VaxStats Repository

LIS 598 J with Dr. Nic Weber

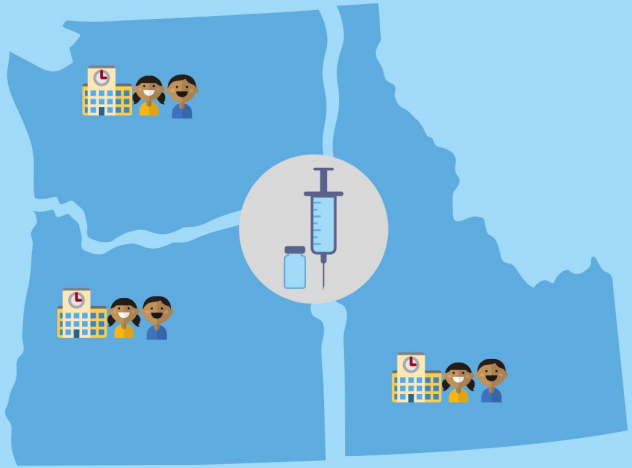
Alexis McClimans, Karalyn Ostler, Kaitlin Throgmorton

VAXSTATS
an open vaccination data repository for the PNW

Brief Overview

VaxStats is an open vaccination data repository for the PNW.

It provides vaccination and immunization data about school-age children in the Pacific Northwest states of Idaho, Oregon, and Washington.



VAXSTATS
an open vaccination data repository for the PNW

Repository Goals

To fully understand how vaccination rates and disease outbreaks are related, and in order to prevent them in future, the VaxStats repository aims to:

- Provide a central repository for all vaccination data on school age children in the PNW, as well as related data like outbreak data and policy data
- Establish collection standards that allow disparate vaccination datasets to be discovered and compared
- Allow users to visualize multiple layers of data at once, through curation of data to improve interoperability and machine readability

VAXSTATS
an open vaccination data repository for the PNW

User Community

- Health officials and medical professionals
- Education administrators, teachers, students, and parents of students
- Policy analysts, legislators, and other citizen interest groups
- Researchers and nonprofit groups
- Journalists
- General public

User Community Features

This user community comes with some unique constraints, including:

- Legal constraints
  - Vaccination law varies from state to state, ranging from stringent to lax
- Privacy constraints
  - Most vaccination data is medical in nature, making it subject to rigid privacy protections
- Interoperability constraints
  - Because of the two constraints above, no state is producing vaccination data in the same way, making interoperability challenging

User Community: Data Requirements

User Group	Data Access	Data Format	Data Analysis
Department of Health	API, File download, GUI browsing	Plain text, non-encoded CSV	See full picture of vaccination data for their area
School Administrator	File download, GUI browsing	Open formats, not specific	Track school vaccination data over time, and compare data from surrounding districts
Policy Analyst	API, File download, GUI browsing	Plain text, non-encoded CSV	View historical and multi-region data, as well as visualize it
Journalist	File download, GUI browsing	Open formats, not specific	Find data easily, read data descriptions, and visualize data
Epidemiologist	API, File download, GUI browsing	Plain text, non-encoded CSV	Load data into statistical processing software and compare with other data
General Public	GUI browsing	Open formats, not specific	Find and browse data quickly and easily

User Community: User Stories

Users Seeking Data

Goal	User Story
Find data about vaccination exemptions	*"As a parent, I want to know what the vaccination exemption rate is in my county."*
Find data about vaccination requirements and recommendations	*"As a policymaker, I want to better understand what other states in the PNW require for vaccination."*
Find data about multiple types of vaccines	*"As a health official, I want to know which vaccines are waived most frequently."*
Download data in machine-readable formats	*"As a journalist, I want to analyze a lot of data at once in order to tell a data-driven story about a recent measles outbreak in my area."*

Users Depositing Data

Goal	User Story
Transform data into an acceptable format for deposit	*"As a public health official, I want to take aggregated data in PDF format, and produce a raw dataset appropriate for deposit."*
Protect sensitive data before placing in public repository	*"As a local government employee, I want to ensure that any data released to the repository won't compromise personally identifiable information of citizens."*
Know how datasets will be stored, secured, and accessed in the future	*"As a researcher, I want to know if this data is secure, and if it will be available persistently in the future."*

AXSTATS
an open vaccination data repository for the PNW

User Community Design Choices

We predicted our user community to be relatively **large and diverse**, because **disease prevention, and thus vaccination and immunization, affects entire communities.** To account for this, we included both **data access requirements** and **user stories** broken down by individual user categories.

We also spent considerable time reviewing the various **constraints** entailed by this data, especially **applicable laws**, and decided to include a **full legislative overview** in our protocol, as well as **specifications for the creation of a legislative tracker** that would also be part of our repository.

Collection Policy

- Submissions-based repository with light curatorial services offered
- Only accepts data about required vaccinations and immunization schedules for school-age children in the PNW states of ID, OR, and WA

Design choices:

- Because of the complex nature of medical vaccination data, the repository is **submissions-based** with only light-touch curation.
- While data about non-required vaccines and adult vaccines exists, it tends to be less complete than data about required vaccines for children. Additionally, immunization for the most severe vaccine-preventable diseases almost exclusively occurs in childhood.

Ingest Policy

VaxStats data must be:

- Machine-Readable
- Human-Readable
- Licensed
- Secure
- Verified

Design choices:

- We sought to **strike a balance** here between making the repository data **open access** while also ensuring that we **protect the sensitive populations** our data is about.

VAXSTATS
an open vaccination data repository for the PNW

Deposit Policy

To deposit data with Vax Stats:

- Prepare data for deposit
- Ensure data is complete, correct, and well-documented
- Submit data and all accompanying documentation

Policy also includes:

- File formatting tips
- Data cleaning tips
- Information about VaxStats' data quality tiers
- Document conversion to machine-readable formats, e.g. Tabula, R, Python

VAXSTATS
an open vaccination data repository for the PNW

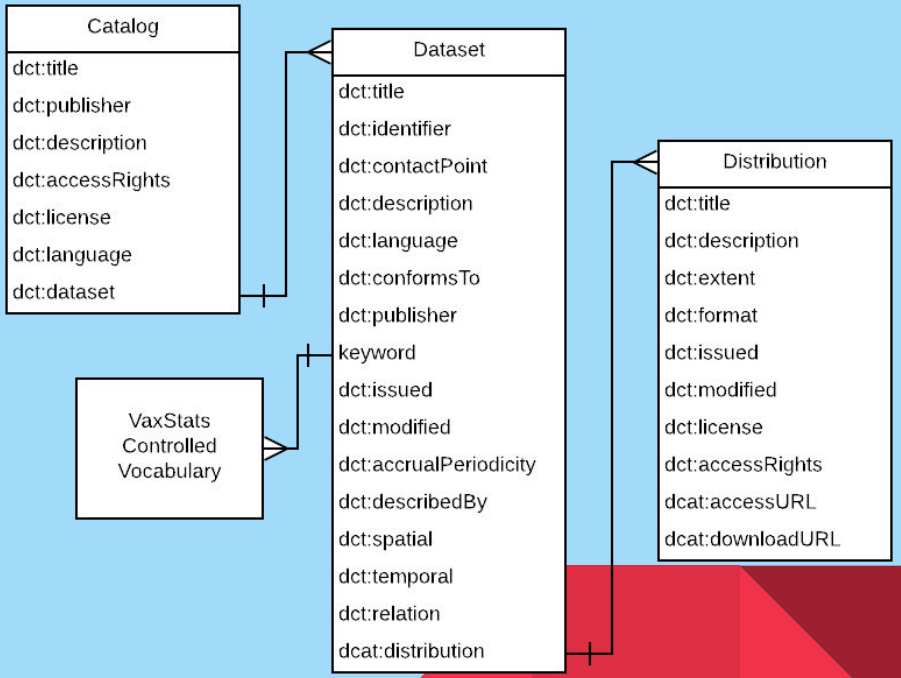
Deposit Policy

Design choices:

- Our deposit policy heavily **emphasizes documentation** (data descriptions, metadata, data dictionaries) in order to demystify the confusing nature of medical data. Vaccination data comes riddled with unexplained acronyms and complicated terminology, so data deposits must be well-explained for the lay user (and non-medical curator).
- The deposit policy also insists on **machine-readable file formats** and **cleaned data** to aid in interoperability. As the repository seeks to make interstate vaccination data more findable, accessible, and reusable, **both human-readable and machine-readable data is essential**.

Metadata: Application Profile

VaxStats' metadata schema is modeled according to the Data Catalog Vocabulary (DCAT), an RDF vocabulary which uses terminology from the Dublin Core Terms vocabulary. It is intended for use in government repositories, and was especially suited to the needs of VaxStats user community.



Metadata: Controlled Vocabulary

An additional controlled vocabulary was needed to exert control over the keywords that were used to describe submissions. As the terminology used to describe submissions was largely medical in nature, VaxStats' leveraged the MeSH vocabulary provided by the National Library of Medicine to control potential keywords. For each individual keyword, there are SKOS preferred and alternate terms that facilitate expanded search, as well as provide control over term assignment.

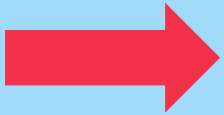
Terms include:

Diphtheria, DT (VACCINE), DTaP (VACCINE), DTP (VACCINE), Hep A, Hep B, HiB, HPV (VACCINE), Immunization, IPV (VACCINE), Measles virus, Meningococcal, MMR (VACCINE), Mumps virus, NIS, PCV, Pertussis, Polio, PV, Rubella Virus, Tdap, Tetanus, Vaccination, Varicella, Varicella (VACCINE)

Datasets

Data from source:

- File Format:
  - 7 Excel Books (.xlsx)
  - 11 PDF files
- Content:
  - Mix of documentation and datasets
- Multiple values per cell
- No standardized formatting



Converted data for deposit:

- File Format:
  - 61 .csv files
- Content:
  - Only data from tables
  - One table per file
- Machine readable
- Columns: variables
- Rows: observations
- Separated documentation

AXSTATS
an open vaccination data repository for the PNW

Data Conversion

To convert the PDF locked data into a machine readable form, we used the browser based tool Tabula.

Steps:

1. Use Tabula to convert as much data as possible from PDF table
2. Export each table in file as a separate .csv file
3. Open using spreadsheet software or text editor
4. Add in data lost in conversion (e.g. vertical text, multiple values in one cell)
5. Add original file name, table name, source information to README file
6. Save extra documentation in PDFs or Excel as separate PDF files

Conclusion

- VaxStats GitHub repository
- VaxStats Gitbook protocol (full written report)

VAXSTATS
an open vaccination data repository for the PNW