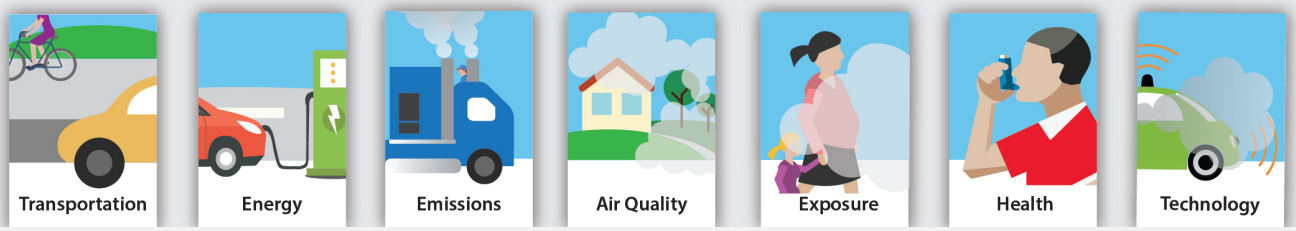
DEVELOPING A TRANSPORTATION, EMISSIONS, AND HEALTH DATAHUB

CARTEEH DataHub

Data for Advancing Research in
Transportation Emissions, Energy, and Health

Discover Data. Exchange Information. Understand Impact.

Browse Topics



August 2019

CARTEEH

Center for Advancing Research in
Transportation Emissions, Energy, and Health
A USDOT University Transportation Center

Texas A&M Transportation Institute
Georgia Tech College of Engineering
Johns Hopkins Bloomberg School of Public Health
College of Engineering The University of Texas at El Paso
UC Riverside CE-CERT

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Developing a Transportation, Emissions, and Health Database		5. Report Date August 2019 6. Performing Organization Code
7. Author(s) Andrew G. Birt, Dan Seedah, Ann Xu		8. Performing Organization Report No. TTI-01
9. Performing Organization Name and Address: CARTEEH UTC Texas A&M Transportation Institute 3135 TAMU College Station, Texas 77843-3135		10. Work Unit No. 11. Contract or Grant No. 69A3551747128
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (USDOT)		13. Type of Report and Period Final May 2017–August 2019 14. Sponsoring Agency Code

15. Supplementary Notes

This project was funded by the Center for Advancing Research in Transportation Emissions, Energy, and Health University Transportation Center, a grant from the U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program.

16. Abstract

The CARTEEH datahub addresses the need for a systematic, cross-disciplinary approach to understanding and integrating transportation and health datasets. This report documents the development of the datahub and includes a technical review of off-the-shelf software solutions for implementing a datahub, as well as a description of high-level goals describing how the datahub should affect CARTEEH research. Based on the technical review and goal statements, the research team implemented a custom datahub using modular software components. The implemented datahub is based on concepts of effectively communicating existing data and other research products within a collaborative research center, and valuing and reusing all research products.

17. Key Words	18. Distribution Statement
Datahub, Information Management System, Information Technology	No restrictions. This document is available to the public through the CARTEEH UTC website. http://carteeh.org

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	32	$0.00

Executive Summary

The Center for Advancing Research in Transportation Emissions, Energy, and Health (CARTEEH) datahub addresses the need for a systematic, cross-disciplinary approach to understanding and integrating transportation and health datasets. It facilitates the sharing and access of data, models, and knowledge products useful for research into the impacts of transportation on human health.

CARTEEH is a U.S. Department of Transportation funded research center focusing on the impact of transportation emissions on human health. The Texas A&M Transportation Institute (TTI) leads the CARTEEH research center, which includes four partner universities: Johns Hopkins University, Georgia Institute of Technology, the University of Texas at El Paso, and the University of California, Riverside.

CARTEEH research focuses on relationships between transportation and health—two core research disciplines that have not traditionally worked together. The transportation-health nexus requires the collaboration of a diverse range of researchers, including transportation engineers, social scientists, toxicologists, epidemiologists, medical health practitioners, computer scientists, and statisticians.

The goal of this project was to develop an information management system (IMS)—the CARTEEH DataHub (https://carteehdata.org/)—capable of improving the collaboration among this geographically and experientially diverse group of researchers. This datahub will become a central repository of data generated by the CARTEEH research center, in addition to existing data useful for transportation-health research. In recognition of the interdisciplinary nature of CARTEEH research, the datahub is designed to bridge knowledge gaps among CARTEEH researchers who have traditionally worked in different domains.

TTI researchers developed the CARTEEH datahub based on four design goals, which were set based on how the datahub could improve research activities:

- Goal 1: Encourage all research products to be reused in future research.
- Goal 2: Improve the meaning and interpretability of all research products in a way that promotes their reuse.
- Goal 3: Encourage researchers to proactively communicate research products.
- Goal 4: Enhance and encourage collaborations by providing a diverse repository of research products and ideas.

The research team used a review of potential off-the-shelf IMSs to guide the development of the datahub. The technical objectives were to design an IMS that could be used to organize and disseminate existing transportation-health research and that was flexible enough to be continually developed to meet future CARTEEH research needs.

The key product from this study is a datahub software and technology platform currently in use by CARTEEH. The software has been developed to meet the unique demands of the CARTEEH research and currently contains over 1000 searchable datasets and other information organized by subject area or information type.

The CARTEEH datahub will enable CARTEEH researchers to conduct more efficient and impactful research. The datahub is currently in use and in continual development.

Table of Contents

List of Figures

List of Tables

Introduction

Transportation is a major source of air pollution, which in turn affects human health and wellbeing. The effects of air quality on human health have been recognized for some time. In the United States, the link between air quality and health was federally recognized through the Air Pollution Control Act (1955) and is currently regulated through the Clean Air Act (1963). However, while this legislation has been demonstrably successful at improving the nation's air quality, the impacts of air quality on human health remain a topic of national and international importance. The World Health Organization estimated that outdoor air pollution was responsible for 7 million annual premature deaths in 2014, with air pollution being the biggest single environmental health risk and the cause of one in eight deaths worldwide. Even in developed countries with a track record of air quality regulation, air pollution has clear, measurable impacts on human health. For example, in the United States, anthropogenic transportation sources were estimated to be responsible for 28 percent of 107,000 premature deaths in 2011 that were related to inhalation of $PM_{2.5}$ (particulate matter with a diameter of less than 2.5 microns) [1].

The Center for Advancing Research on Transportation Emissions, Energy, and Health (CARTEEH) is a Tier 1 University Transportation Center (UTC) focused on addressing emissions in the context of public health. CARTEEH is led by the Texas A&M Transportation Institute in consortium with Johns Hopkins University, the Georgia Institute of Technology, the University of Texas at El Paso, and the University of California, Riverside. CARTEEH's focus areas include policy and decision-making, emissions and energy estimation, alternative technologies, exposure assessment, public health impacts, and data integration. A major component of CARTEEH is collaboration between researchers at different universities and operating in different research fields. CARTEEH's goal is to use this diverse expertise to address research gaps and unify the fields of transportation and public health research and policy.

This project deals with the development of a web-based datahub that will underpin research undertaken within CARTEEH. The datahub is an information management system (IMS) that centralizes and organizes existing transportation-health information. The principal goal of the datahub is to encourage the reuse of existing transportation-health research products. The definition of reuse is broad and includes physically reusing datasets, analyses (code), or other research products in new studies, or using existing research projects to understand the principles of a research method. In all cases, the rationale is that reusing existing research products will improve the efficiency and quality of future transportation-health research.

This document describes the development of the CARTEEH datahub. It begins with a review of existing IMS software solutions and outlines the design goals and rationale of the CARTEEH datahub. This information is followed by sections describing the functionality of the CARTEEH datahub, as well as future work and opportunities.

Methodology

This section of the report describes the methods and rationale used to develop the CARTEEH datahub. First, we describe a technical review of existing software solutions that offered potential solutions for the CARTEEH datahub. Second, we describe some design goals for the datahub based on how it would positively affect CARTEEH research. Finally, we provide a short, non-technical description of the methods used to implement the CARTEEH datahub.

Potential Software Solutions for the CARTEEH Datahub

Four off-the-shelf IMS solutions were evaluated in line with the design principles of the CARTEEH datahub. Table 1 summarizes these solutions. These solutions were selected based on their current popularity, ease of installation, available documentation, and our initial knowledge of their capabilities and features.

Table 1. IMS Software Solutions Evaluated by the Research Team

Product	Description
ckan https://ckan.org/	The Comprehensive Knowledge Archive Network (CKAN) is an open-source, free-to-use data management system designed to make data accessible by providing tools to streamline publishing, sharing, finding, and using data. CKAN has an active community of developers who continually maintain its core technology and add extra functionality through CKAN extensions. CKAN is mainly used by public institutions seeking to share their data with the general public.
dkan https://getdkan.org/	DKAN is an open data portal based on CKAN and is written as an extension of Drupal, a free and open-source content management framework. The main difference in technologies between CKAN and DKAN is that CKAN is written in the Python programming language and DKAN is written in the hypertext preprocessor language (PHP). DKAN provides a suite of cataloging, publishing, and visualizing features that enable organizations and individuals to share data and write stories describing the data.
The Dataverse® Project https://dataverse.org/	Dataverse is an initiative of the Institute for Quantitative Social Science, Harvard University Library, and Harvard University Information Technology. It is an open-source web application built for sharing, preserving, citing, exploring, and analyzing research data. A key design feature of Dataverse is its ability to host multiple virtual archives called dataverses. Each dataverse contains datasets and descriptive metadata and can host other dataverses.
data.world https://data.world/	Data.world is a subscription-based IMS solution for organizations and individuals to catalog their data and share with others. It includes tools that facilitate data exploration, data profiling, data quality control, and integration with a number of third-party applications.

We modified an evaluation matrix developed by the Texas Digital Library Data Management Working Group (TDL-DMWG) [2] to explore the software. The TDL-DMWG matrix was created by its members to select a software solution for the Texas Digital Library. To assist with the selection process, TDL-DMWG developed an evaluation matrix made up of four main functional categories (ingestion, processing, curation, and access). Within each main functional category, TDL-DMWG defined a more detailed list of evaluation criteria.

We used the TDL-DMWG evaluation matrix to score the potential IMS software summarized in Table 1. Because our design objectives were to develop a flexible, adaptable IMS, we added hosting and pricing and the ability to be customized to the evaluation categories. For each function, the IMSs were scored on a 4-point scale (Table 2). Five functional groups were used to evaluate existing IMSs for the CARTEEH datahub: data ingestion, data processing, data curation, data access, and hosting and pricing.

Table 2. Scoring the Functionality of IMS Software

●	Complete (i.e., the solution fully implements the function described).
◔	Somewhat Complete (i.e., the solution implements the function described but lacks a minor desired feature for the CARTEEH datahub). Additional information is provided when this score is used.
◐	Partially Complete (i.e., the solution implements the function described but lacks a significant desired feature for the CARTEEH datahub). Additional information is provided when this score is used.
○	The software solution does not implement the function described.

The following sections provide an overview of each functional category, with a review of the evaluation criteria and evaluation outcomes.

Data Ingestion

Data ingestion is the transfer of data from assorted sources to a storage medium where it can be accessed, used, and analyzed by individuals or an organization. For IMS solutions such as the CARTEEH datahub, data ingestion occurs when a user selects a file to upload into the system. Data ingestion also involves gathering and attaching metadata, copyright information, or access permissions to the data. At the end of the data ingestion process, the transferred data should be accessible, reusable, well documented, version controlled, and searchable. The data ingestion functionality and descriptions evaluated are as follows.

- **File Uploads:** File uploads involve a user transferring a file to the IMS by selecting the file from the user's computer.
- **Link to External Dataset:** To minimize dataset duplication across multiple systems, data contributors should have the option of linking to other datasets.
- **File Formats:** A critical component of data ingestion is the ability for a system to accept various file formats, including but not limited to geographical information system (GIS) files, images (uncompressed and compressed), videos (uncompressed and compressed), text files, R files, and Microsoft Excel files.
- **Controlled Vocabulary:** Controlled vocabulary enables system administrators to set predefined keywords, tags, topics, or terminologies such that all data contributors use the same standardized list in describing their data. Controlled vocabulary ensures uniformity across the entire platform such that it is easier for users to search for related content and link datasets together.
- **Copyright Permissions and Notification:** Typically, web applications that allow users to upload and share data serve a term of use and copyright permission agreement to the user during user registration or before file upload.
- **File Size Limits:** File size upload limits are set by system administrators to control upload and download bandwidths and ensure that adequate disk space is available.
- **Custom Metadata Schema:** Metadata is documentation about data. A metadata schema describes the fields and format of metadata.
- **Data Reuse Information:** Data reuse involves using research data for a research activity or purpose other than that for which it was originally collected.
- **Content Licensing:** Content licenses define the terms under which data or other research products can be reused—for example, for commercial or non-commercial purposes.
- **Required Metadata:** Well-defined metadata schemas enable system administrators to set required versus optional metadata fields.
- **Preservation Storage Notification:** Data preservation involves the implementation of regular backups of all data and directories. Backup scheduling is determined by the software application vendor or the system administrator of the IMS software.

Table 3 provides notes on the performance of each IMS. Table 4 provides an overall summary of the data ingestion evaluation of each evaluated system.

Table 3. Comments on Data Ingestion Functionality

Function	Evaluation Notes
File Uploads	The file upload function is available in all the IMSs tested. DKAN and Dataverse provide drag-and-drop file upload option.
Link to External Datasets	CKAN, DKAN, and data.world allow linking to external datasets.
File Formats	Each IMS was able to ingest a variety of file formats and retrieve them in native form.
Controlled Vocabulary	Dataverse allows administrators to edit and add controlled vocabularies and instructional text, set required fields, and define the user interface display for the field.
Copyright Permissions and Notification	DKAN and CKAN notify users of copyright permissions during file uploads. Dataverse and data.world do not provide alerts but instead include these in the website's terms of use policy.
File Size Upload Limits	Data.world has a default maximum file upload size of 2 GB (but this can be increased by request). DKAN, CKAN, and Dataverse file size limits are set by the system administrator. The upper limits of uploads are most likely limited by the web browser used to perform uploads. Most browsers have a maximum file size upload limit of 2 GB, while Google Chrome and Opera allow for file uploads greater than 4 GB.
Custom Metadata Schema	DKAN provides the simplest metadata definition schema. New fields can be added or removed by the system administrator. Dataverse allows administrators to customize dataset-level metadata using templates. CKAN's metadata definitions are difficult to customize. Data.world's metadata fields cannot be modified.
Data Reuse Information	All IMS solutions provide clear content licensing agreements for users, data cataloging, easy search interfaces, and the ability to find related datasets through controlled vocabulary. Each IMS also allows data contributors to select from a list of content licensing options during the data ingestion process. However, CKAN and DKAN provide more licensing options out of the box than Dataverse and data.world.
Content Licensing	Each IMS solution allows data contributors to select from a list of content licensing options during the data ingestion process.
Required Metadata	DKAN, CKAN, and Dataverse enable system administrators to set required fields during the metadata schema definition phase.
Preservation Storage Notification	Data.world's backup frequencies are unknown. Open-source systems like DKAN, CKAN, and Dataverse require system administrators to conduct backups on a regular schedule.

Table 4. Summary of Data Ingestion Evaluation

Function	CKAN	DKAN	Dataverse	Data.world
File Uploads	●	●	●	●
Link to External Datasets	●	●	○	●
File Formats	●	●	●	●
Controlled Vocabulary	●	●	●	◐
Copyright Permissions and Notification	●	●	◔	◔
File Size Limits	●	●	●	◔
Custom Metadata Schema	◐	●	●	◐
Data Reuse Information	●	●	●	●
Content Licensing	●	●	◐	◐
Required Metadata	●	●	●	●
Preservation Storage Notification	These are generally defined in the system administrator's terms of use agreement			

Data Processing

Data processing is the ability of a system to transform ingested data into meaningful information—for example, the ability to open GIS datasets and automatically generate a map of the data or the ability to understand tabular data and compute parameters such as number of records or number of fields. Advanced data processing systems also facilitate integration of ingested data with other datasets through extraction, transformation, and loading procedures.

- **Native GIS Data Processing:** GIS data processing involves the ability for a system to detect GIS files and display the content of the files on a map.
- **Interoperability via APIs:** Application programming interfaces (APIs) are communication protocols that enable external applications to access, update, or retrieve information from an IMS solution. APIs are commonly used by software developers to add functionality to an application without direct access to the code base. For data analytics and data interoperability tasks, APIs are useful for transferring information from an IMS to another analysis system, such as Microsoft Excel, R, Tableau, and Power BI.
- **System Notifications:** The system sends alerts to all administrators, team members, and/or content owners when data or metadata have been edited.
- **Version Control:** Version control is defined here as the task of keeping data consisting of many versions in a well-organized manner such that previous versions of the data can be easily tracked or identified.
- **Digital Object Identifiers:** A digital object identifier (DOI) is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the internet.
- **File Format Specifications:** This function involves the platform alerting administrators and data contributors when ingested content does not conform to a preset list of preferred file formats. Restricting which files can be uploaded can be done by administrators on all the evaluated platforms.
- **User Authentication:** User authentication is the ability of a system to approve or deny access until a user's registration or login is confirmed.

- **Access Levels for All Users:** In addition to user authentication, user permissions and restrictions need to be set at different levels on the platform. This function is required to ensure that only specific individuals can have access to certain information.
- **System Logs:** System logs keep track of activities on the platform, including number of file uploads, metadata on file size, date of deposit, depositor, and others. This information is used in understanding system performance issues, error messages, number of users, number of visitors to the site, and so forth.

Table 5 and Table 6 show the results of the data processing functionality evaluation for each IMS.

Table 5. Comments on Data Processing Functionality

Function	Evaluation Notes
Native GIS Data Processing	GIS data processing is natively supported by CKAN and Dataverse. CKAN's GIS support is provided by two CKAN extensions: ckanext-spatial and ckanext-geoview. Similarly, Dataverse provides native geospatial support through its GeoConnect platform. These extensions enable users to visualize their datasets once ingested into the system. DKAN allows users to draw geographic areas through its mapping interface but not process ingested GIS datasets. Data.world, on the other hand, allows for GIS file uploads only.
Interoperability via APIs	APIs are strongly supported by all four IMSs to perform various tasks, such as linking datasets, downloading datasets, and making updates to an existing dataset.
System Notifications	Notifications in DKAN can be set through the DKAN Workflow component. DKAN Workflow creates a moderation queue so that content is published to the live site only after a designated supervisor or group moderator has reviewed and approved it. CKAN notifications send emails when there is new activity on the user's dashboard and can send email notifications to users, for example, when a user has new activities on his/her dashboard. Notifications in Dataverse are also available for different actions.
Version Control	IMS solutions like Dataverse provide automatic version control where data are appended with a version number each time they are uploaded onto the platform. Version control enables both data contributors and users to be notified of changes to a dataset and any updates that have been made to the dataset.
Digital Object Identifiers	CKAN facilitates automatic DOI assignment through the installation of ckanext-do. Both CKAN and Dataverse require the administrator to have an account with a DOI service provider, for example, DataCite [3]. Data.world allows users to assign DOIs to a dataset but does not automatically generate or create DOIs upon data ingestion [4].
File Format Specifications	DKAN by default does have some restrictions on what files can be uploaded, compared to the others, which currently do not have any restrictions.
User Authentication	This functionality is available on all four IMS solutions evaluated. User authentication is required to create, edit, delete, or upload data on an IMS platform.
Access Levels for All Users	This function is available on all four IMS solutions evaluated.
System Logs	All four IMSs provide system logs.

Table 6. Summary of Data Processing Evaluation

Function	CKAN	DKAN	Dataverse	Data.world
Native GIS Data Processing	●	◐	●	○
Interoperability via APIs	●	●	●	●
System Notifications	●	●	●	●
Version Control	○	●	●	●
Digital Object Identifiers	●	○	●	◐
File Format Specifications	●	●	●	●
User Authentication	●	●	●	●
Access Levels for All Users	●	●	●	●
System Logs	●	●	●	●

Data Curation

Data curation involves the implementation of a workflow that ensures that prior to publishing, the data have gone through quality control steps. Data curation functions include:

- **Workflow:** This function allows users at various permission levels to view and approve content prior to publishing.
- **Messaging and Commenting:** This function allows users to communicate with other members using a messaging or commenting feature.
- **To-Do Lists:** This function allows users to generate to-do lists for team work on active data projects.
- **Collaborative Working Spaces:** This function allows team leaders to share data with a select group of team members so the data can be revised by the entire team prior to publishing.
- **Access Levels for Team Members:** This function allows team leaders to establish differing access levels for each team member. Access levels include public access, private access, view-only access, and editing access.

Table 7 and Table 8 show the results of the data curation evaluation.

Table 7. Comments on Data Curation Functionality

Function	Evaluation Notes
Workflow	DKAN is the only platform found to have a workflow feature. DKAN Workflow allows site managers and administrators to determine which users can add, edit, and delete content, as well as which users can view and approve content prior to publishing. It creates a moderation queue so that content is published to the live site only after a designated supervisor or group moderator has reviewed and approved it.
Messaging and Commenting	Each of the platforms evaluated has some form of messaging or commenting feature. However, data.world has the most advanced features of the four. CKAN enables users to make comments on datasets by installing a third-party extension [5]. DKAN's integration with Drupal allows registered users to make comments about a dataset. Dataverse, on the other hand, only allows direct messaging via the dataset webpage to the data publishers. Thus, comments are not visible on the public website.
To-Do Lists	None of the applications evaluated has an out-of-the-box solution for to-do lists. Data.world's discussion interface can be used by team members to generate to-do lists and follow-up items.
Collaborative Working Spaces	All four platforms evaluated provide users with the ability to collaborate, though the implementation varies from one platform to another.
Access Levels for Team Members	Of the four platforms, DKAN has the most advanced user access settings, with the ability to define custom roles and access levels.

Table 8. Summary of Data Curation Evaluation

Function	CKAN	DKAN	Dataverse	Data.world
Workflow	○	●	○	○
Messaging and Commenting	◐	◐	◐	●
To-Do Lists	○	○	○	◐
Collaborative Working Spaces	●	●	●	●
Access Levels for Team Members	●	●	●	●

Data Access

Data access covers a number of software features, such as dataset linking, metadata and data reuse information, data export, search engine optimization, and licensing:

- **Dataset Linking:** Dataset linking is the ability for a system to manually or automatically link a dataset to grant number, ORCID, object DOI, and related content (including publications produced from the data). Dataset linking is an advanced feature requiring a predefined metadata schema that enables a dataset to be linked to other resources.
- **Metadata Reuse:** Metadata reuse is the ability for a system to export metadata in various outputs for reuse in other systems. Metadata reuse is implemented via API calls where developers can request machine readable data, which include a dataset's metadata fields.
- **Data Reuse Information:** Data reuse information provides contextual information needed to reuse data, in the form of instructions on how users can use the metadata APIs to retrieve information or perform queries on individual variables. Well-documented data reuse information makes it easier for developers to integrate data from an IMS solution into external applications.

- **Data Reuse and Export:** Well-designed IMS solutions allow for the export of data into various formats for reuse in external applications.
- **Search Engine Optimization:** Search engine optimization (SEO) is the process of maximizing the number of visits to a website through search engines such as Google, Yahoo, or Microsoft Bing. SEO ensures that the IMS appears high on the list of results returned by a search engine.
- **Licensing Terms and Disclaimers.**

Table 9 and Table 10 summarize the results of the data curation evaluation.

Table 9. Comments on Data Access Functionality

Function	Evaluation Notes
Dataset Linking	Simple dataset linking can be achieved through manual user input of a resource's URL. Of the four systems evaluated, only Dataverse provides a built-in DataCite DOI generator. The DOI generator for CKAN is available through the ckanext-doi extension published by the National History Museum [6]. Both applications require the administrator to have a valid DataCite account, which is used in creating and synching the DOIs. DKAN and data.world [4] provide users with input parameters to attach manually generated DOIs to their content. In addition to DOIs, all four systems provide permalinks to their content, which can be shared publicly.
Metadata Reuse	CKAN, DKAN, and Dataverse provide the ability for users to export metadata separately. This feature was not found for data.world during the evaluation.
Data Reuse Information	All four IMS solutions evaluated provide adequate documentation on how to access and use their APIs.
Data Reuse and Export	All the platforms evaluated were found to provide the capability to export data from the platform. Data.world delivers the most advanced features for data reuse and exports through its data integration connectors, which out of the box support analytical platforms like Tableau, PowerBI, and R, among others.
Search Engine Optimization	All the platforms evaluated were found to include this feature.
Licensing Terms and Disclaimers	All the platforms evaluated were found to include the ability for system administrators to define licensing terms associated with datasets. Furthermore, information regarding appropriate use of the data can be included in the terms of conditions page of the website.

Table 10. Summary of Data Access Evaluation

Function	CKAN	DKAN	Dataverse	Data.world
Dataset Linking	◐	○	●	○
Metadata Reuse	●	●	●	○
Data Reuse Information	●	●	●	●
Data Reuse and Export	●	●	●	●
Search Engine Optimization	●	●	●	●
Licensing Terms and Disclaimers	●	●[7]	●	●

Hosting and Pricing

Table 11 provides a summary of the evaluation results for hosting and pricing. CKAN, DKAN, and Dataverse are open source and can be self-hosted, while data.world is a commercially cloud-hosted platform that cannot be self-hosted. Self-hosted

systems enable the owner of the IMS to retain *full* control over both the IMS installation and its stored data. However, they must also be installed and maintained by the host—involving financial costs (for hardware) and human resource costs (such as backups, disaster recovery, and upgrades). On the other hand, solutions such as data.world provide an immediate solution for hosting DIK, but at the cost of losing full control over the information within the system.

Data.world's standard pricing is $100/month for up to five members, and then $15/month for additional members. This service includes an organizational account for members to view activity feeds and get access to team resources, as well as permission to upload an unlimited number of datasets (with a limit of 1 GB per dataset). Its enterprise pricing is $25 per member per month, with a minimum of 25 members and 1-year subscription [7]. The enterprise package service includes a number of additional advanced features, such as big data support (increased storage limits), virtualized datasets, and data warehousing.

Table 11. Summary of Data Hosting Evaluation

Function	CKAN	DKAN	Dataverse	Data.world
Self-Hosting	●	●	●	○
Authentication	◐	●	●	○
Backups	●	●	●	●

Because CKAN, DKAN, and Dataverse are open-source applications, the software is free, but administrators are required to pay for hardware (or cloud-based hardware solutions) to host the software. While the solutions do not offer on-demand support, each system provides sufficient documentation to install and configure an instance of the software. The recommended hardware requirements for a small to medium CKAN instance are two CPU cores, 4 GB of RAM, and 60 GB of disk space. An IMS with heavy traffic will require two servers with quad core processors with 8 GB of RAM (one for web and one for database) and 160 GB of disk space (if the system will host large files, more drive space is recommended) [8]. DKAN has the following minimum system requirements: 2 GB of RAM for production and 1 GB disk space. However, for data storage in MySQL, 100 GB of disk space is recommended. The Harvard University hosted Dataverse instance (see https://dataverse.harvard.edu) is run by four Amazon Web Services (AWS) server nodes: two m4.4xlarge instances (64 GB/16 vCPU) as web frontends, one 32 GB/8 vCPU (m4.2xlarge) instance for the search engine, and one 16 GB/4 vCPU (m4.xlarge) instance for R. The PostgreSQL database is served by the Amazon Relational Database System, and physical files are stored on Amazon S3 [9]. Table 12 shows the estimated cost of these hardware requirements based on cloud-based virtual machine hosting. These costs are for a high-traffic, high-usage datahub with regular daily backups. Systems expecting smaller traffic will be able to reduce costs by reducing the hardware requirements.

For authentication, CKAN provides Lightweight Directory Access Protocol (LDAP) and Single Sign On (SSO) authentication, while DKAN provides LDAP, SSO, and Open Authorization (OAuth) authentication. Dataverse uses Shibboleth and OAuth support for single sign on [10].

CKAN's backup functionality is provided through command line methods. DKAN has backup functionality on the user interface provided by the Drupal Backup and Migrate module. Information on data.world's backup schedule is not available.

Table 12. Harvard Dataverse Hosting Cost Estimates (sourced using 2019 AWS infrastructure costs matched to Dataverse system requirements)

Quantity	Service Type	Monthly Price	1-year All Upfront Reserved Instance Monthly Price
2	Amazon EC2 Service—m4.4xlarge	$ 1171.20	$ 958.58
1	Amazon EC2 Service—m4.2xlarge	$ 292.80	

1	Amazon EC2 Service—m4.xlarge	$ 146.40	
500 GB	Amazon S3 File Storage	$ 11.50	$ 11.50
200 GB	Amazon RDS—PostgreSQL—db.m1.xlarge	$ 381.68	$ 381.68

Summary of Evaluations

The off-the-shelf IMS solutions evaluated all provided functionality and features useful for developing the CARTEEH datahub. CKAN is a mature product with an open-source community of developers and multiple years of product development. DKAN is a derivative of CKAN and introduces the concept of data stories into IMS solutions. Dataverse was developed to facilitate the publication and subsequent tracking of original content created by researchers. All three products are relatively easy to install, configure, and customize. Because they are open source, they are essentially free to use apart from the costs of the hosting hardware and costs associated with maintaining or updating the IMS. We estimate that any of the open-source IMSs can be installed and maintained on hardware for $1000–$2000 per month.

Self-hosted solutions also allow system administrators to control the information uploaded to the system. This includes not only the original data uploaded to the system but, just as importantly, the connections and metadata associated with these uploaded data and information. In contrast, data.world is a commercial platform, with limited opportunity for customization or control over uploaded data.

Each of the open-source platforms was built for a particular purpose, and each has specific strengths. For example, data.world provides workflow functionality designed to enable teams of researchers to analyze and visualize datasets through the IMS. Dataverse was designed primarily for the research community to improve the visibility of research data and to ensure that any reuse of that data is controlled and credited. DKAN and CKAN were developed as traditional data portals designed to centralize, organize, and communicate datasets.

Design Goals for the CARTEEH Datahub

This section describes high-level design goals for the CARTEEH datahub. These design goals are outlined with a rationale for how they will improve CARTEEH research activities:

- **Goal 1: Encourage all research products to be reused in future research.**
- **Goal 2: Improve the meaning and interpretability of all research products in a way that promotes their reuse.**
- **Goal 3: Encourage researchers to proactively communicate research products.**
- **Goal 4: Enhance and encourage collaborations by providing a diverse repository of research products and ideas.**

Goal 1: Encourage All Research Products to Be Reused in Future Research

Figure 1 illustrates a conceptual view of data, information, and knowledge management in a traditional closed research environment. Typically, a research problem is formulated based on the accumulated knowledge within a research group. The same knowledge base is used to design experiments and collect, organize, and analyze data. Many of the products of research, for example, data or analyses (models or computer code), are stored locally. At the end of the research process, tangible research products are exported from the system via reports and academic papers.

The processes illustrated in Figure 1 lead to data and knowledge silos. The CARTEEH datahub is designed to break down data and knowledge silos by providing a centralized location to store and organize transportation-health research products. While traditional research tends to focus on the end products of research, the CARTEHH datahub is designed to encourage researchers to share intermediate research products such as data, analyses (e.g., code), reports, and data visualizations.

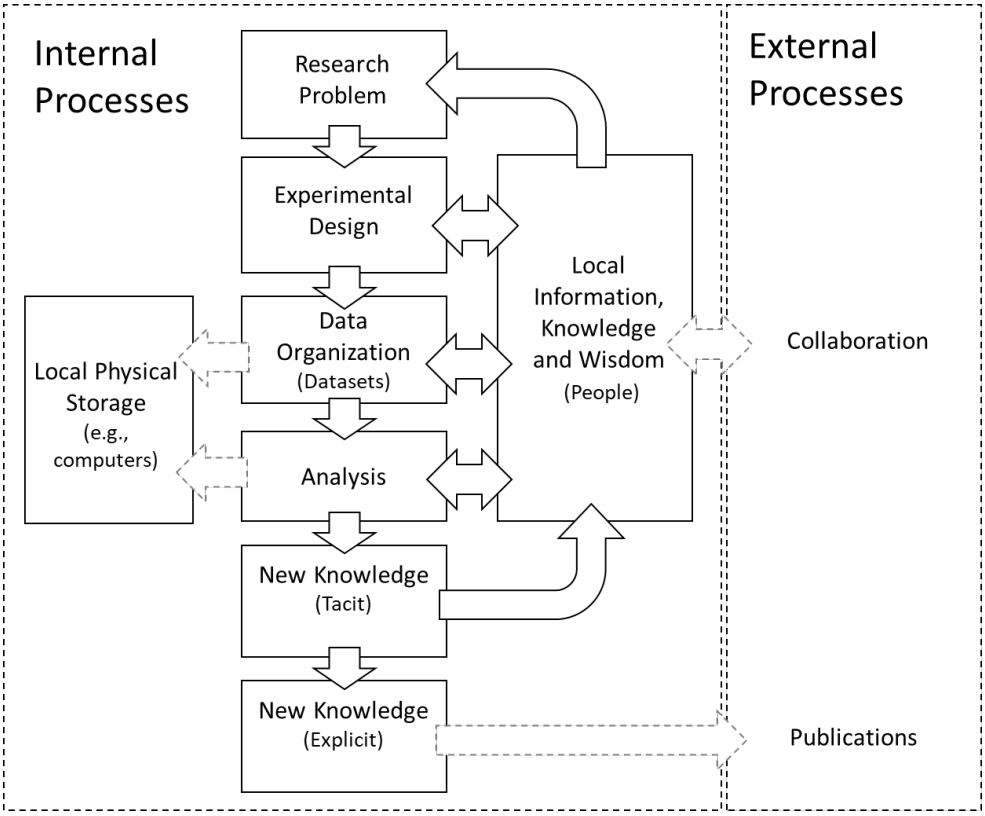


Figure 1. Conceptual model of data, information, and knowledge generation in a closed research environment.

Goal 2: Improve the Meaning and Interpretability of All Transportation-Health Research Products in a Way That Promotes Their Reuse

Figure 2 shows a a conceptual model of information management known as the Data-Information-Knowledge-Wisdom (DIKW) hierachy [11–13]. The DIKW model implies continuous pathways that transform raw data to information, information to knowledge, and knowledge to wisdom. The various entities in the hierarchy are defined relative to each other:

- **Data** are symbols that represent properties of objects, events, and their environment. They are the products of observation but are of no use until they are in a usable (i.e., relevant) form. To be made usable, data must be transformed to a form that provides meaning and context.
- **Information** is contained in descriptions and answers to questions that begin with such words as *who*, *what*, *when*, and *how many*.
- **Knowledge** is know-how and enables the transformation of information into instructions or understanding. Knowledge is data **and** information that have been organized and processed to convey understanding.
- **Wisdom** is the ability to think and act using knowledge, experience, understanding, common sense, insight, and judgment.

The DIKW hierachy suggests that there are implicit connections between entities at each level of the hierachy. For example, the data generated during a study are transformed to information products (figures, maps, summaries) and then to knowledge products (study reports, publications). Even after research is complete, these connections between raw data and knowledge remain important and useful. However, in conventional research environments, intermediate research products and their connections are often lost from the system.

A central idea underpinning the CARTEEH datahub is to acknowledge and retain the connections between different types of content generated through research activities. The goal is for researchers to use the datahub to deposit datasets, data visualizations, computer code, and other content, such as study reports, peer-reviewed papers, and presentations, in a way that helps maintain connections among data, information, and knowledge. Maintaining these connections will improve the

resusability of datahub content—especially data. The connections will enable other researchers to understand the meaning of past research products, increasing the likelihood that they will be reused in future research.

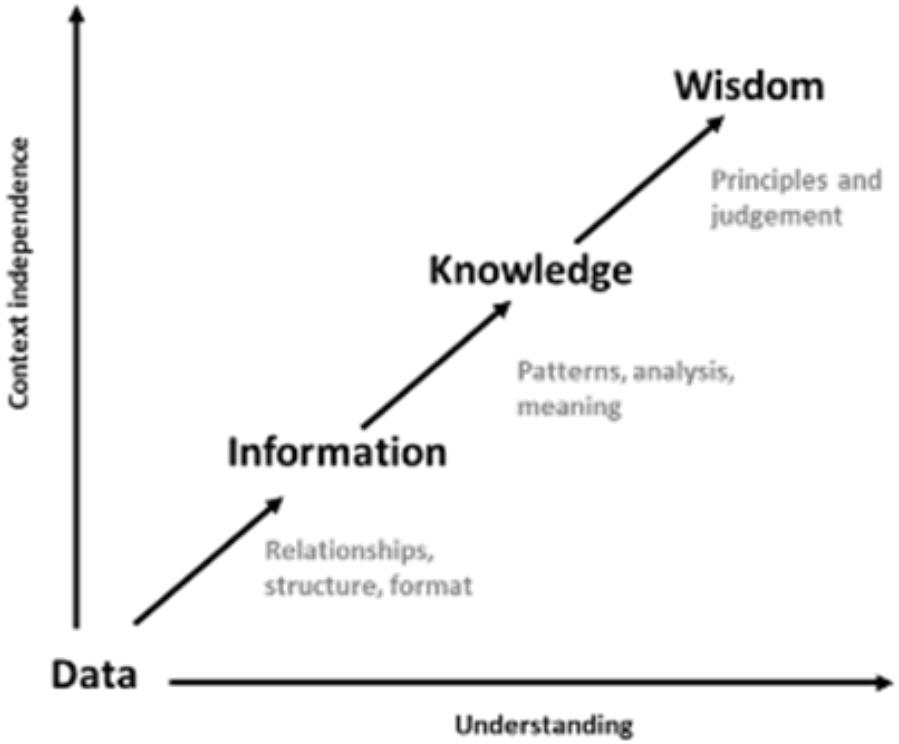


Figure 2. Conceptual model of the relationships among data, information, and knowledge.

Goal 3: Encourage Researchers to Proactively Communicate Research Products

Following the mathematical theory of communication, Weaver modeled information transfer in the context of communication. He defined communication as "all the procedures by which one mind may affect another" and further defined three levels of a communication problem [14]:

- Level A—How accurately can the symbols of communication be transmitted? (The ***technical*** problem).
- Level B—How precisely do the transmitted symbols convey the desired meaning? (The ***semantic*** problem).
- Level C—How effectively does the received meaning affect conduct in the desired way? (The ***effectiveness*** problem).

The goal of reusing data, information, and knowledge can be reframed in the context of communication (i.e., the challenge of communicating existing research products in a way that promotes their reuse). In the age of computers, the technical problem of accurately transmitting information has been largely solved. Datasets, documents, and computer code can be stored, opened, and transmitted with ease. In many respects, the simplicity and routineness of this technical problem prevents effective information reuse. For example, it is easy to save information in a highly accessible location and common digital format and therefore believe it is available for reuse. However, the reality is that it is unlikely to be reused unless the its meaning and context can be effectively communicated to other researchers (the semantic and the effectiveness problems of communication).

Encouraging other researchers to reuse datasets and other research products will require existing research to be proactively communicated (i.e., in a way that promotes understanding and that highlights its value for future research). We envisage the datahub to be a technological solution that will help researchers better communicate research products to their peers, as well as a tool that will help researchers *find* research products capable of positively influencing their future research.

Goal 4: Enhance and Encourage Collaborations by Providing a Diverse Repository of Research Products and Ideas

Researchers and funding agencies have begun to acknowledge the limitations of conducting research and analysis in the way outlined in closed research environments (i.e., as per Figure 1). Analyses of the impacts of scientific research have shown that scientists participating in larger and more diverse collaborative teams are more likely to earn prestigous awards and have increased federal funding, higher productivity, and higher research impacts [15, 16].

Figure 3 illustrates a modeling chain for transportation-health research [17]. It begins with the design and use of a transportation system; continues through models of emissions or energy use, with their effects on air quality; and finally moves to the effect of air quality on human health. This research pathway requires a considerable diversity of skills, knowledge, data, and understanding. CARTEEH research therefore requires collborations involving a wide range of sub-disciplines, for example, transportation engineering, health science, atmospheric and air quality science, statistics, toxicology, and so forth.

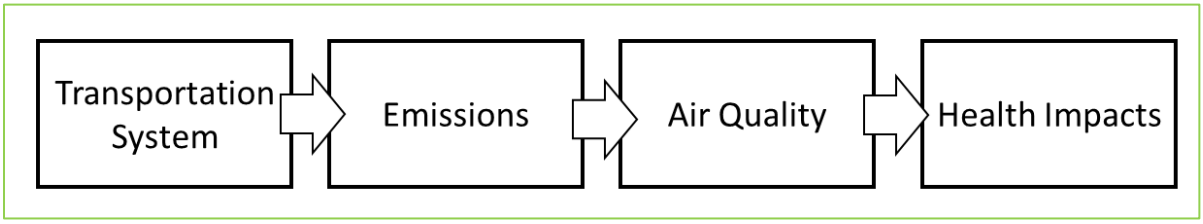


Figure 3. Conceptual transportation–air quality–health modeling chain.

A central idea of the CARTEEH datahub is to make collaborative research easier and more efficient. Effective collaborations occur through sharing of ideas [18] or through the effective division of scientific skills, knowledge, or resources [19]. In conventional research, these collaborations tend to involve interpersonal communication—of ideas, skills, resources, or data. However, traditional collaborations formed through interpersonal relationships often take time to develop and are frequently unavailable to junior researchers. They are often based around qualitative discussions of capabilities rather than tangible research products (e.g., data or models). Some potential collaborations may never be realized because research products are not effectively communicated or visible to the broader research community [20].

Stember defined different types of collaboration [21]. Intradisciplinary collaborations occur when multiple researchers work together within a single research discipline, for example, within a single step of the modeling chain shown in Figure 3. Multidisciplinary collaborations occur when researchers with different skill sets and expertise use existing methods to work on different components of a study. Multidisciplinary collaborations are useful for addressing large problems that extend beyond a single discipline (for example, across the full transportation-health modeling chain). Interdisciplinary collaborations seek to integrate and synthesize knowledge across disciplines by modifying conventional approaches and methods. For instance, integrating a new pollutant into transportation-health research (e.g., based on health data) could require modification of all steps in the modeling chain.

The CARTEEH datahub should promote all types of collaborations but recognize the special importance of interdisciplinary collaborations. While multidisciplinary collaboration can be made more efficient by encouraging the physical reuse of data (e.g., an existing dataset), interdisciplinary collaboration requires a different type of information reuse focused on creativity and exploration. Interdisciplinary research requires that individual researchers broaden their knowledge of research methods across the transportation-health nexus so that novel integrations of data and analyses can be developed and explored.

Implementing the CARTEEH Datahub

Following the technical review of existing IMS solutions and their functionality, as well as an analysis of the research issues likely to be faced by CARTEEH researchers, we developed our own IMS (datahub). The high-level design goals described in the previous section guided how data, information, and knowledge should be stored and organized within the software,

while the technical review (and detailed analysis of the source code provided in the open-source IMS) helped provide the technical scope.

Table 13 lists the core technologies used to develop the datahub. The core of the datahub was developed using NodeJS and ExpressJS—two software libraries that allow developers to seamlessly combine client and server-side code. Where possible, we integrated open-source libraries to ease code development. Open-source libraries also enhance long-term flexibility—new features can be added as required. Strapi was used to develop content storage, search, and retrieval modules that link to a Mongo-DB database. A separate MySQL database was implemented as storage for large, traditionally structured datasets.

The CARTEEH datahub is hosted on an AWS EC2 server running the Linux (Ubuntu) operating system. Additional cloud resources include a separate Windows EC2 server (hosting Microsoft Internet Information Services, MySQL database) and an Amazon Simple Storage Service (S3). Source code is maintained in a centralized GIT repository.

Table 13. Key Technologies Used in the CARTEEH Datahub

Technology	Description
NodeJS	Cross-platform JavaScript run-time environment that executes JavaScript code for server-side scripting. JavaScript is used primarily for client-side scripting for web browsers.
ExpressJS	Web-server framework for NodeJS used in developing web and mobile applications.
Strapi	Customizable and fully extensible headless content management system with an admin panel. Strapi addresses REST API handling, user authentication, and different content type creation. It was selected for its ease of use, default security features, speed, and well-written documentation.
Mongo-DB	Cross-platform document-oriented (No-SQL) database program. It is one of three database systems supported by Strapi and was selected because of its strong compatibility with Strapi's architecture.
LokiJS	Lightweight and fast in-memory, document-oriented datastore used by the datahub for caching content from Strapi.
Apache HTTP Server	Cross-platform web server that is used as a proxy server to redirect traffic to ExpressJS.
Amazon EC2	Cloud-hosted Linux and Windows servers to host the datahub and other custom web applications and services that can be linked back to the datahub. The datahub is served from AWS infrastructure and hosted at https://carteehdata.org.
Amazon S3	Cloud storage service used by the CARTEEH datahub application to store and retrieve uploaded files.
MySQL and Amazon Redshift (PostgreSQL)	Relational database management systems available to CARTEEH researchers via the datahub.

Results

The CARTEEH datahub was developed using open-source technologies to ensure that its functionality could be extended in line with the growing needs of CARTEEH. Figure 4 shows an annotated screenshot of the CARTEEH datahub home page (https://carteehdata.org/). The basic functionality of the datahub involves exploring and searching datahub content (annotations 1 and 2); authenticating and registering users (annotation 3); and uploading and editing data (annotation 4).

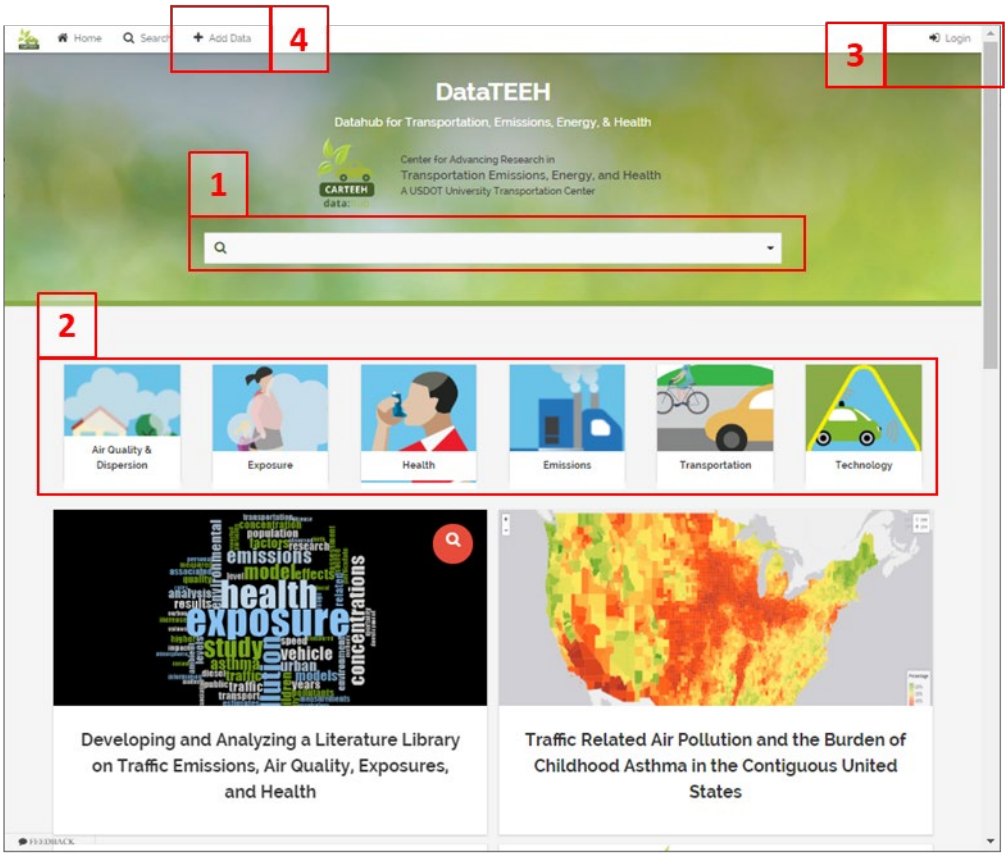


Figure 4. Annotated screenshot of the datahub home page: (1) search control; (2) controls to browse by topic; (3) authentication via login and registration views; and (4) controls to add or edit data.

Exploring and Searching Datahub Content

The CARTEEH datahub search interface allows users to perform data or metadata searches using a predetermined taxonomy of keywords and classifiers defined by the data (and information) within the datahub [22]. Results from text searches can be further refined by categories such as subject area, content type, author, or study type. Figure 5 shows the results of a CARTEEH datahub search. Relevant content is displayed in a paginated list with a user-defined title and a brief description. Clicking on the title of a content type takes the user to another tabbed view providing more information about the content (the second panel in Figure 5). Additional information is divided into the following categories:

- **Overview:** A brief full-text description of the data, including geographic coverage.
- **Files:** A list of all the electronic files that comprise the content in their original formats.
- **Metadata:** The metadata provided for the content (either when uploaded or subsequently edited).
- **Terms of Use:** A description of the terms of use or licenses associated with the content.
- **Reviews:** Comments on the content as provided by other registered CARTEEH datahub users.

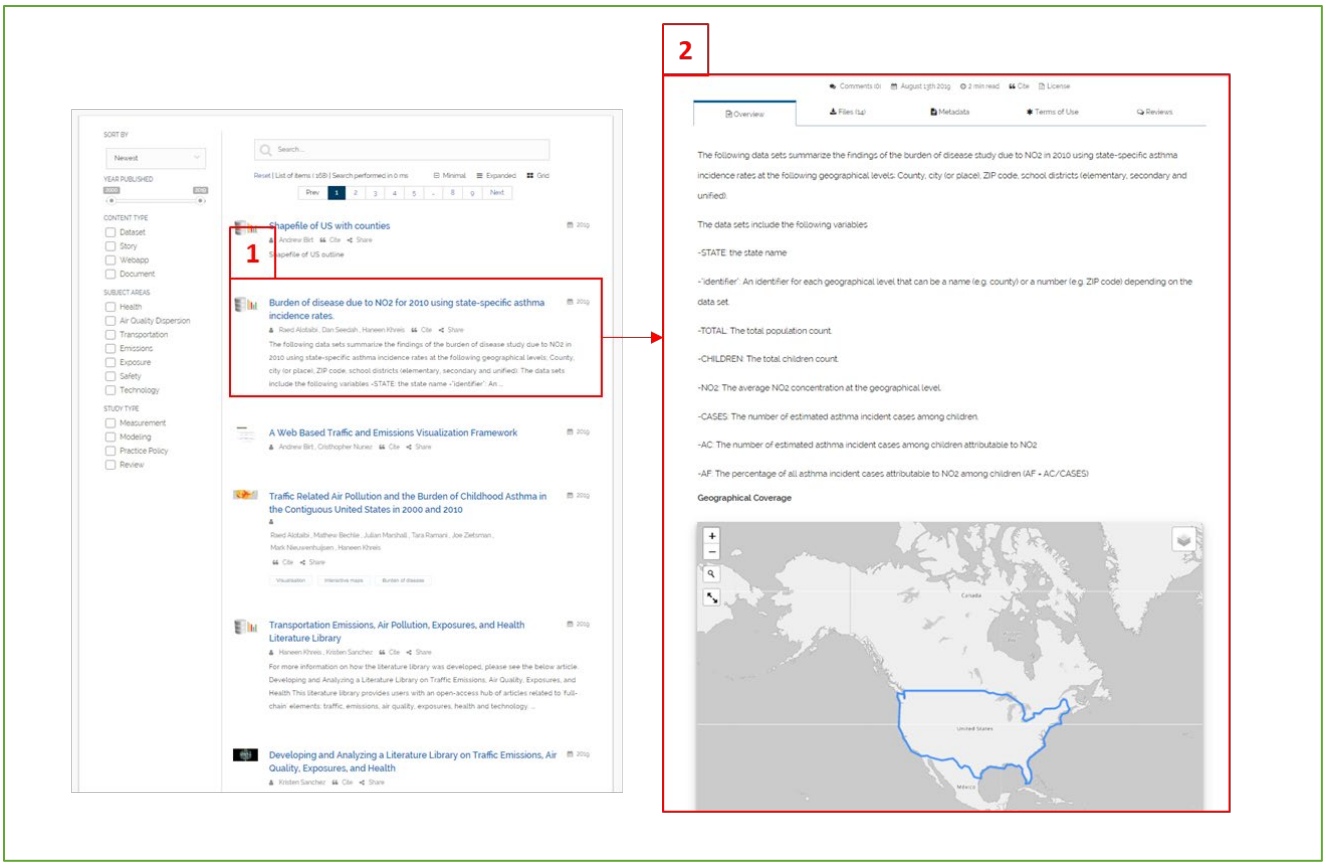


Figure 5. Annotated screenshot of CARTEEH datahub search and browse view: (1) relevant content listed in a paginated view; and (2) details of uploaded content.

Loading and Editing Content

Before content can be added to the datahub, users must register and log in (browsing or searching for content does not require authentication). Registration is designed to be as streamlined as possible (users provide an email, password, and first and last names, and then agree to terms and conditions of the datahub). After registering, users can log in using a standard login interface (which also includes password retrieval and profile edit functionality). As well as controlling who can upload data, authentication within the datahub is used to enable content owners to share their content with different groups of users: public (registered and unregistered users), all registered users, select registered users, or select unregistered users (via an explicit email invite).

The CARTEEH datahub allows registered users to create or upload one of three basic content types: **datasets**, **data stories**, or **webapps**. In abstract terms, these content types are containers that hold different types of information. Users choose the type of container based on their own classification of the type of information they wish to upload. The **dataset** content type is intended as a container for one or more datasets or other research products (presentations, analysis code, documents, videos, etc.). **Data stories** are complete study narratives or analysis ideas and are designed to contain a mix of file types. **Webapps** are dynamic visualizations of data created by explicitly linking code and analyses together. All content types also contain metadata. In addition to uploading electronic files (or providing links to an existing web-based resource), users are prompted to provide a title and free text description for the content, as well as metadata. The metadata forms use a controlled vocabulary to ensure consistency of search terms and tags within the datahub. The system also allows users to specify the geographical extents and granularity of their content, and to set terms of use of the content. Figure 6 shows screenshots of the forms (webpages) used to submit content to the datahub.

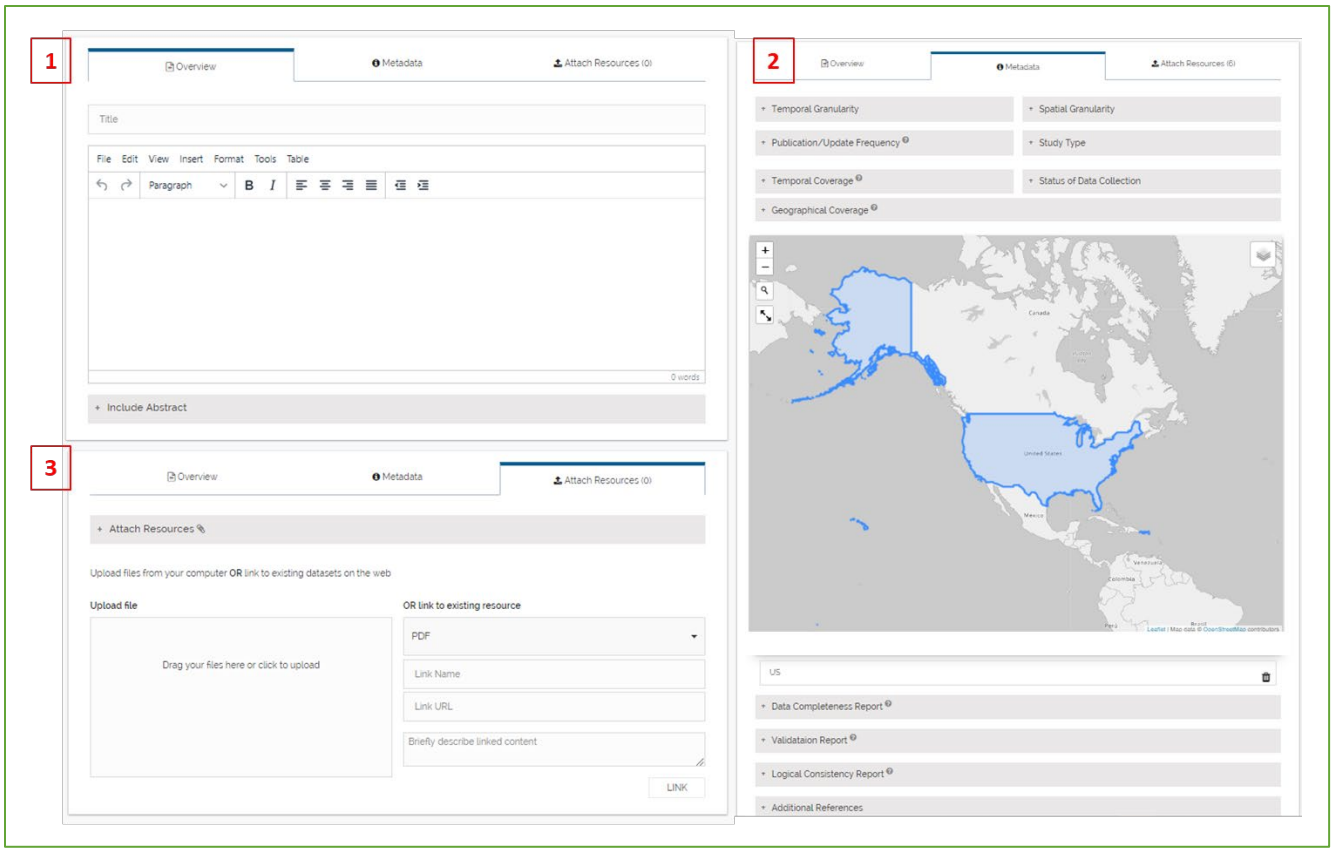


Figure 6. Annotated screenshots of the upload functionality: (1) the overview tab is used to title and describe the content; (2) the metadata tab enables a user to attach metadata; and (3) the resource view defines the location of data to be uploaded.

The datahub offers the following additional functionality during loading and editing tasks:

- **Data Reuse and Preservation**—Uploaders can use the metadata and content description fields to describe how others can reuse their datasets or other information. The datahub's preservation policy is defined on the preservation policy page and in the CARTEEH Data Management Plan, which is embedded in the datahub.
- **File Formats**—The datahub supports any file formats—including GIS, images (uncompressed and compressed), videos (uncompressed and compressed), and text files. Files are stored in their original formats.
- **Native GIS Data Processing**—The datahub includes native GIS support for GeoJSON and Google KML (keyhole markup language) files. It is also preloaded with country-, state-, and county-level GeoJSON geographical boundaries that users can select for their datasets. Drawing tools are available for users to define custom geographical boundaries.
- **Administrative Metadata**—The datahub captures metadata such as file upload size, date of deposit, and depositor (content owner). The datahub also records when content was created or last updated (accessible by administrators).
- **Messaging and Commenting**—Data depositors have the option of receiving email messages from individuals interested in learning more about their data by checking the "Allow Reader Emails" control on the content editing page. When this option is checked, a mail icon is shown on the public content page.

CARTEEH Data Stories

Data stories are intended as a way of linking multiple different content types into a single narrative or story about CARTEEH research or other activity (e.g., technology transfer, policy, education). The key idea of the data story is that linking different types of informational content into a single narrative will increase the contextual meaning of content such as data or analysis code. For example, a data story written to explain a particular study can link together one or more datasets, code that performs a particular analysis on the data, documents describing the study, and links to key references. Readers of the data story could then decide to download the code and data, for example, to rerun or modify an analysis, explore citations

used in the story, or simply use the narrative, along with complete descriptions, and data visualizations or analyses to better understand a study. The data story narrative also enables authors to link visualizations (e.g., graphs and maps) or code sections to further contextualize linked information or the narrative itself.

Figure 7 illustrates an example of a data story that explicitly links data and analysis code into a text-based narrative. The story outlines a study performed by a CARTEEH intern who developed models to explore how the randomness of traffic activity affects the temporal pattern of generated emissions. The study used software called Simulation of Urban Mobility (SUMO), an open-source traffic microsimulation model, to simulate traffic in a simplified network (signaled intersection). Second-by-second vehicle locations were extracted from the SUMO output files using Python code and were used to populate a MySQL database. The Environmental Protection Agency's MOtor Vehicle Emissions Simulation (MOVES) emission model was then used to generate a database of second-by-second emission rates. Finally, R code was written to calculate second-by-second emissions on the network based on the speed of each simulated vehicle and an appropriate emission rate. The data story shown in Figure 7 provides a narrative that offers a rationale for the study and an overview of the methods used to conduct the research (left-hand side of the figure). The narrative is intended to explicitly include references to data or code directly available for download from within the data story view. In Figure 7, the left-hand part of the panel shows the main view of the data story, which consists of formatted text, visualizations, and code snippets. The right margin provides links to datasets, code, or other content relevant to the data story. Links are shown to (1) the R code used to perform the analysis, and (2) a webapp that dynamically shows the traffic activities and associated emissions.

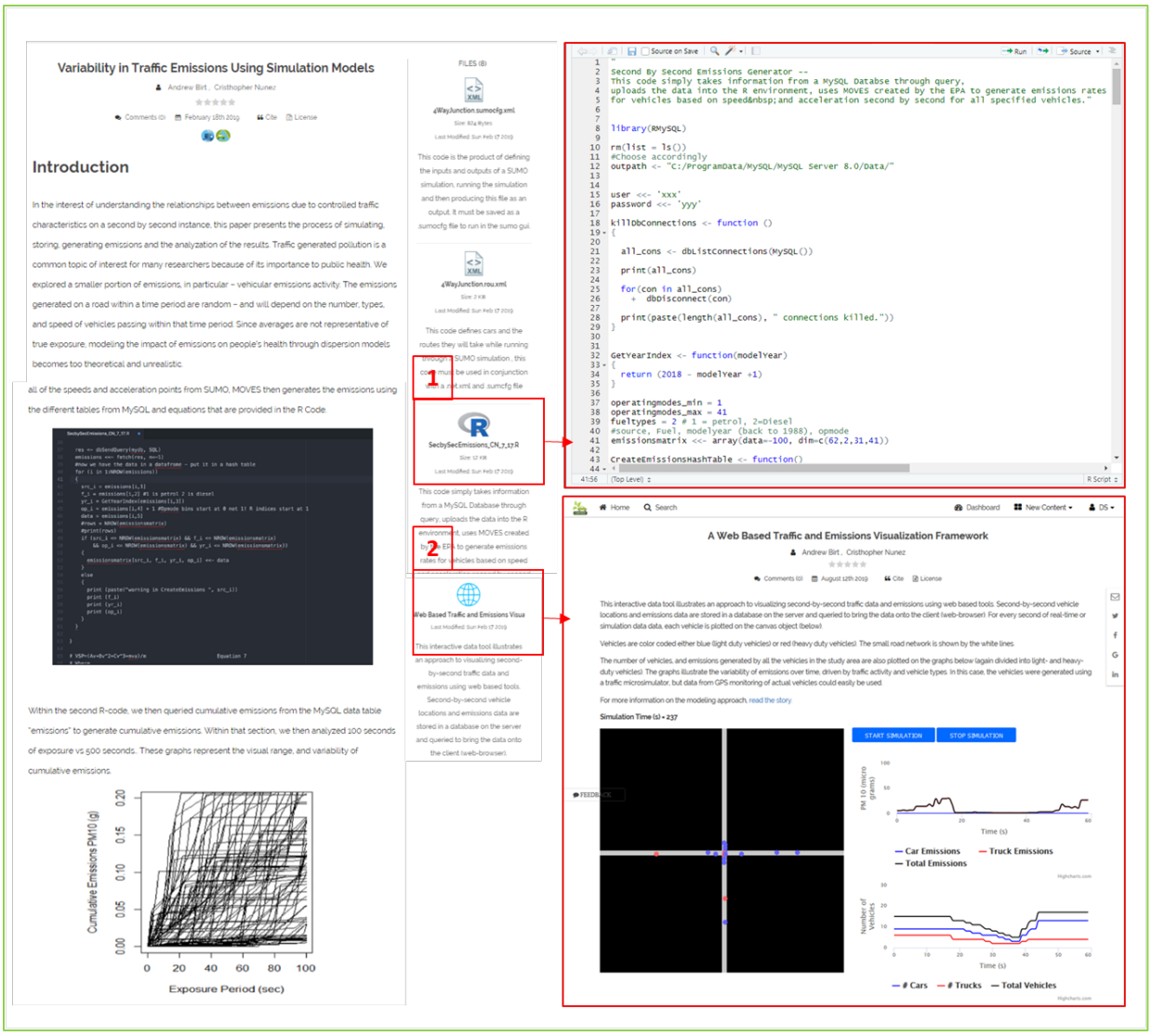


Figure 7. Example of a data story content type: (1) R code that accompanies the data story, and (2) a webapp content type linked to the data story.

CARTEEH Data Applications (Webapps)

CARTEEH webapps enable authors and programmers to create interactive data visualizations for the CARTEEH datahub. Figure 8 provides screenshots of a CARTEEH-funded study relating traffic pollutants to incidents of childhood asthma between 2000 and 2010 [23]. The webapp displays a county-level map illustrating the percentage of childhood asthma cases estimated to be attributed to different traffic-related pollutants. The map can be adjusted to show asthma cases for one of three modeled pollutants (NO_2, $PM_{2.5}$, and PM_{10}), can be panned and zoomed, and shows the data for any county in a pop-up dialog box. The interactive tool also links to the original dataset, which can be downloaded, explored, or reused.

The datahub currently contains built-in functionality to develop webapps using HTML5, JavaScript, and CSS. However, webapps developed in applications such as R, Python, Tableau, or Microsoft PowerBI can also be embedded as a CARTEEH webapp. The datahub architecture also includes the ability to generate bespoke MySQL databases (schema) that can be used to store large, highly structured datasets for the purpose of advanced visualizations and data retrieval.

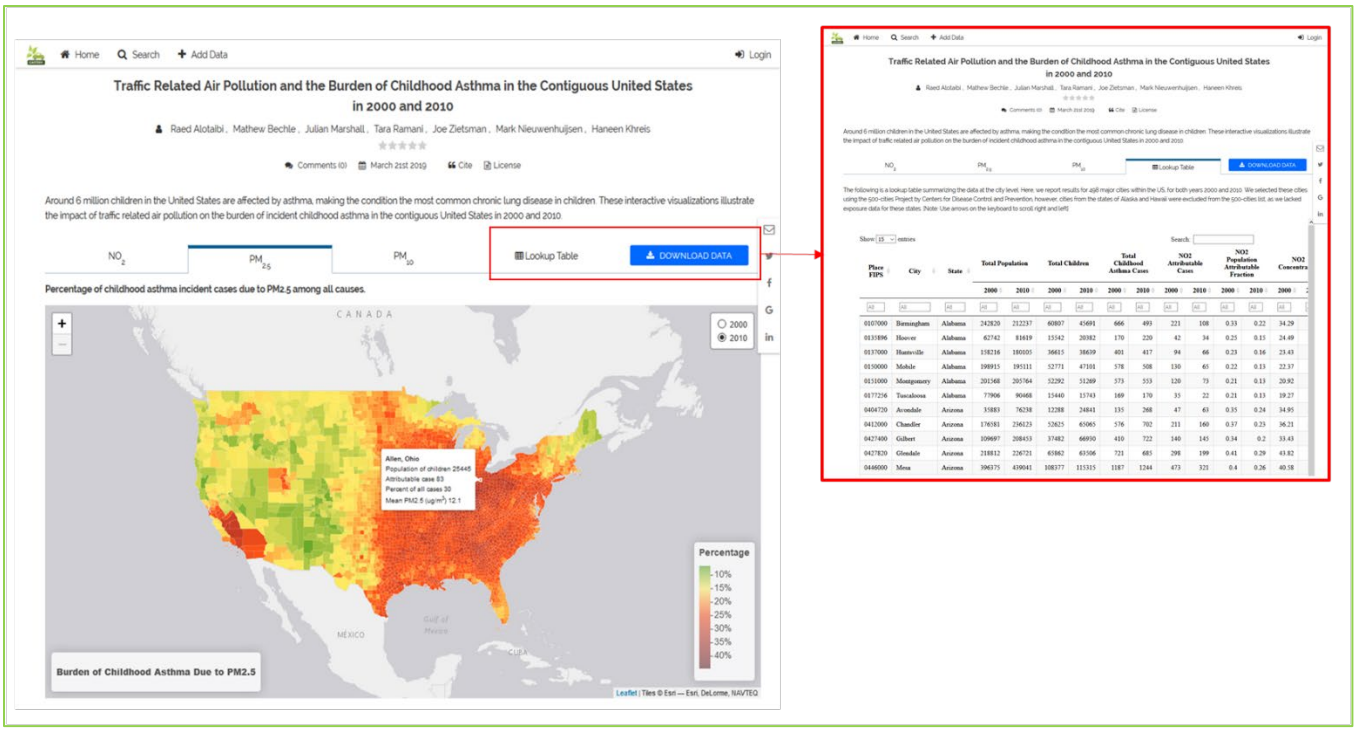


Figure 8. Annotated screenshot of webapp content illustrating how the data can be downloaded directly from the webapp.

Conclusions

This report summarizes the rationale and design of an IMS developed for a unique, collaborative research center focused on transportation emissions, energy, and public health. CARTEEH research involves collaborations among researchers from multiple domains, including clinical health, epidemiology, toxicology, civil engineering, and environmental science. The CARTEEH datahub is designed to support and enhance these collaborations and promote more efficient and impactful research. To achieve this, we developed an IMS that CARTEEH researchers can use to store, organize, and communicate datasets and other research products. The CARTEEH datahub has been active for less than a year, and there are currently over 1000 contributed research products within the datahub. CARTEEH is just nearing its first complete research cycle, and the content of the datahub is expected to grow considerably as researchers begin to understand the collaborative value of sharing and communicating their research.

The datahub's design principles and goals are founded on encouraging researchers to communicate all products of their research to their peers. The CARTEEH datahub currently allows researchers to do this by providing rich metadata descriptions and through data stories and webapps. Connecting data, information, and knowledge-based research products is essential for ensuring research products retain meaning and value through time. For this to occur, researchers will be required to proactively communicate existing research products to their peers, and the datahub is designed to be a tool that will help this process. This approach of proactively communicating research rebalances existing conventions and perceptions of data reuse, where often a researcher must work hard to find existing data for novel research. In a truly collaborative environment, equal responsibility should be placed on effectively communicating the meaning and value of existing research products to other researchers.

The collaborative, communicative approach will require a different mindset for most researchers. Generally, researchers are trained to share the end products of their research—such as peer-reviewed papers or reports. Sharing other intermediate research products is usually considered to be less important. However, this view is changing. Piwowar [24] suggested that funding agencies are increasingly interested in all the research products of a potential researcher—not just peer-reviewed publications—and that in the future, it will become routine for researchers to track and value a wider range of metrics for impactful research. Initiatives such as open science [25] and science 2.0 [26, 27] are also gaining traction. Donaho noted that due to the complexity of modern research, it is often difficult to fully communicate research methods

and findings in traditional research articles and suggested that often, such articles serve only as advertisements for the true scholarship involved in research [28]. To effectively communicate research, modern science will require information management tools such as the CARTEEH datahub that are capable of maintaining transparent and explicit linkages among data, information, and knowledge generated through research. These linkages will be useful not only in the context of scientific replicability but also to help researchers understand the research methods of peer scientists.

Outputs, Outcomes, and Impacts

CARTEEH's research focuses on understanding the links between transportation and health. The center's work involves collaborations among researchers with a wide variety of expertise in the areas of clinical health, epidemiology, toxicology, civil engineering, and environmental science. The CARTEEH datahub supports and enhances these collaborations by allowing its research scientists to share data, models, and ideas. Using the datahub, researchers will be able to acknowledge and understand the types of data generated by other research fields, as well as the models used to analyze them. The datahub not only supports the immediate data storage and organizational needs of CARTEEH but also helps generate new research ideas and increase the cost effectiveness of data management by promoting reuse.

Outcome

This project resulted in a web-based platform that underpins research activities undertaken within CARTEEH. It allows CARTEEH and non-CARTEEH researchers to explore the types of data available for transportation-health research and thus helps promote effective exploration in these areas. The code and methods used to develop the datahub, along with its broad vision as a repository for not only data but also models and ideas, will be useful for other organizations seeking to develop effective, research-oriented data management practices.

Impact

The long-term goal is for the datahub to be used to organize and share data, models, and ideas among transportation and health researchers. This ongoing self-organization of materials will promote data reuse and encourage the exchange and harmonization of research techniques and models used by the different research areas.

Research Outputs, Outcomes, and Impacts

In January 2019, we presented at two Transportation Research Board (TRB) committee meetings during the TRB's annual meeting in Washington, DC:

- Transportation and Air Quality, ADC20—January 14, 2019.
- Artificial Intelligence and Advanced Computing Applications, ABJ70—January 15, 2019.

We also presented the datahub at the CARTEEH Symposium on February 18, 2019. A copy of the presentation is available on the datahub [29].

References

[1] A. L. Goodkind, C. W. Tessum, J. S. Coggins, J. D. Hill, and J. D. Marshall, "Fine-scale damage estimates of particulate matter air pollution reveal opportunities for location-specific mitigation of emissions," *Proc. Natl. Acad. Sci.*, vol. 116, no. 18, pp. 8775–8780, 2019.

[2] "TDL Data Management Working Group Report," 2015.

[3] "Welcome to DataCite." [Online]. Available: https://datacite.org/. [Accessed: 21-Aug-2019].

[4] "Create dataset DOI—DOIs—API Reference—data.world for developers—Build data.world apps." [Online]. Available: https://apidocs.data.world/api/dois/adddoi. [Accessed: 21-Aug-2019].

[5] "CKAN hosting guidelines—ckan/ckan Wiki—GitHub." [Online]. Available: https://github.com/ckan/ckan/wiki/CKAN-hosting-guidelines. [Accessed: 21-Aug-2019].

[6] "GitHub—NaturalHistoryMuseum/ckanext-doi: CKAN extension for assigning a DOI to datasets." [Online]. Available: https://github.com/NaturalHistoryMuseum/ckanext-doi. [Accessed: 21-Aug-2019].

[7] "Pricing—data.world." [Online]. Available: https://data.world/pricing. [Accessed: 21-Aug-2019].

[8] "Database Dumps—CKAN Data Management System Documentation 1.7.4 documentation." [Online]. Available: https://docs.ckan.org/en/ckan-1.7.4/database-dumps.html. [Accessed: 21-Aug-2019].

[9] "Preparation—Dataverse.org." [Online]. Available: http://guides.dataverse.org/en/latest/installation/prep.html. [Accessed: 21-Aug-2019].

[10] "OAuth Login: ORCID, GitHub, Google—Dataverse.org." [Online]. Available: http://guides.dataverse.org/en/latest/installation/oauth2.html. [Accessed: 21-Aug-2019].

[11] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *J. Inf. Sci.*, vol. 33, no. 2, pp. 163–180, 2007.

[12] A. G. Birt, A. Calixto, M. Tchakerian, A. Dean, R. N. Coulson, and M. K. Harris, "Harnessing information technology (IT) for use in production agriculture," *J. Integr. Pest Manag.*, vol. 3, no. 1, pp. 1–8, 2012.

[13] M. Frické, "The knowledge pyramid: a critique of the DIKW hierarchy," *J. Inf. Sci.*, vol. 35, no. 2, pp. 131–142, 2009.

[14] W. Weaver, "Recent contributions to the mathematical theory of communication," *ETC a Rev. Gen. Semant.*, pp. 261–281, 1953.

[15] J. D. Adams, G. C. Black, J. R. Clemmons, and P. E. Stephan, "Scientific teams and institutional collaborations: evidence from US universities, 1981–1999," *Res. Policy*, vol. 34, no. 3, pp. 259–285, 2005.

[16] J. Adams, "Collaborations: the fourth age of research," *Nature*, vol. 497, no. 7451, p. 557, 2013.

[17] H. Khreis, K. de Hoogh, and M. J. Nieuwenhuijsen, "Full-chain health impact assessment of traffic-related air pollution and childhood asthma," *Environ. Int.*, vol. 114, pp. 365–375, 2018.

[18] D. C. Pelz and F. M. Andrews, "Scientists in organizations: productive climates for research and development," 1966.

[19] L. L. Hargens, "Relations between work habits, research technologies, and eminence in science," *Sociol. Work Occup.*, vol. 5, no. 1, pp. 97–112, 1978.

[20] K. J. Boudreau *et al.*, "A field experiment on search costs and the formation of scientific collaborations," *Rev. Econ. Stat.*, vol. 99, no. 4, pp. 565–576, 2017.

[21] M. Stember, "Advancing the social sciences through the interdisciplinary enterprise," *Soc. Sci. J.*, vol. 28, no. 1, pp. 1–14, 1991.

[22] D. Tunkelang, “Faceted search,” *Synth. Lect. Inf. Concepts, Retrieval, Serv.*, vol. 1, no. 1, pp. 1–80, 2009.

[23] R. Alotaibi *et al.*, “Traffic related air pollution and the burden of childhood asthma in the contiguous United States in 2000 and 2010,” *Environ. Int.*, vol. 127, pp. 858–867, 2019.

[24] H. Piwowar, “Value all research products,” *Nature*, vol. 493, no. 7431, p. 159, Jan. 2013.

[25] M. Woelfle, P. Olliaro, and M. H. Todd, “Open science is a research accelerator,” *Nat. Chem.*, vol. 3, no. 10, p. 745, 2011.

[26] B. A. Nosek *et al.*, “Promoting an open research culture,” *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015.

[27] B. Shneiderman, “Science 2.0,” *Science*, vol. 319, no. 5868, pp. 1349–1350, 2008.

[28] D. Donoho, “50 years of data science,” in *Tukey Centennial Workshop*, 2015, pp. 1–41.

[29] “Developing the carteeh datahub.” [Online]. Available: https://carteehdata.org/library/document/developing-the-carteeh-datahub.