Use of an innovative meta-data search tool improves variable discovery in large-p data sets like the Simons Simplex Collection (SSC)

Leon Rozenblit, JD, PhD

Presenter Disclosures

Leon Rozenblit

(1) The following personal financial relationships with commercial interests relevant to this presentation existed during the past 12 months:

- Employment by commercial entity, Prometheus Research, LLC
- Stock ownership, Prometheus Research, LLC

Objectives

- Describe the process for developing an agile software tool that promotes **variable discovery** in large data sets
- Assess the value of a technological approach for facilitating autism research and promoting data sharing
- Discuss how researchers who work with large, complex data sets can adopt this approach

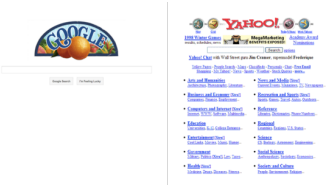
Background

- Simons Foundation Autism Research Initiative (SFARI): Simons Simplex Collection (SSC)
- Data collected for over 2600 families with at least one child affected with Autism Spectrum Disorder (ASD)
- 13 sites, lots of attention to consistency
- Repository for genetic samples and phenotype data, other linked data

Problem

- The SSC contains many thousands of variables
- Challenging for researchers to identify variables relevant for their projects
- Possible solutions
  - Ontologies
  - Data Dictionary
  - Variable browser approach

Google versus Yahoo



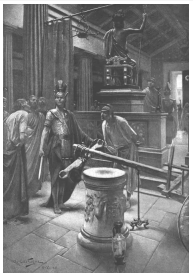
What are you looking for?

Verbal IQ Search

First Attempt

- Keep data in relational database
- Model meta-data as a separate schema in same database
- New data model tries to support complex search features (synonyms, concept weights)
- Resulted in poor performance

Simple Solution?

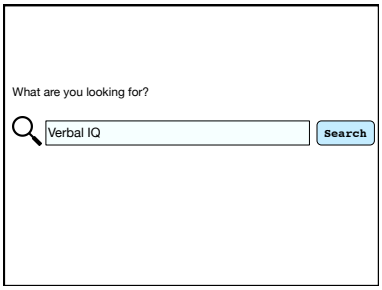


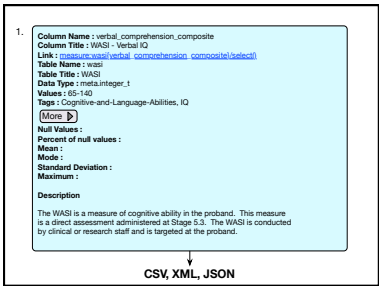
Solution

- **Pretend** each variable is a document
- Use standard document search techniques to find and rank variables

Solution

- Pretend each variable is a document
  - Build a "search report" (a structured index) for each variable
  - Build an output report for each variable
  - Store both reports as attributes in a searchable database where each row is a "variable"
- Use standard full-text search tools to find search report
  - Run full-text search function on search report attribute
  - Rank output
  - Return corresponding stored output report attribute





Challenges

- Search ranking algorithm tuning
  - Ranking by values
  - Appropriate weighting for different kinds of tags (manual, meta-data-derived, value-derived)
- Parser
  - Recognizing variants of Boolean operators
  - Stop word removal

Approach

- Agile software development
- Iterated over a 2 week cycle for 3 months
- Incorporated feedback from test users

Enabling Technology

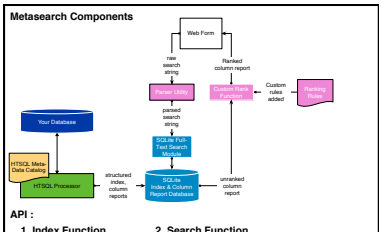
- SQLite
- SQLite Full Text Search
- **HTSQL**

What Does HTSQL Get You?

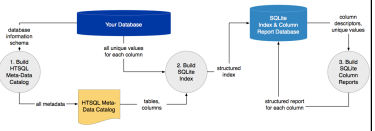
1. A relational database Web gateway: http://demo.htsql.org/school
2. An advanced query language where **the URI is the query**

   /course?credits>3&department.school='eng'

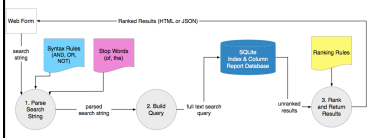
   /school{name, count(program),count(department)}
3. A REST-ful API for relational databases*
4. A way to build maintainable web-apps quickly and cheaply
5. A communication tool for developers, analysts, and end-users



Prepare for Search



Execute Search



What are you looking for?

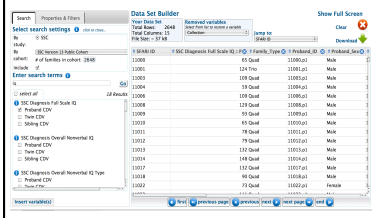
ASD Symptoms **Search**

1.

Column Name : scq_life_item_1
Column Title : SCQ Item 1 - Is she/he able to talk using phrases/sentences?
Link : ssc:commonly_used(scq_life_item_1)/select(i)
Table Name : commonly_used
Table Title : Other Commonly Used Variables
Data Type : measure.yes_no
Values : Yes (4016), No (228)
Tags : ASD Symptoms, Language

More

CSV, XML, JSON



Value of Approach

- Lightweight, easy to implement, low cost
- Fast!
- Accessible via the Web (via HTSQL)
- Improved usability
- Promotes data use, reduces support burden
- Potential for better data integration
- Can be used on top of <u>any</u> relational database

Implications for Public Health Practice

- Optimizes data retrieval
- Promotes interdisciplinary data usage
- Aids in analytical studies
- In use in SFARI Base, a data dissemination system
  - Over 120 research projects
  - Distributed more than 130,000 biospecimens

Acknowledgements

- Prometheus Research
  - Alexey Voronoy
  - Matthew Peddle
  - Clark Evans
  - Naralys Sinanis
- Weill Cornell Medical College
  - Stephen Johnson

Different from Spotlight?

	Document Search	Meta-Search
Prepare to Search	Scan file-system for documents	Scan a database for columns (across all tables)
	Build index entry for each document	Build index entry for each column (structure entry to support different search strategies, take into account related meta-data like table names, as well as values stored in each column)
Execute Search	Parse search string	Parse search string
	Compare terms in search string to index	Compare terms in search string to index
	Identify matching documents	Identify matching columns
	Rank matching documents	Rank matching columns (utilizing differential ranking weights from different search strategies)
	Return ranked list of documents	Return ranked list of "column reports"