Data quality, transparency and reproducibility in large bibliographic datasets

Angela Zoss[1], Trevor Edelblute[2], Inna Kouper[2]
[1]Duke University
[2]Indiana University

http://bit.ly/IASSIST-bibdata

What are bibliographic datasets?

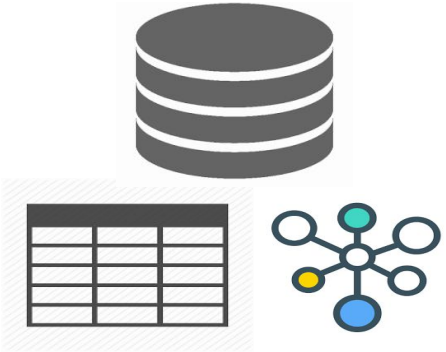
Data about formal products of scholarly communication (esp., conference papers, journal articles, monographs)

Can include:

- publication metadata
- citation information
- author affiliation information
- publication venue information
- full text of publications

Bibliographic database

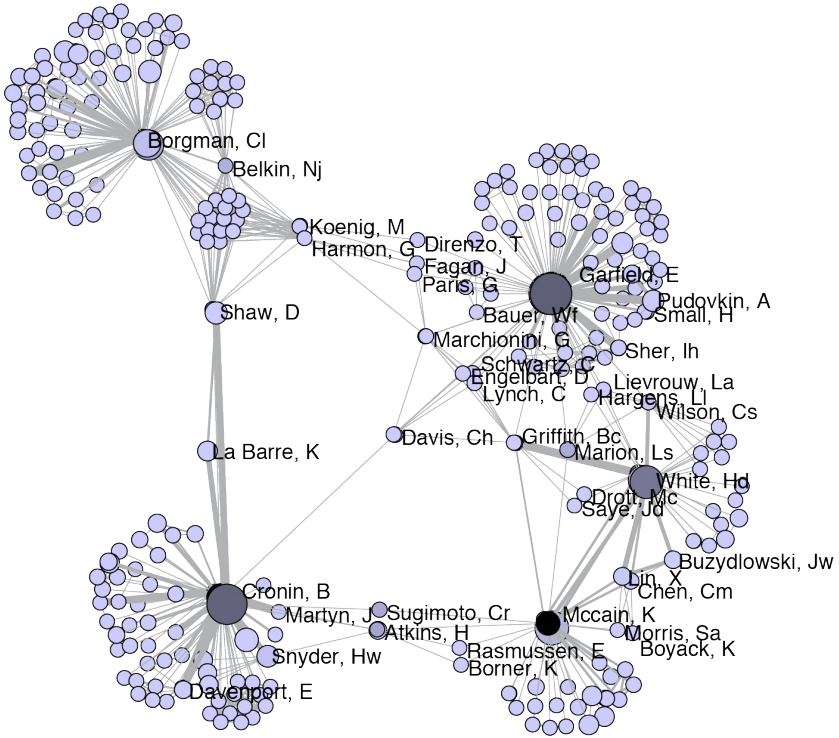
e.g., individual journals, aggregators, library collections, citation indexing services



Bibliographic datasets

Why use bibliographic datasets?

- Study the processes of science
  - the growth of a field
  - the birth of an innovation
  - the influence of certain theories or researchers on other schools of thought
  - changes in communities and collaborations
- Interest cuts across disciplines



http://dx.doi.org/10.1002/bult.2012.1720380608

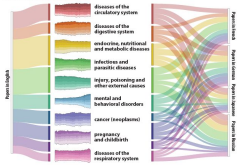
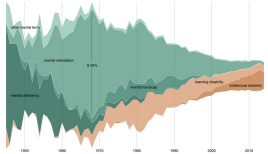
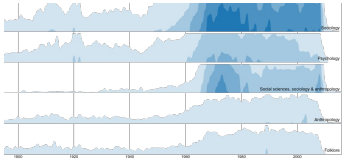
Reproducible workflows

- Same databases are used in multiple projects by same or other teams
- How can we save time preparing and using datasets?
- How can we ensure datasets are reliable and reusable?

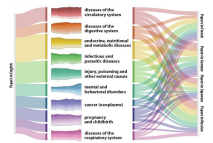
Database comparison

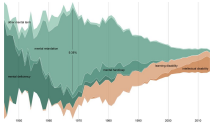
	PubMed	**dblp**	**ACM DL**	**HathiTrust**
Content type	Biomedical	Computer science	Computer / Info. science	Library collections
Number of records	~ 27 million	~ 4 million	~ 3 million (< 500k full text)	> 14 million
Coverage	1800s - current	1930s - current	1950s - present	900s - current
Non-bibliographic metadata	MeSH subject headings, abstracts	Venue information	Author affiliations, paper refs & citations	LoC subject headings
Download	XML / Text format, 10K batches	XML / JSON, whole dataset or batches of 1k	**Not allowed**	JSON, batches

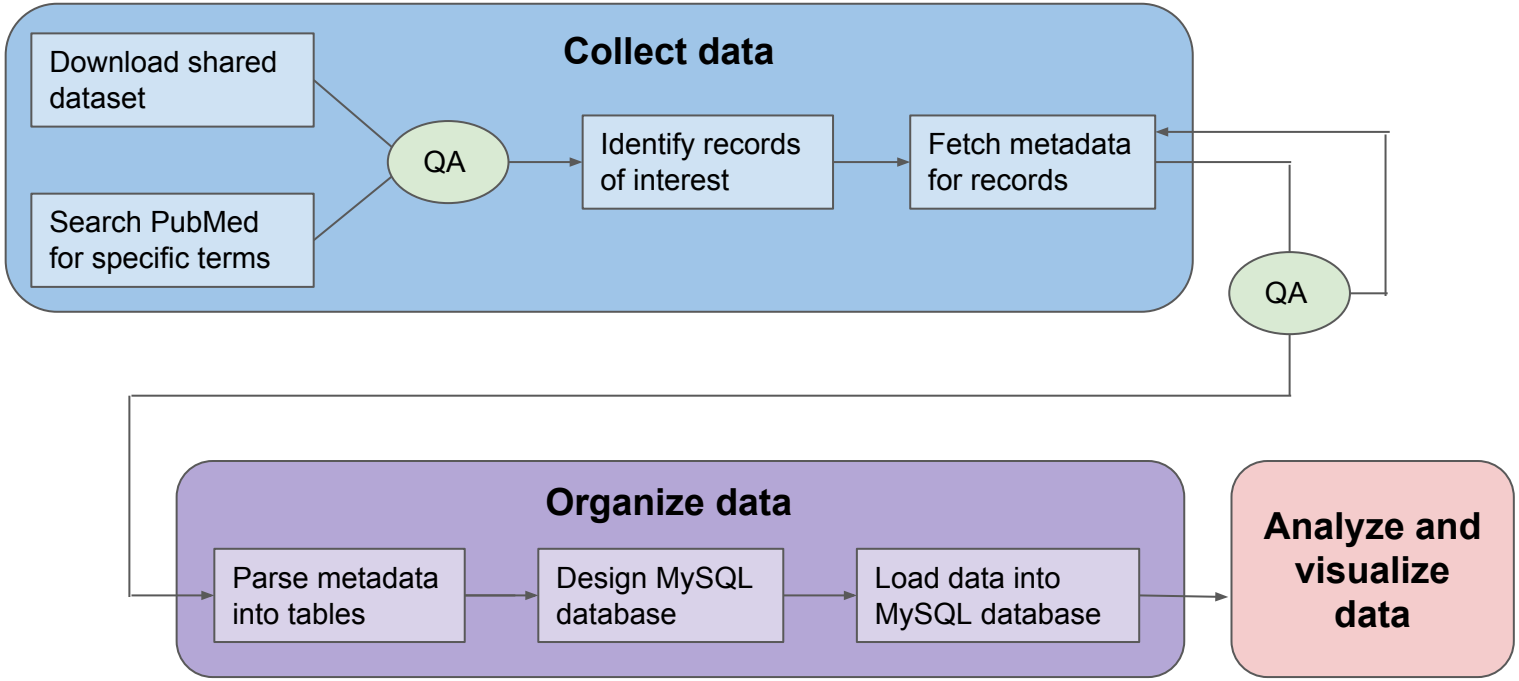
Dataset comparison

	WebSci '14	**Mental Disorders**	**SIGWEB communities**	**Social Science Lit.**
				
Content type	Biomedical	Biomedical	Computer science	Library collections
Number of records	~ 22 million	~930K	~ 10K	~ 14 million
Coverage	1800s - current	1945 - current	1987 - current	900s - current
Original database	PubMed	PubMed	dblp / ACM DL	HathiTrust DL
Acquisition method	Other group	Direct download	Other group	Direct download

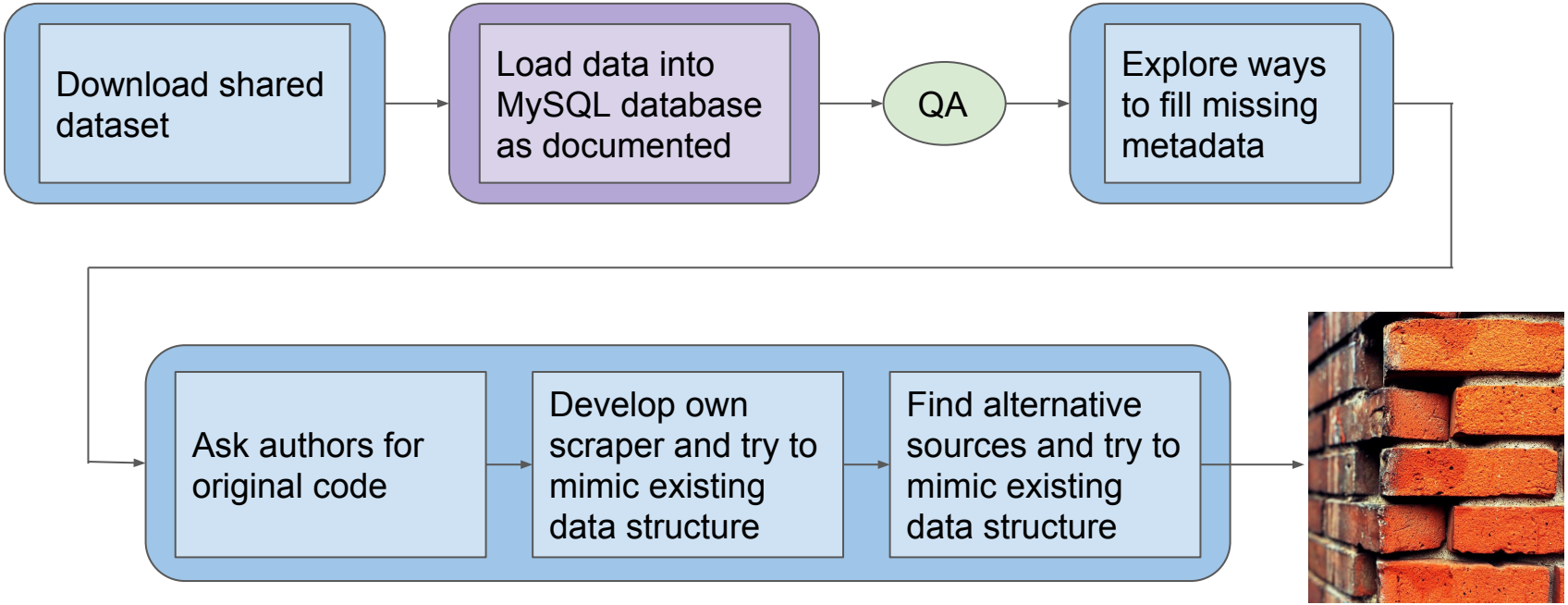
WebSci'14 and Mental Disorders datasets workflow





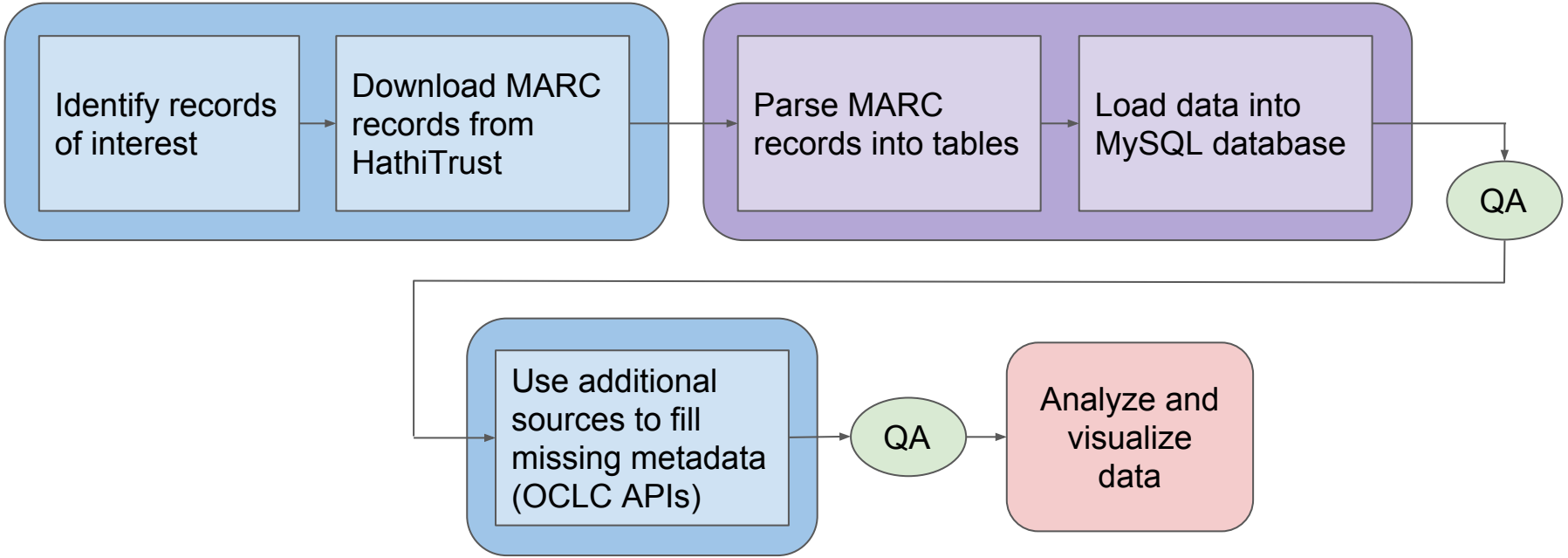


SIGWEB dataset workflow



https://data.mendeley.com/datasets/dn5d8fbkb9/1

Social Science Lit. dataset workflow



Challenges with large bibliographic datasets

- Need skills to edit and use **scripts**, multiple file formats, databases
- **Scaling up** takes extra time, resources, and effort in management of infrastructure
- The larger you get, the more prone to **errors** your dataset is (hard to check)
- Automation doesn't prevent failure; sophisticated **QA** procedures are needed
- Different APIs have different **features**
- Data*set* **quality** depends on data*base* quality
- Processes (scripts, QA, etc.) have to work for **non-English languages**

Comparing bibliographic databases

Database	**Quality** (data errors or gaps, type of fields available)	**Transparency** (documentation, ease of access and download)	**Reproducibility** (ability to automate retrieval)
PubMed	high	medium	high
dblp	medium	high	medium
ACM	high	none	none
HathiTrust	medium	medium	medium

Challenges with datasets from other groups

- May not have all **fields** needed/possible
- May be difficult to reproduce **original collection process**
- Bibliographic databases are **live** and being updated, compromising reproducibility of original dataset

How do we resolve these issues?

- Open more databases for mining
- Standardize across databases and their APIs
- Create reproducible workflows
  - Use open tools
  - Share scripts, database definitions
  - Document manual steps (e.g., search terms, database design, QA, curation)
  - Create infrastructure to assign PIDs to queries
- Assign proper licenses
- Track versions

Experiment data files

Download all files (1)

DBLP-SIGWEB.zip 6 MB

Steps to reproduce

Commands to import data in MySQL

mysql -u root -p; \\log-in to mysql
enter your password
create database sigweb; \\create a new schema
use dblp; \\change the database
source filename.sql; \\import sql file

This repository Search Pull requests Issues Gist

pmvis / pmvis Private

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master pmvis / create_all_tables.sql Find file Copy path

amzoss adding SQL files for creating/loading data tables fd5b6ab on Oct 20, 2014

0 contributors

36 lines (31 sloc) 730 Bytes Raw Blame History

```sql
CREATE TABLE metadata (
pmid VARCHAR(8) NOT NULL PRIMARY KEY,
lang VARCHAR(30),
country VARCHAR(40),
title VARCHAR(200),
vern_title VARCHAR(200),
abstract VARCHAR(10000),
prim_abs_y_n, VARCHAR(1),
issn VARCHAR(9),
vol VARCHAR(20),
issue VARCHAR(20),
pub_year INT(4),
journal VARCHAR(200)
) ENGINE = MYISAM;

CREATE TABLE mesh (
pmid VARCHAR(8) NOT NULL PRIMARY KEY,
heading_num VARCHAR(3),
term VARCHAR(120),
desc_or_qual VARCHAR(4),
major_or_no VARCHAR(1),
term_type VARCHAR(11)
) ENGINE = MYISAM;


CREATE TABLE keywords (
pmid VARCHAR(8) NOT NULL PRIMARY KEY,
term VARCHAR(210),
```

This repository Search Pull requests Issues Gist

pmvis / pmvis Private

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master pmvis / multi_parser.py Find file Copy path

amzoss condense abstracts 9a5895c on Sep 13, 2014

0 contributors

151 lines (112 sloc) 4.95 KB Raw Blame History

```
#!/usr/bin/python
import csv
import codecs
import sys
reload(sys)
sys.setdefaultencoding("utf-8")  # This block helps write the csv in utf-8

try:
        import xml.etree.cElementTree as ET
#       import cElementTree as ET
except ImportError:
        import xml.etree.ElementTree as ET
import re
import os

meta_csv = codecs.open('metadata.txt', encoding='utf-8', mode='a') #append mode
meta_writer = csv.writer(meta_csv, delimiter="|", quotechar='"', escapechar='\\')

mesh_csv = codecs.open('mesh.txt', encoding='utf-8', mode='a') #append mode
mesh_writer = csv.writer(mesh_csv, delimiter="|", quotechar='"', escapechar='\\')

key_csv = codecs.open('keywords.txt', encoding='utf-8', mode='a') #append mode
key_writer = csv.writer(key_csv, delimiter="|", quotechar='"', escapechar='\\')

affil_csv = codecs.open('affiliations.txt', encoding='utf-8', mode='a') #append mode
affil_writer = csv.writer(affil_csv, delimiter="|", quotechar='"', escapechar='\\')

#ref_csv = codecs.open('references.txt', encoding='utf-8', mode='a') #append mode
```

Links to projects

Diseases across the Top Five Languages of the PubMed Database: 1961-2012
http://bit.ly/2qQTMY0

Mental Disorders over Time: A Dictionary-Based Approach to the Analysis of Knowledge Domains
http://bit.ly/2qQKb3n

Questions?

angela.zoss@duke.edu

http://bit.ly/IASSIST-bibdata