Monitor; using data provenance for ML debugging)
- **Causality-Inspired Machine Learning** (e.g., using ideas from
  causal inference to improve the robustness and generalization
  of ML algorithms, especially in cases of distribution shift; domain adaptation)



Link prediction



Highlights for the DBDBD audience

About GitTables gittables.github.io

GitTables is a dataset of currently 1.7M relational tables extracted from CSV files in GitHub. Our continuing curation aims at growing the dataset to at least 20M tables. Table columns in GitTables have been annotated with more than 2K different semantic types from Schema.org and DBpedia. Our column annotations consist of semantic types, hierarchical relations, range types and descriptions.

The high-level pipeline in Figure 1 illustrates how GitTables was created.

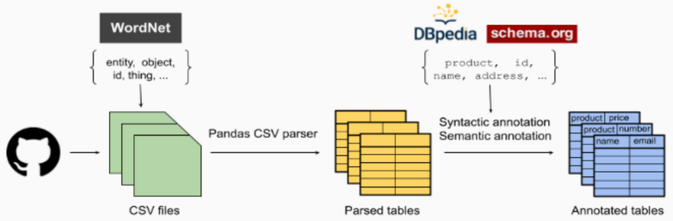
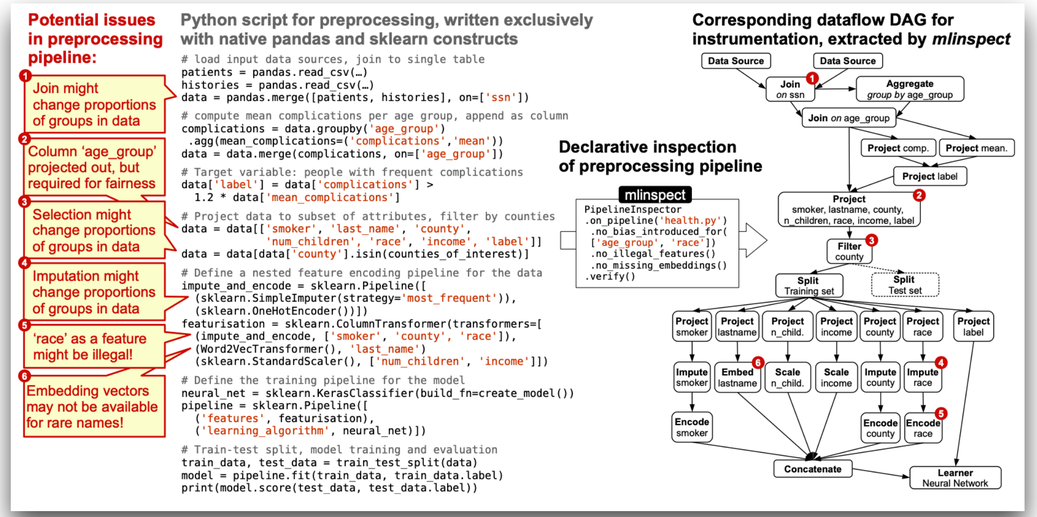


Figure 1: high-level pipeline of the process of constructing GitTables.

mlinspect https://github.com/stefan-grafberger/mlinspect



Analysis of how data reuse and search in the wild

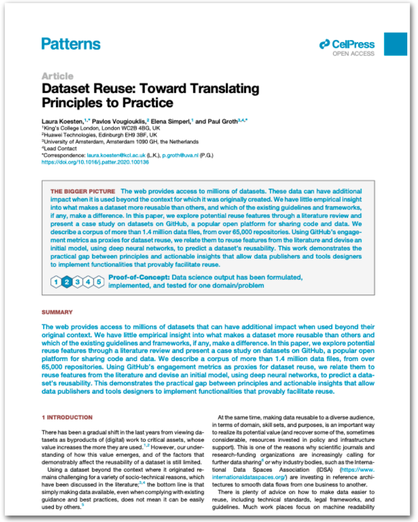
Harvard Data Science Review

Lost or Found? Discovering Data Needed for Research

Kathleen Gregory[1], Paul Groth[2], Andrea Scharnhorst[1], Sally Wyatt[3]

[1]Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences,
[2]Informatics Institute, University of Amsterdam,
[3]Faculty of Arts and Social Sciences, Maastricht University

Published on: Apr 30, 2020
Updated on: May 07, 2020
DOI: 10.1162/99608f92.e38165eb



Patterns

CellPress OPEN ACCESS

Article

Dataset Reuse: Toward Translating Principles to Practice

Laura Koesten,[1,*] Pavlos Vougiouklis,[2] Elena Simperl,[1] and Paul Groth[3,4,*]
[1]King's College London, London WC2B 4BG, UK
[2]Huawei Technologies, Edinburgh EH9 3BF, UK
[3]University of Amsterdam, Amsterdam 1090 GH, the Netherlands
*Lead Contact
*Correspondence: laura.koesten@kcl.ac.uk (L.K.), p.groth@uva.nl (P.G.)
https://doi.org/10.1016/j.patter.2020.100136

THE BIGGER PICTURE The web provides access to millions of datasets. These data can have additional impact when it is used beyond the context for which it was originally created. We have little empirical insight into what makes a dataset more reusable than others, and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This work demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

1 2 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

The web provides access to millions of datasets that can have additional impact when used beyond their original context. We have little empirical insight into what makes a dataset more reusable than others and which of the existing guidelines and frameworks, if any, make a difference. In this paper, we explore potential reuse features through a literature review and present a case study on datasets on GitHub, a popular open platform for sharing code and data. We describe a corpus of more than 1.4 million data files, from over 65,000 repositories. Using GitHub's engagement metrics as proxies for dataset reuse, we relate them to reuse features from the literature and devise an initial model, using deep neural networks, to predict a dataset's reusability. This demonstrates the practical gap between principles and actionable insights that allow data publishers and tools designers to implement functionalities that provably facilitate reuse.

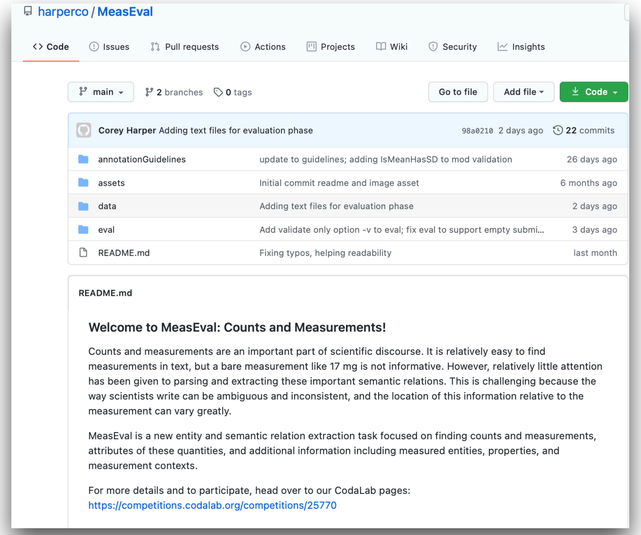
1 INTRODUCTION

There has been a gradual shift in the last years from viewing datasets as byproducts of (digital) work to critical assets, whose value increases the more they are used.[1,2] However, our understanding of how this value emerges, and of the factors that demonstrably affect the reusability of a dataset is still limited.

Using a dataset beyond the context where it originated remains challenging for a variety of socio-technical reasons, which have been discussed in the literature,[3,4] the bottom line is that simply making data available, even when complying with existing guidance and best practices, does not mean it can be easily used by others.[5]

At the same time, making data reusable to a diverse audience, in terms of domain, skill sets, and purposes, is an important way to realize its potential value (and recover some of the, sometimes considerable, resources invested in policy and infrastructure support). This is one of the reasons why scientific journals and research-funding organizations are increasingly calling for further data sharing[6] or why industry bodies, such as the International Data Spaces Association (IDSA) (https://www.internationaldataspaces.org/) are investing in reference architectures to smooth data flows from one business to another.

There is plenty of advice on how to make data easier to reuse, including technical standards, legal frameworks, and guidelines. Much work places focus on machine readability



FYI:

- Hiring: assistant prof. in data management methodologies
- Special Issue JWS on Knowledge Engineering for large scale knowledge graphs