Pacific Northwest
NATIONAL LABORATORY

PNNL-31960 Rev 0
DVZ-RPT-066 Rev 0

Evaluation of Data Catalog Software for Hanford Site Environmental Datasets

September 2021

Kenneth D Ham
Dustin L Crockett

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-31960 Rev 0
DVZ-RPT-066 Rev 0

Evaluation of Data Catalog Software for Hanford Site Environmental Datasets

September 2021

Kenneth D Ham
Dustin L Crockett

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

Environmental information and data underpin achievement of the U.S. Department of Energy (DOE) Office of Environmental Management (EM) mission at the Hanford Site. The Hanford Environmental Data Management (HEDM) Program is the DOE Richland Operations Office (RL) approach to develop and implement a formal program for managing environmental data and the associated records, materials, and systems at the Hanford Site. The current project, contract, organization, and contractor-specific efforts at managing environmental data sets are insufficient to provide orderly, long-term, site-wide access. A vital element to be created within the HEDM program plan is a catalog of data sources, called the Hanford Environmental Information and Data Index (HEIDI), that will enable long-term access and retrievability for the multiple independent sources of data that might otherwise be difficult to discover.

This report compares leading open source and commercial data catalog platforms using criteria to assess the functionality needed to develop the HEIDI catalog of Hanford data sources that connects and exchanges data with established Hanford Local Area Network (HLAN) enterprise information technology systems. Proprietary platforms evaluated included ArcGIS Enterprise Sites, Junar, OpenDataSoft, and Socrata, and non-proprietary platforms included Energy Data eXchange (EDX), Comprehensive Knowledge Archive Network (CKAN), and DKAN (a Drupal-based open data portal based on CKAN). Capabilities supporting data discoverability, retrieval, and archival, as well as metadata standard requirements and integration into the HLAN were rated as either failing to meet requirements (F), meeting requirements (M), or exceeding requirements by delivering additional desired features (E). The lowest rating for any capability area was assigned as the overall rating for the platform. These findings enable DOE-RL and the contractors implementing the HEDM plan to focus on candidate tools likely to meet the requirements for implementing HEIDI.

All of the platforms receiving an overall rating of ‘F’ were unable to be deployed on Hanford infrastructure or within dedicated cloud resources. A propriety software-as-a-service (SaaS) model of delivering a data catalog (e.g., found in software such as Junar and OpenDataSoft) favors consistency across customers at the expense of customization and configurable roles that are needed for Hanford work. Hosting data on a shared commercial platform places limits on dataset size (maximum of 240 Mb for OpenDataSoft), a significant limitation for HEIDI implementation. EDX, a government data catalog based on CKAN, received the ‘F’ rating due to an inability to incorporate authentication from HLAN into the system.

Among platforms rated ‘M’ or ‘E’, only the Socrata platform had a SaaS delivery model. In contrast to other SaaS platforms, Socrata provided custom roles and gateways that allow local datasets to be incorporated into an online catalog. Socrata also complies with the Federal Risk and Authorization Management Program, a significant benefit for cloud-based management of Hanford data. The other platforms rated ‘M’ or ‘E’, ArcGIS Enterprise Sites, CKAN, and DKAN, provide fully self-hosted options, allowing for greater control and flexibility with the HEIDI catalog. These widely used tools have supportive communities of practice, extensive customization options, and demonstrated deployments that provide evidence that they can meet requirements, often deliver additional desired features, and work well with federal government systems. Completely customized alternatives built on a collection of applications were not evaluated because achieving similar performance to CKAN or DKAN requires substantial resources, especially in the absence of the active communities that have grown to support these tools.

ArcGIS Enterprise Sites, Socrata, CKAN, and DKAN were evaluated as strong candidates for successful implementation with HEIDI. Further evaluation is needed to assess licensing terms, costs of acquisition, maintenance requirements, and sustainability through long-term stewardship at Hanford.

Acknowledgments

We acknowledge the insightful inputs on Hanford Site environmental data management from Bill Webber, Joanne Rieger, and Len Habel of the Central Plateau Cleanup Company and Ken Kapsi of the DOE Richland Operations Office. We thank Jason Altman of Hanford Mission Integration Solutions for educating us on geographic information management at the Hanford Site and the software and hardware infrastructure underlying those capabilities.

Funding for this work was provided by the U.S. Department of Energy Richland Operations Office under the Deep Vadose Zone – Applied Field Research Initiative. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the Department of Energy under Contract DE-AC05-76RL01830.

Acronyms and Abbreviations

AD	Active Directory
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
CSA	Cloud Security Alliance
CSDGM	Content Standard for Digital Geospatial Metadata
DCAT	Data Catalog Vocabulary
DKAN	a Drupal-based open data portal based on CKAN
DOE	U.S. Department of Energy
EDX	Energy Data eXchange
EM	Office of Environmental Management
FedRAMP	Federal Risk and Authorization Management Program
FEMA	Federal Emergency Management Agency
FGDC	Federal Geographic Data Committee
HEDM	Hanford Environmental Data Management
HEIDI	Hanford Environmental Information and Data Index
HHS	U.S. Department of Health and Human Services
HLAN	Hanford Local Area Network
HMIS	Hanford Mission Integration Solutions
INSPIRE	A metadata format required by the European Union
ISO	International Organization for Standardization
NQAP	Nuclear Quality Assurance Program
REST	REpresentational State Transfer
RL	Richland Operations Office
SaaS	software-as-a-service
USDA	U. S. Department of Agriculture

Contents

Figures

Tables

1.0 Introduction

Environmental information and data underpin the U.S. Department of Energy (DOE) Office of Environmental Management (EM) mission at the Hanford Site. The Hanford Environmental Data Management (HEDM) Program is the DOE Richland Operations Office (RL) approach to develop and implement a formal program for managing environmental data and the associated records, materials, and systems at the Hanford Site. A vital element of the HEDM program plan is a catalog of data sources, called the Hanford Environmental Information and Data Index (HEIDI), that will enable long-term access and retrievability for the multiple independent sources of data that might otherwise be difficult to discover.

The mission of the Hanford Site involves near-term environmental restoration and long-term stewardship, so long-term maintenance of environmental data is a priority (Atz et al. 2020). Software becomes obsolete over time, so it is important to select a data management alternative that is likely to be updated and maintained for the foreseeable future. Support of shared standards is also desirable to facilitate a transition of the catalog to another tool if the need arises. Licensing terms and costs can also influence whether it is feasible to implement a software solution for an extended number of years, as will the costs of maintaining the software. While it is not known what the future will bring, the software candidates for this evaluation were selected because they are currently in broad use and are well supported commercially or through the open-source community.

This report focuses on the suitability of available software tools to provide the critical functions of cataloging, storing, retrieving, and capturing metadata for data sources, and managing data access. Because data transparency is of keen interest in many countries, particularly for government data, a software industry has developed to provide data catalog platforms to the open data portal market (Schrack 2021; Grzenda and Legierski 2021). These data catalog software platforms provide many of the functions needed to implement HEIDI, even though it is not envisioned as an open data portal. The ancillary functions of the platform become important in determining whether that software can handle the types of data found at the Hanford Site and whether interoperability with other systems in the Hanford Local Area Network (HLAN) environment is feasible.

This report evaluated leading open source and commercial data catalog platforms to assess whether they possess the functionality needed to catalog Hanford data sources and connect and exchange data with established HLAN enterprise systems. This initial evaluation allows DOE-RL and the contractors implementing the HEDM plan to weigh the costs and benefits of these candidate tools for HEIDI implementation.

2.0 Background

Managing environmental data and information to support the DOE Office of Environmental Management (EM) mission at large and complex sites such as Hanford is challenging, not only because of the variety of data and information collected over decades by different contractors and stakeholders, but also because it must be managed to be readily accessible and retrievable over the long term. Hanford Site contractors currently use multiple, independent approaches to manage the ongoing acquisition of environmental data while accommodating decades of project, organization, and contractor-generated environmental data. Users and stakeholders have difficulty finding and accessing the data they need because there is no sitewide catalog of data sources. Without such a tool, it is challenging to identify the dataset integrity in terms of completeness and accuracy

The 2017 HEDM assessment report (CHPRC-03503 2017) states that the current environmental data systems will become increasingly incapable of providing the environmental information needed to support key remedial decisions. Continued reliance on the existing interrelated environmental data systems and records will dramatically increase the costs for DOE and reduce the effectiveness of this approach. DOE has established the HEDM Program to address the management of environmental data at Hanford in a more integrated and systematic manner. The HEIDI catalog of all Hanford Site environmental data sources is a key component of the program plan.

HEIDI is envisioned as a tool that makes Hanford Site environmental data findable, accessible, interoperable, and reusable (FAIR) by implementing the principles of Wilkinson et al. (2016). Achieving these principles is expected to reduce the costs of finding and using data and to increase the ability of existing environmental data to support key remedial decisions. The value of an effective data catalog extends beyond characterization and remediation phases to support long-term stewardship of the site.

3.0 Data Catalog Requirements and Desired Features

The basic functions of a data catalog are cataloging, finding, and retrieving datasets for examination and use. Within those functions, there are numerous capabilities that can improve the user's ability to interact with the datasets and allow data catalog software to integrate into existing workflows more easily. The implementation of the data catalog can be simplified by a software platform that is easy to use, integrates well with existing systems, and gives administrators the ability to secure datasets from unauthorized access or modification. Some of those capabilities are required, while others are desired features. This section first describes relevant terminology, then identifies requirements and desired features related to catalog functions and implementation in the HLAN environment.

3.1 Data Catalog Terminology

The concept of a data catalog is deceptively simple, but the details can be critically important to the usefulness and functionality of such a system. To help keep those details in clear focus, terms are defined in the table below to aid in a discussion of functional requirements.

Table 3.1. Data Catalog Terms Defined

Term	Definition
Catalog	A collection of catalog records
Catalog record	Metadata for a dataset. The catalog record contains a link to the dataset.
Controlled vocabulary	The list of acceptable values for a metadata element
Dataset	A collection of information in a machine-readable format
Hosting	A data catalog must be hosted to become available on a network. Hosting can be commercial, self -hosted, or free.
Machine-readable	A format that can be processed by a machine, without human intervention
Metadata	Information to describe multiple aspects of a dataset
Metadata schema	A structured set of data elements that describe a dataset
Naming authority	A source of acceptable names for a data element
Open data	Data that can be freely used, reused, and shared

3.2 Catalog Functions

To provide access to datasets, data catalog software must manage a set of records to enable users to find datasets of interest and to provide a way to retrieve the dataset for a selected catalog record. How those functions are implemented is important for providing value to users.

3.2.1 Catalog a Dataset

A catalog record is created by compiling the metadata for a dataset, including a link to the dataset. The catalog record is the basis for all other functions. When metadata entries include rich, consistent descriptions of datasets and their contents, users can apply search criteria to find the appropriate data for their intended use. Standards for the metadata are important to maintain consistent structure and values across records, even when data comes from different organizations.

3.2.1.1 Metadata Standards

Several metadata standards have found favor for various fields of study or in various regions of the world. In the U.S., Executive Order No. 12906 requires all federal agencies and organizations receiving federal funds for the creation of geospatial data to document that data using a metadata standard endorsed by the Federal Geographic Data Committee (FGDC). FGDC-endorsed metadata standards include the FGDC-developed Content Standard for Digital Geospatial Metadata (CSDGM), Version 2 (FGDC-STD-001-1998), as well as the ISO 19XXX series of standards. Any data catalog software used for Hanford environmental data should be capable of importing and exporting metadata in a format endorsed by the FGDC. The HEIDI catalog itself may operate on a superset of this format, incorporating additional fields where needed to support full functionality for the Hanford Site mission.

Requirement: Support FGDC-endorsed metadata standard(s).

3.2.1.2 Metadata Creation Wizard

Ideally, the catalog software will provide a human-friendly interface to enable users to create a machine-readable catalog entry that is complete and consistent. Consistency with metadata standards is a start, but it will be necessary to establish and enforce additional conventions, such as controlled vocabularies and name authorities. The consistency that results from these efforts will make searching the catalog for datasets more effective. In the case of Hanford Site environmental data, for example, it will be important to follow naming conventions for spatial objects such as operating units, wells, buildings, and so on that have been established by the Geospatial Information Technology Services group, which is part of Hanford Mission Integration Solutions (HMIS), the current holder of the Hanford Mission Essential Services Contract at the Hanford Site. The benefits of such efforts will be evident when the data catalog is integrated into other data systems on the HLAN. It may be necessary to develop a customized metadata creation wizard to enable additional metadata elements and to enforce acceptable entries consistent with Hanford Site naming conventions. The option to validate metadata to a standard or even to custom entry lists would be useful for maintaining consistency within the catalog and for connections to other data systems.

Requirement: Metadata validation tool for standard schemas.

Desired Feature: Metadata validation tool customizable for Hanford-specific schemas.

Desired Feature: Metadata creation wizard with customizable elements and name authorities.

3.2.1.3 Standards-based Metadata Import and Export

To minimize manual data entry and transformation, the data catalog software should make it easy to import and export metadata in standard schemas. A schema defines the data elements and how they are used to describe a resource that is the subject of the metadata.

Importing and Exporting Metadata

Some datasets will have a standards-based metadata entry available to load or access via the Application Programming Interface (API). The data catalog software should allow population of the catalog record through such importable information to the extent possible, with missing fields made evident for manual entry.

It may be useful to import or export metadata to standard schemas, by converting between native formats and standard schemas.

Desired Feature: Import and export metadata in popular standard schemas.

Metadata Transfer via API

An API is a structured way to transfer information between the data catalog and other software. Among many possible functions, data catalog software often includes APIs that provide documented ways to share datasets, including sharing metadata. Given the number of datasets and variety of sources found at Hanford, this functionality is needed to enable datasets to be integrated from a variety of organizations and software platforms. For transfer of metadata or data from the catalog, REpresentational State Transfer (REST) APIs are preferred to enable these links to occur over HTTP protocols and standardize the methods used to make the requests.

Requirement: API connectivity for input and output of datasets, including metadata.

Desired Feature: REST APIs.

3.2.1.4 Archive Dataset Snapshot

A special function for the Hanford data catalog is to archive a snapshot of a dataset to preserve the data queried at a specific time in support of analysis and reporting that leads to a deliverable. While data catalogs are typically focused on providing the most current version of a dataset, Hanford Site deliverables need to be associated with the version of the database used to support the deliverable. The ability to archive the snapshot of a dataset is needed for quality requirements associated with reporting.

Requirement: Software must support the archiving of multiple dataset snapshots to support reporting of quality requirements.

3.2.2 Find a Dataset

Another elementary function of a catalog is the ability to find a dataset. Finding a dataset relies on imposing selection criteria on the metadata fields of catalog records. As mentioned earlier, consistency of metadata structure and content is key to effective searching.

3.2.2.1 Filter Criteria

A list of datasets can be narrowed by applying criteria to values contained in metadata fields. For example, a user may request a list of datasets for a date range of interest, and a particular contaminant measured in groundwater. Careful selection of criteria can deliver a focused list of datasets.

Requirement: Search options that provide result lists built from filtering metadata field entries.

3.2.2.2 Interactive Map Selection of Datasets

Another way to limit a selection of datasets is to interact with a map of points, lines, or polygons. Those spatial objects, in the case of the Hanford Site, might indicate items such as wells, tanks, buildings, waste sites, or operable units. By selecting those objects, or an area containing objects, it should be possible to obtain a list of datasets that contain relevant data.

Requirement: Search for dataset by selecting spatial objects or geographically defined areas.

3.2.2.3 Index on Metadata Values

Data catalog performance depends in large part on the speed of metadata queries. Query speed depends on indexing of metadata fields to enable the catalog to be quickly partitioned to select a relevant list of datasets. Searching based on dataset content (versus the metadata) is not expected to be practical.

Desired Feature: Indexed, rapid searching through catalog entries by metadata terms.

3.2.3 Retrieve Data Set

Another elemental function of a data catalog is to retrieve datasets that may fit the intended use. When a search returns a list of one or more datasets based on metadata fields, the user will often be interested in exploring whether the data is suitable for the intended use and then downloading the content of the dataset. Data can be stored locally to the catalog or federated from other systems through API links. This distinction should be of little importance to the end user.

Some of the possible ways of interacting with the dataset are described below, with R and D used as shorthand to indicate a Requirement or a Desired feature, respectively, if the dataset supports such an interaction.

- View the dataset on a map (**R**)
  - Interact with the data on the map, selecting areas or sites, making values visible upon mouseover, etc. (**R**)
  - Export the map and displayed data to an image (**D**)
- View the dataset as a table (**R**)
- Export the dataset to a file, either in its native format or a common file type such as CSV or Excel (**R**)
- Identify the API link to the dataset metadata for use in other software (**R**)
- View metadata (**R**)
  - Examine appropriateness for the intended use
    - Quality (**R**)
    - Timeframe (**R**)
    - Spatial extent (**R**)
- Identify related datasets (**D**)
- Graph the data (**D**)
- Perform simple data analyses, providing basic statistics such as means and ranges (**D**)
- Data extraction and filtering (**D**)
  - Limiting columns (**D**)
  - Filtering columns or rows on values (**D**)

3.2.4 Large Dataset Storage

Environmental datasets can grow to large file sizes. Incorporating the content of those files into an active database is impractical, so the catalog should provide links to large datasets, with the understanding that such data may only provide limited opportunities for mapping, plotting, or filtering within the user interface of the data catalog.

Requirement: For large datasets, catalog entries will link to file locations.

3.3 Implementation on HLAN

HEIDI is planned to be a catalog of all environmental datasets available for the Hanford Site. As such, HEIDI will be implemented on the HLAN to allow seamless data transfer interactions with other applications managing Hanford environmental data through APIs. Hanford data is collected by multiple contractors (who may change over time), so interoperability will be an essential part of any successful data catalog. For manageability, the data catalog should be implemented as a dedicated website (i.e., not as part of an existing website), available for linking from other Hanford sites. This frees the catalog from restrictions and logistical challenges that may arise from linkages with elements that follow a different development cycle. Requirements and features related to using a data catalog in the HLAN environment are described in the sections below.

3.3.1 Site Software Approval

Any software implemented on HLAN must complete a software approval process that involves grading the software to determine the required level of risk management review and precautions needed in case of a failure. Ideally, potential impediments to approval should be identified in the initial evaluation process.

Requirement: Catalog software must run on a Hanford-approved operating system (Linux, macOS, Unix, or Windows) and have no characteristics that would disqualify it for use on the HLAN.

3.3.1.1 Self-Hosted

Catalog solutions implemented as software-as-a-service (SaaS) typically host the catalog and datasets on commercial cloud servers. While there are advantages to such an approach, the proprietary approach common to all customers accessing the commercial cloud presents additional challenges for the customization, interoperability, and security that are necessary for integration into the HLAN. Solutions that can be implemented in dedicated HLAN cloud infrastructure are not necessarily subject to the same challenges. This evaluation assumes that solutions that cannot be hosted within the HLAN infrastructure, whether that be on hardware or dedicated cloud, are not viable at this time.

Requirement: Catalog software must be able to be hosted on HLAN hardware or HLAN dedicated cloud.

3.3.2 Single Sign-On

Incorporating the catalog software into HLAN workflows will be much more straightforward if it supports authentication using credentials already created for users. This capability also facilitates assigning user roles for multiple data systems.

Requirement: Support for HLAN single sign-on.

3.3.3 Role-Based Access Control

Roles enable catalog administrators to grant permissions to users according to how they will be interacting with the catalog. Although the names and responsibilities vary among software solutions, typical roles include:

- Catalog administrator
- Dataset publisher
- Dataset editor
- Dataset viewer
- Dataset API consumer

Role-based access control organizes the many users expected to access the data catalog into a manageable number of groups (roles) that share the same permissions.

Requirement: Support for role-based access control.

Desired Feature: Customizable roles.

3.3.4 Linking to Actively Maintained Hanford Databases

Hanford environmental datasets are actively maintained and updated, which creates a need to keep the catalog up to date. Two options exist for incorporating an actively managed dataset into a data catalog:

- Mirror the dataset: Create a copy and set a schedule for updating the copy (**Requirement**).
- Federation: Link to the data through APIs so that the metadata and data are always up to date (**Desired Feature**).

The efficacy of each approach will vary by the size and frequency of updates for a dataset, but the ability to support federation can ease the burden of keeping the linked data fresh.

3.3.5 Deployment

For ease of deployment on the HLAN, the data catalog software must fit easily into existing HLAN workflows and procedures. Flexibility of configuration is one way to minimize custom coding while achieving the desired integration and interoperability. Where customizations are required to deliver catalog functions or facilitate interoperability, it should be possible to complete them with familiar programming languages.

Requirement: Customizations using programming languages commonly used on other Hanford software.

Desired Feature: Flexible, straightforward configuration, minimizing requirements for custom coding.

4.0 Candidate Data Catalog Software Products

A select group of data catalog software products likely to provide the capabilities needed for cataloging Hanford Site environmental data and information were evaluated. The following sections provide some details on each solution, show examples of the user interface of a data catalog implemented using the solution, and rate the solution relative to its ability to deliver the requirements and desired features described in Sections 3.2 and 3.3. Each platform was rated for its ability to deliver the requirements and desired features within each major topic areas identified by the subsections in Sections 3.2 and 3.3. The possible ratings were: failing to meet requirements (F), meeting requirements (M), or exceeding requirements by meeting requirements and delivering additional desired features (E).

The lowest rating for any major topic area was assigned as the overall rating for the platform. This overall rating identifies which candidate software platforms can meet the requirements for HEIDI implementation on HLAN. After screening out candidates that fail to deliver the required capabilities, the remaining candidates can be evaluated for other considerations that may drive a final selection.

4.1 ArcGIS Enterprise Sites

ArcGIS Enterprise Sites is the enterprise version of ArcGIS Hub, for which publicly available examples are numerous. Esri introduces their ArcGIS Enterprise Sites product by stating: "ArcGIS Enterprise Sites allows you to create a tailored web page experience for your users to help you share your portal's authoritative GIS data to other departments more easily, even if they are not used to working in your GIS. Instead of learning to navigate the portal and access groups, members can go directly to the custom web page you create to navigate the content relevant to them."[1]

While self-hosted implementations of ArcGIS Enterprise Sites are designed to deliver information within the enterprise, its capabilities largely mirror those of the SaaS offering called ArcGIS Hub (https://www.esri.com/en-us/arcgis/products/arcgis-hub/overview), which focuses on open data. Enterprise Sites is included in the licensing of ArcGIS Enterprise, with premium features (social media interaction) available at additional cost. Selected specifications are provided in Table 4.1.

[1] About ArcGIS Enterprise Sites, https://enterprise.arcgis.com/en/sites/latest/get-started/about-this-application.htm. Accessed 9/2/2021.

Table 4.1. ArcGIS Enterprise Sites Specifications

Specification	Result
Software	ArcGIS Enterprise Sites
Website	https://enterprise.arcgis.com/en/sites/
Hosting models	Self/on-premises or third party (SaaS available using ArcGIS Hub)
Vendor organization	Esri
Latest version	6 May 2021: ArcGIS Enterprise 10.9 17 July 2018: Initial availability in 10.6.1.
Browser support	Chrome; Firefox; Safari; iOS Safari; Chrome for Android
Operating system	Linux or Windows
Customization programming languages	Python through the ArcGIS REST API for configuring the catalog. HTML and CSS for web page customization.
License	Included in ArcGIS Enterprise, premium functions at additional cost
Metadata standards support	• FGDC CSDGM metadata • INSPIRE, based on a directive of the European Union • ISO 19139 Metadata Implementation Specification • ISO 19139 Metadata Implementation Specification GML3.2 • North American Profile of ISO 19115 2003
Recent open-source activity	NA

4.1.1 User Interface: ArcGIS Enterprise Sites

The site used to evaluate ArcGIS Enterprise Sites is the Federal Emergency Management Agency (FEMA) Geospatial Resource Center (https://gis-fema.hub.arcgis.com/), a website hosting spatial data related to emergency management. The associated home page includes logos, visuals, and entry point menus for data, dashboards, applications, and other resources (Figure 4.1). To address the varied nature of disasters, the home page includes entry points that lead to dashboards and curated topic areas, and featured applications that would help the user zero in on their area of interest. Not included in the screenshot are additional home page options to access data by category or region.

The FEMA Geospatial Resource Center query page for wildfire data, Figure 4.2, includes a preview image, a description, a revision date, and source information. Icons provide similar information on source and access, with additional indications of how the dataset can be viewed. The column on the left provides ways to filter the category and types of datasets that are visible.

The dataset detail page of the FEMA Geospatial Resource Center website (Figure 4.3) provides information on dates of creation and update history, plus extensive descriptive information about the contents of the dataset in a column on the left. A column on the right provides a variety of ways to view and interact with the data and provides supplemental information on the dataset. There is no listing of related datasets, but the listing of tags provided in the right-hand column can serve a similar purpose. At the bottom of the right-hand column is a URL link to the dataset. That link takes the user to a landing page that provides links to versions [JavaScript Object Notation (JSON), Simple Object Access Protocol (SOAP), and extensible markup language (XML)] of the data for download, an option to export the image in one of several formats, and several other tools for image analysis, query, or re-projection.

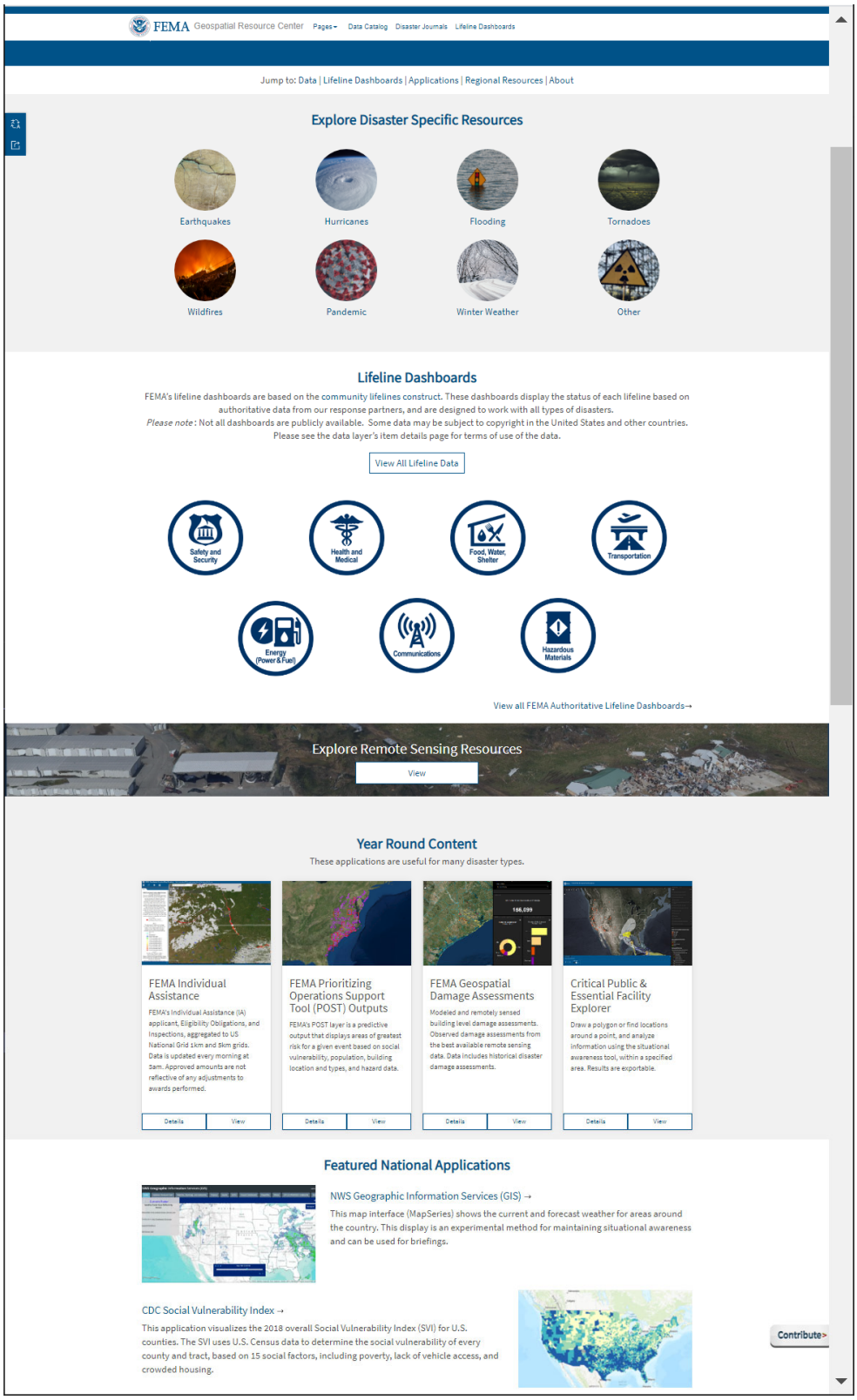

Figure 4.1. Home Page of the FEMA Geospatial Resource Center Web Page (https://gis-fema.hub.arcgis.com/) Implemented Using ArcGIS Hub

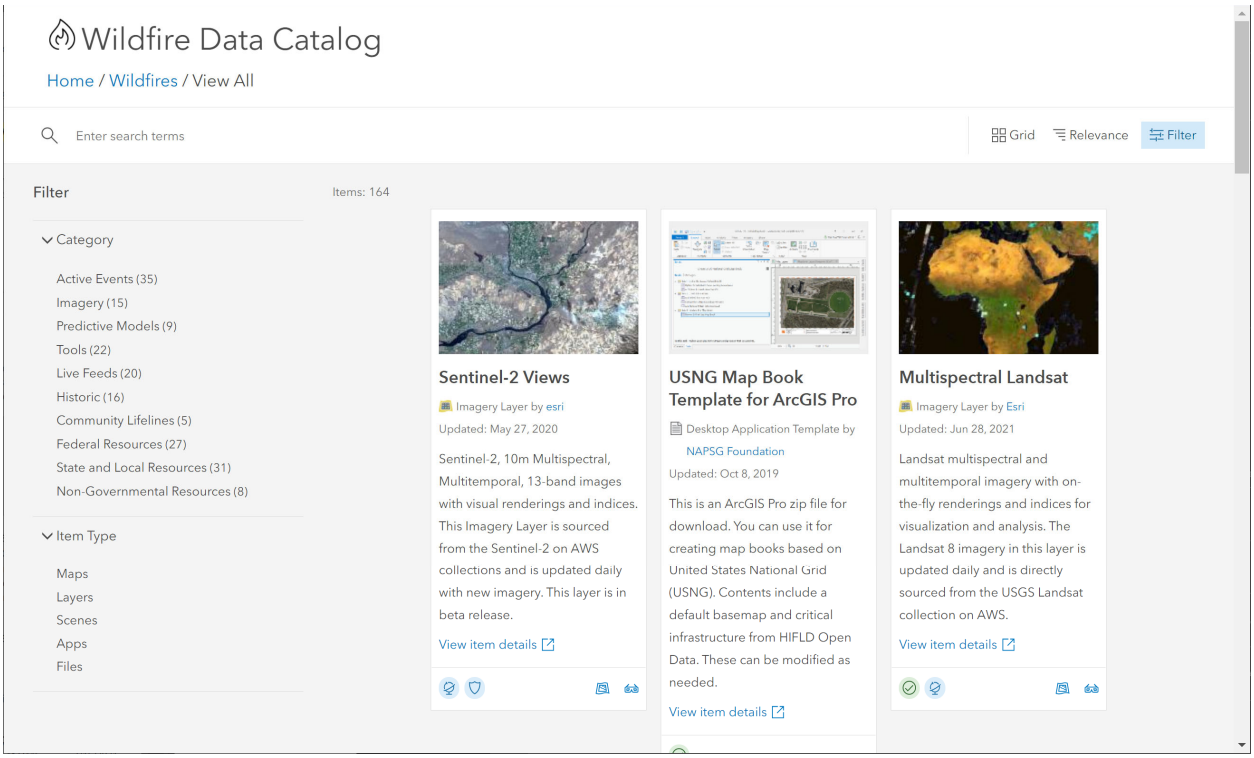

Figure 4.2. Dataset Query Page of the FEMA Geospatial Resource Center (https://fema.maps.arcgis.com/apps/FilterGallery/index.html?appid=4da9f50af45e47f281a1a0e4bdcc11af) Implemented Using ArcGIS Hub

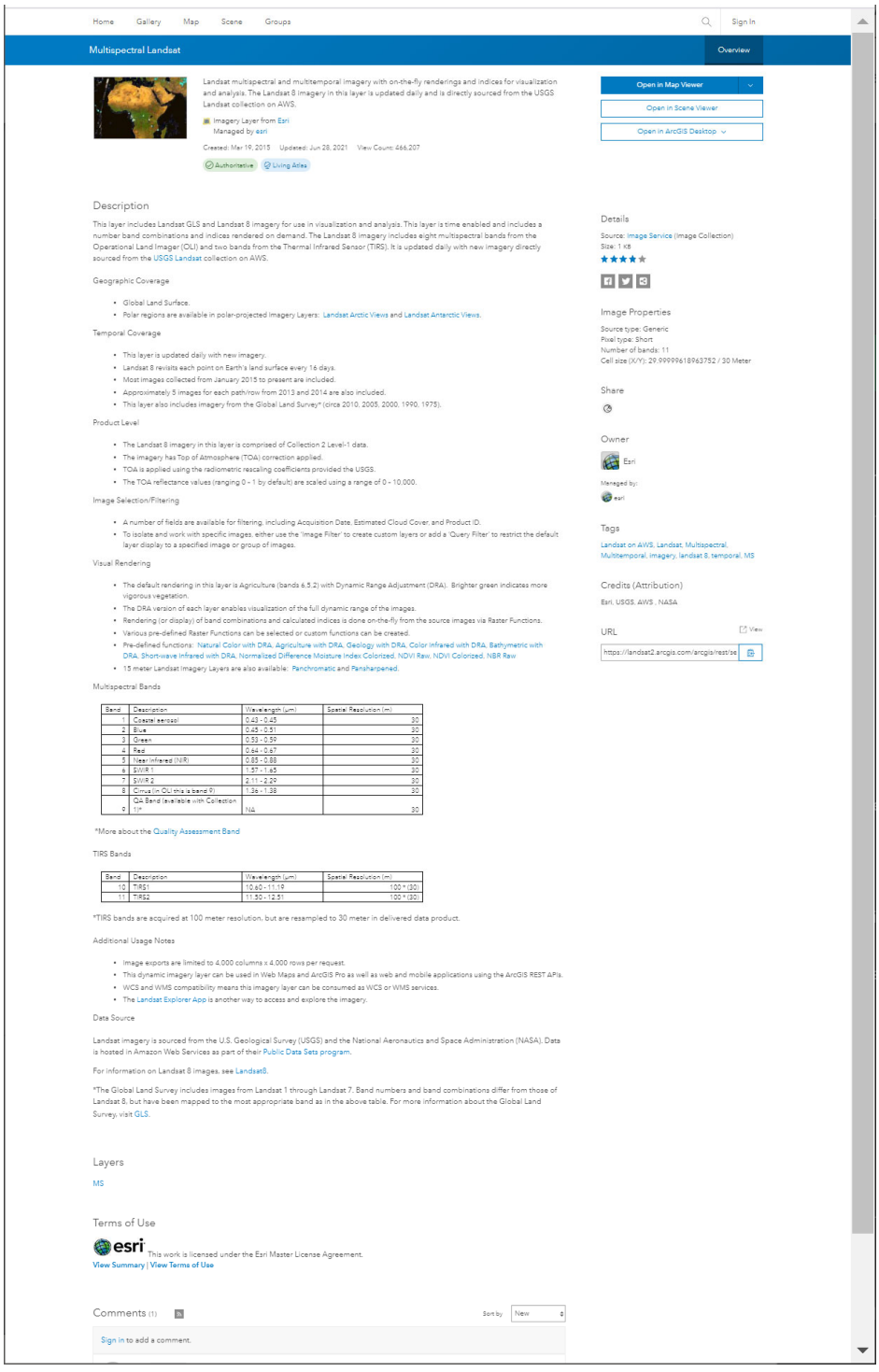

Figure 4.3. Dataset Detail Page of the FEMA Geospatial Resource Center (https://fema.maps.arcgis.com/home/item.html?id=d9b466d6a9e647ce8d1dd5fe12eb434b) Implemented Using ArcGIS Hub

4.1.2 Ratings: ArcGIS Enterprise Sites

ArcGIS Enterprise Sites was rated according to its ability to fulfill the requirements and deliver the desired features described in Sections 3.2 and 3.3. Table 4.2 lists the ratings and justification for each category of requirements/desired features.

Table 4.2. Ratings for ArcGIS Enterprise Sites

Section	Description	Rating	Justification
3.2.1	Catalog	E	Strong metadata standards support and flexibility
3.2.2	Find	E	Themes, stories, keywords, related
3.2.3	Retrieve	E	Table, map, plot, download, API, filter
3.2.4	Large dataset storage	E	Offline, but available by map layer or area
3.3.1	Hanford Site software approval	E	In current use on HLAN
3.3.2	Single sign-on	E	Windows Active Directory (AD) adaptor built in
3.3.3	Role-based access	E	User types, custom roles, and privileges
3.3.4	Linking/Federation	E	Linking and federation among ArcGIS servers
3.3.5	Deployment	E	On premises, with existing infrastructure
	Overall rating	E	

4.2 CKAN (Comprehensive Knowledge Archive Network)

Datopian and Link Digital, the co-stewards of the CKAN project, describe the software as follows: "CKAN is an open-source DMS (data management system) for powering data hubs and data portals. CKAN makes it easy to publish, share, and use data. It powers hundreds of data portals worldwide."[1]

CKAN identifies governments and enterprise as important areas of adoption. As a popular open-source offering, CKAN benefits from an active user community. Though there is no subscription fee for the basic software, the CKAN market is large enough that paid consultants are available to implement, for example, on-premise installations for organizations that prefer to outsource. A broad array of extensions has been developed to bolster the functionality of CKAN. Selected specifications are provided in Table 4.3.

[1] CKAN - The open source data management system, https://ckan.org/. Accessed 9/2/2021

Table 4.3. CKAN Specifications

Specification	Result
Software	CKAN
Website	https://ckan.org/
Hosting models	Self or SaaS
Vendor organization	Datopian and Link Digital are co-stewards of the CKAN project
Latest version	18 May 2021: CKAN 2.9.3 19 October 2012: Initial availability as CKAN 1.7.2
Browser support	Chrome, Firefox, Safari, Edge
Operating system	Linux
Customization programming languages	Python
License	Affero General Public License
Metadata standards support	CKAN, a form of the Data Catalog Vocabulary (DCAT), other standards through extensions
Recent open-source activity	
Contributors (1 month)	16 authors
Commits (1 month)	127 commits
Open tickets (1 month)	39 active issues

4.2.1 User Interface: CKAN

The flagship site for CKAN is DATA.GOV, which catalogs the U.S. government's open data. Though other CKAN sites may include more striking visual imagery, the DATA.GOV interface is uncluttered and straightforward (Figure 4.4). This approach provides a direct path to the abundant datasets cataloged there.

The dataset query page for DATA.GOV (Figure 4.5) includes a simple search bar but allows searching within results or filtering of the results by location, category, type, or tag. In addition to the title and description, the list of datasets includes visual cues to help users find datasets of the desired format and source.

The DATA.GOV dataset detail page (Figure 4.6) focuses on downloading or otherwise obtaining the dataset. The metadata is easily viewable or downloadable and the keywords from the metadata are presented as clickable links to search for other datasets having that keyword. Access and use terms are prominently displayed.

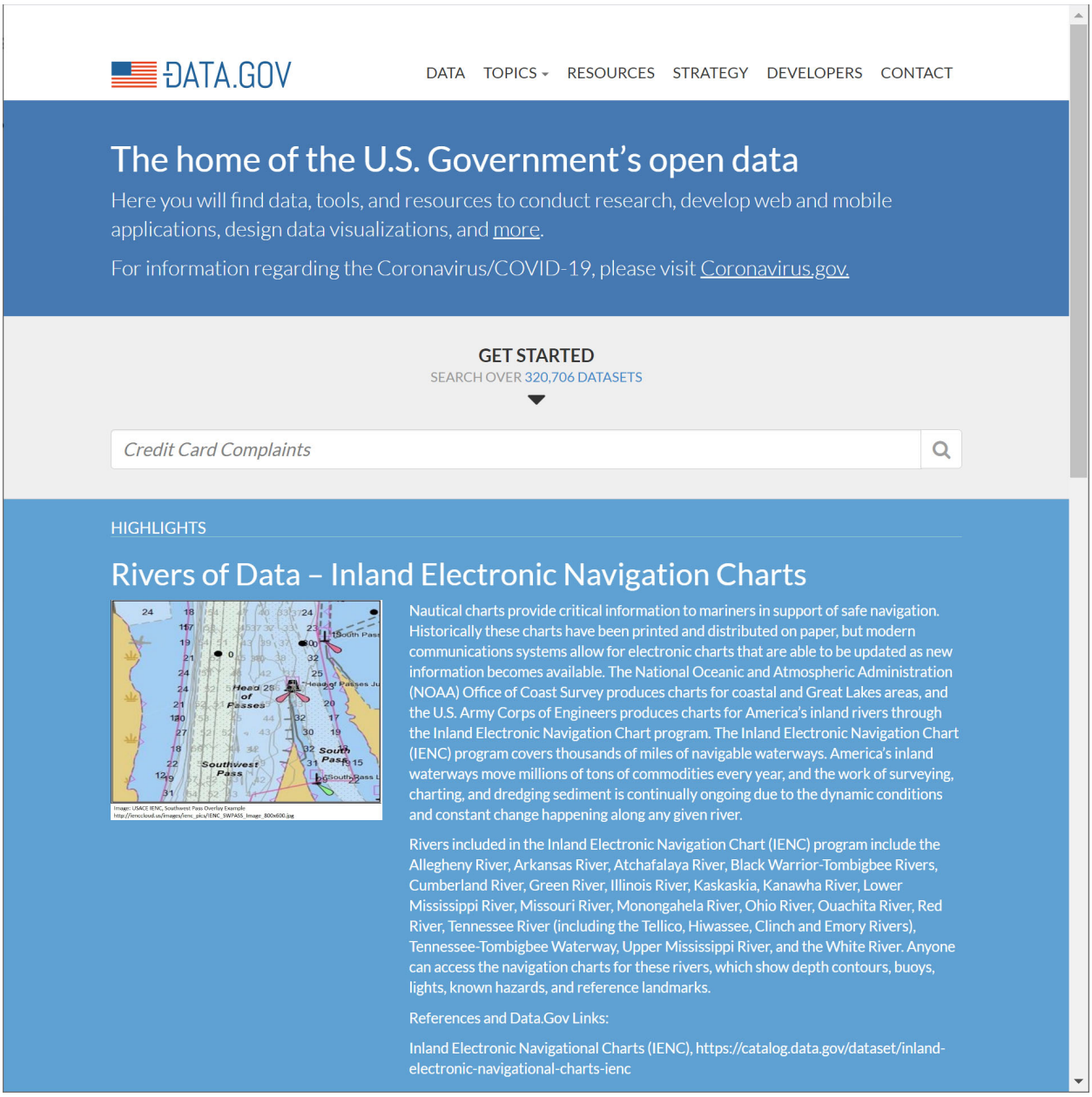

Figure 4.4. Home Page of DATA.GOV (https://www.DATA.GOV/), a Data Catalog Implemented Using CKAN

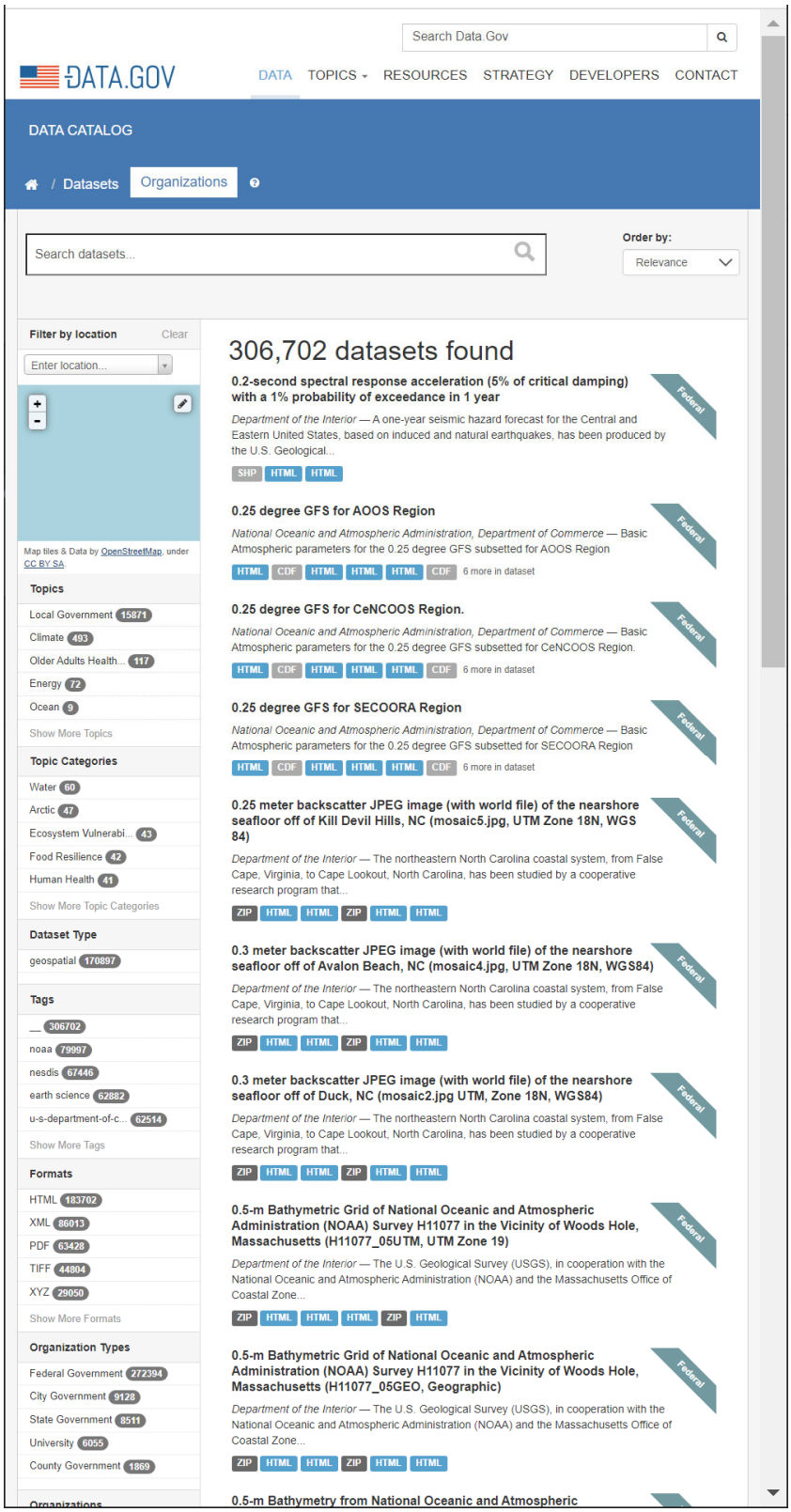

Figure 4.5. Dataset Query Page of DATA.GOV (https://catalog.DATA.GOV/dataset) Implemented Using CKAN

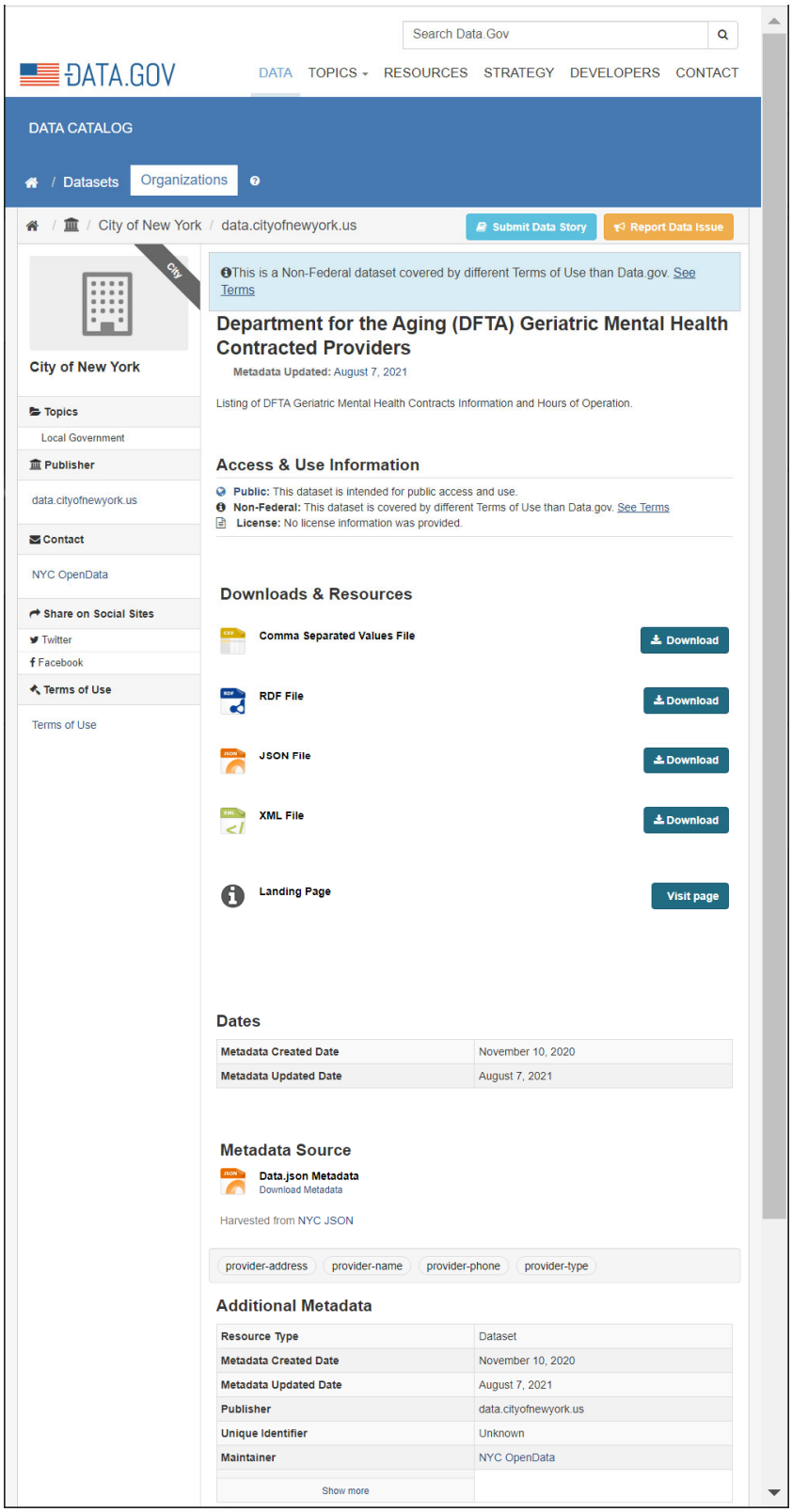

Figure 4.6. Dataset Detail Page of DATA.GOV (https://catalog.DATA.GOV/dataset/department-for-the-aging-dfta-geriatric-mental-health-contracted-providers) Implemented Using CKAN

4.2.2 Ratings: CKAN

Table 4.4 lists the CKAN ratings and justification information for each category of requirements/desired features.

Table 4.4. Ratings for CKAN

Section		Rating	Justification
3.2.1	Catalog	E	Native tools limited, but customization and extensions fill gaps
3.2.2	Find	E	Themes, keywords, related
3.2.3	Retrieve	E	Table, map, plot, download, API
3.2.4	Large dataset storage	E	Multiple options through extensions
3.3.1	Hanford Site software approval	E	In use by U.S. federal government (General Services Administration)
3.3.2	Single sign-on	E	Through extension
3.3.3	Role-based access	E	User types, custom roles, and privileges
3.3.4	Linking/Federation	E	Data harvesting and federation
3.3.5	Deployment	M	On-premises with new infrastructure
	Overall rating	M	

4.3 DKAN

CivicActions, the stewards of the open data project DKAN, describe it as follows: “DKAN is a community-driven, free and open-source open data platform that gives organizations and individuals ultimate freedom to publish and consume structured information.”[1]

DKAN is based on CKAN, but it incorporates it into the Drupal content management system to simplify the creation of pages, blogs, and other content. As was the case for CKAN, there are paid consultants available to help with implementation, if needed. Selected specifications are provided in Table 4.5.

[1] DKAN Open Data Platform, https://getdkan.org/. Accessed 9/2/2021.

Table 4.5. DKAN Specifications

Specification	Result
Software	DKAN
Website	https://getdkan.org/
Hosting models	Self or SaaS
Vendor organization	CivicActions
Latest version	6 May 2021: DKAN 2.10.3 4 April 2014: Initial availability as DKAN 1.0
Browser support	Google Chrome; Firefox; Safari; Microsoft Edge; Opera
Operating system	Linux
Customization programming languages	PHP: Hypertext Preprocessor
License	GNU Public License v2
Metadata standards support	DKAN provides a "data.json" index to satisfy the U.S. federal government's Project Open Data requirements. More information about the "slash data" or "data.json" requirements can be found in Project Open Data's Open Data Catalog Requirements and Common Core Metadata Schema pages. The exact mapping of data (specifically, Drupal data tokens) from your DKAN site to the data.json index can be customized using the Open Data Schema Mapper.
Recent Open-Source Activity	
Contributors (1 month)	8 authors
Commits (1 month)	70 commits
Open Tickets (1 month)	32 active issues

4.3.1 User Interface: DKAN

The U.S. Department of Agriculture's (USDA's) Ag Data Commons provides a good example of a site built using DKAN. The homepage (Figure 4.7) presents data themes and programs, along with some featured datasets and a news feed.

The USDA Ag Data Commons dataset query page provides numerous options for searching, and these options can be chosen in sequence to drill down quickly to a list of datasets (Figure 4.8). Visual cues are included in dataset listings to indicate what formats are available.

The USDA Ag Data Commons dataset detail page (Figure 4.9) example presents a description and a table of metadata. Keywords are presented in three groups, ISO Topic(s), National Agricultural Library Thesaurus Term, and Ag Data Commons Keywords. Clicking on a keyword for either of the first two groups initiates a "drill up" search for all datasets associated with that keyword.

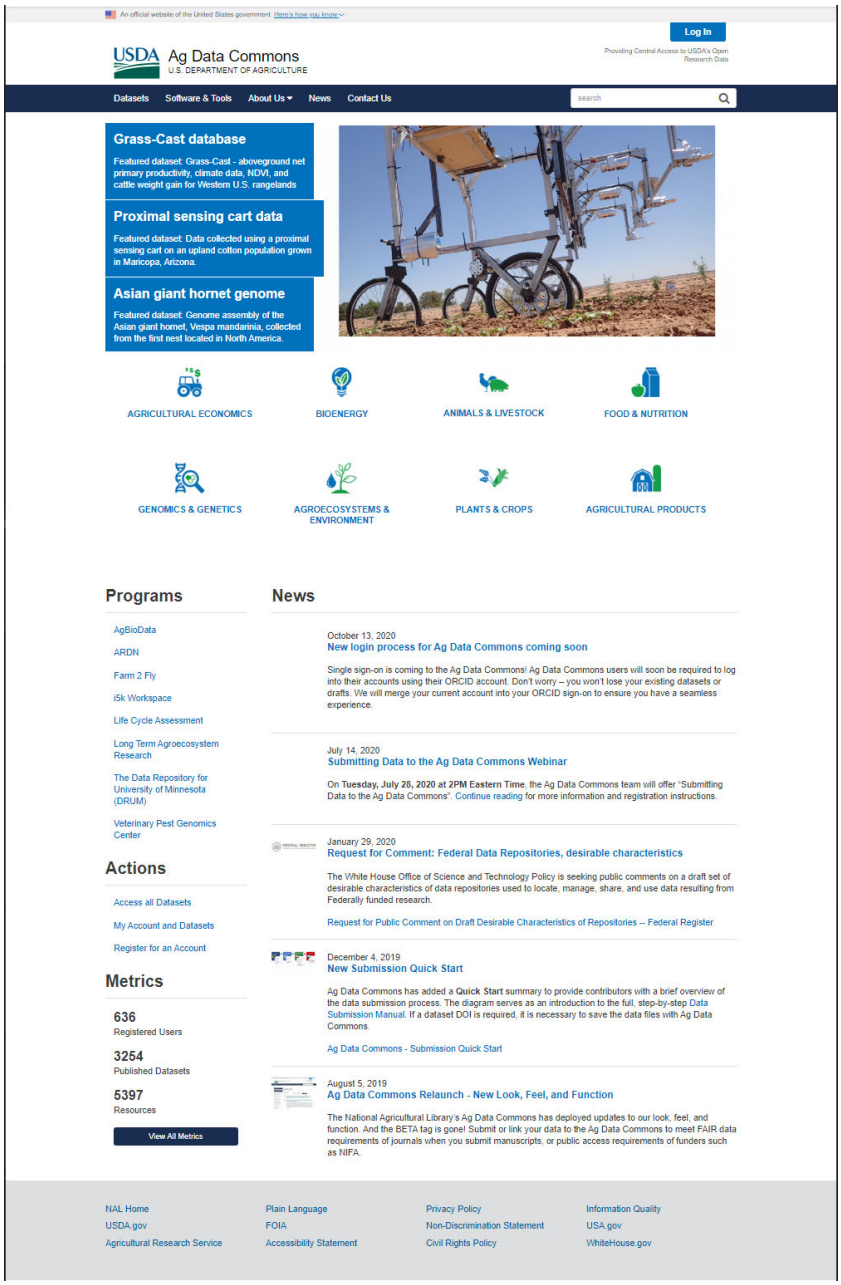

Figure 4.7. Home Page of the USDA Ag Data Commons (https://data.nal.usda.gov/) Implemented Using DKAN

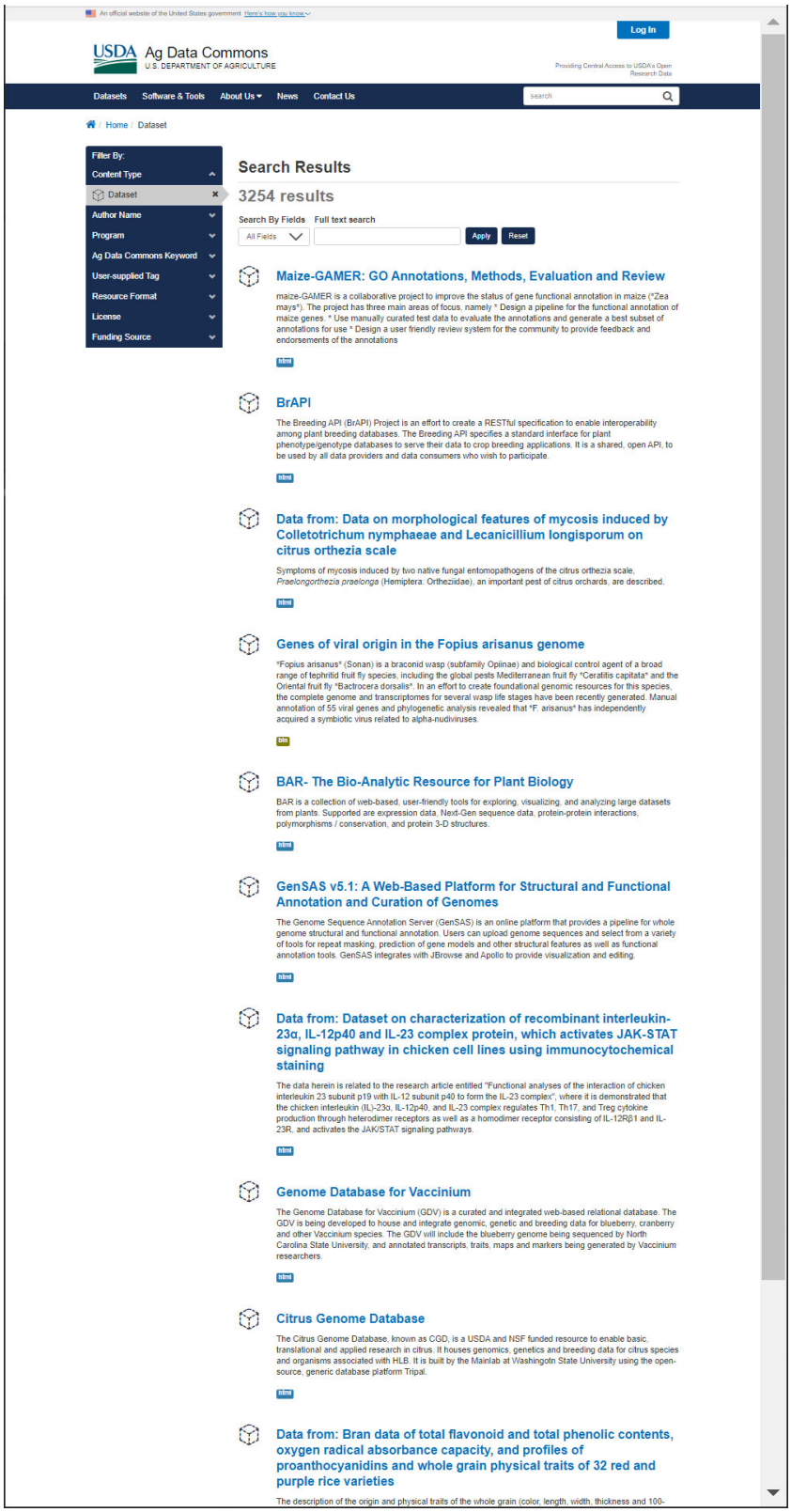

Figure 4.8. Dataset Query Page of the USDA Ag Data Commons (https://data.nal.usda.gov/search/type/dataset) Implemented Using DKAN

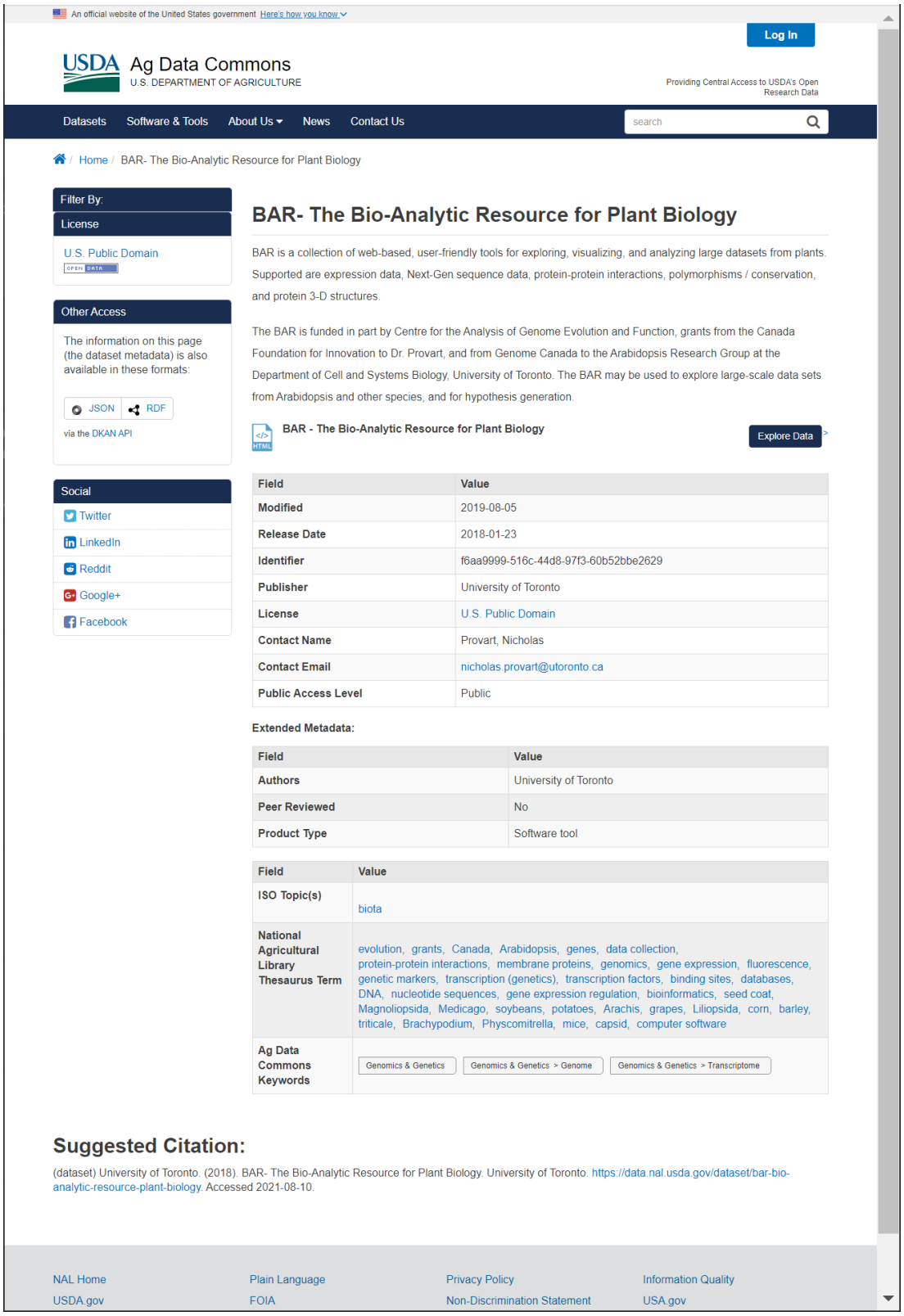

Figure 4.9. Dataset Detail Page of the USDA Ag Data Commons (https://data.nal.usda.gov/dataset/bar-bio-analytic-resource-plant-biology) Implemented Using DKAN

4.3.2 Ratings: DKAN

Table 4.6 lists the DKAN ratings and justification information for each category of requirements/desired features.

Table 4.6. Ratings for DKAN

Section	Description	Rating	Justification
3.2.1	Catalog	E	Custom metadata options
3.2.2	Find	E	Themes, stories, keywords
3.2.3	Retrieve	E	Table, map, plot, download, API
3.2.4	Large dataset storage	E	DataStore was refactored in 2019 to handle large files
3.3.1	Hanford Site software approval	E	In use by U.S. federal government (Health and Human Services, USDA)
3.3.2	Single sign-on	E	Drupal supports AD login
3.3.3	Role-based access	E	Account types, roles, groups, and privileges
3.3.4	Linking/Federation	E	Data harvesting and federation
3.3.5	Deployment	M	On-premises with new infrastructure
	Overall rating	M	

4.4 EDX (Energy Data eXchange)

The Energy Data eXchange (EDX), developed and maintained by the National Energy Technology Laboratory, is "the Department of Energy (DOE)/Fossil Energy's (FE) virtual library and data laboratory built to find, connect, curate, use and re-use data to advance fossil energy and environmental R&D. EDX supports the entire life cycle of data by offering secure, private collaborative workspaces for ongoing research projects until they mature and become catalogued, curated, and published. EDX adheres to DOE Cyber policies as well as domestic and international standards for data curation and citation."[1]

EDX is an existing open data portal custom built using the CKAN software. Certain challenges arise when using a catalog run by others for purposes other than your own. In the case of EDX, for example, the metadata requirements are straightforward and simple, which is appropriate to a focus on sharing and reuse of data sets. Incorporating datasets into HLAN workflows, in contrast, would require a more structured approach to metadata and greater control on the contents of metadata fields to avoid ambiguity when combining data from multiple HLAN systems. It is not reasonable to expect a public resource such as EDX to alter its configuration to meet HLAN objectives, but it is worth considering whether the benefits of using an existing catalog outweigh those challenges. Selected specifications are provided in Table 4.7.

[1] About - EDX - Energy Data eXchange, https://edx.netl.doe.gov/about. Accessed 9/2/2021.

Table 4.7. EDX Specifications

Specification	Result
Software	EDX
Website	https://edx.netl.doe.gov/
Hosting models	SaaS
Vendor organization	National Energy Technology Laboratory
Latest version	NA
Browser support	Internet Explorer; Firefox; Chrome
Operating system	Customized website based on CKAN
Customization programming languages	NA
License	NA
Metadata standards support	Unknown
Recent open-source activity	NA

4.4.1 User Interface: EDX

Unlike websites used to illustrate how the other tools might implement a data catalog user interface, the EDX website is a data catalog operated for other purposes that is being considered as a candidate for delivering the HEIDI functions. Rather than evaluating whether a tool can be used to implement a data catalog that meets the requirements, we are evaluating whether incorporating HEIDI data into the existing EDX catalog can meet requirements. The homepage presents information about EDX, some login information, and a menu for accessing the various function areas of the site (Figure 4.10).

The EDX dataset query page provides the list of all dataset with options for filtering by source, type, group, keyword, format, or location. (Figure 4.11). Visual cues are included in dataset listings to indicate what formats are available, the source, the size, and how many files are included.

The EDX dataset detail page (Figure 4.12) presents the resources available for a dataset, with an option to preview or go to that resource. The citation for the dataset is made available to copy. Keywords from the metadata are presented as links to search for other datasets having that keyword. The spatial extent of the data, when available, is mapped. The ability to follow a dataset, nominate a dataset to a group, join a related group, and see activity are ways for users to maintain a connection with a topic area. A link to the metadata opens the file in its native format, which is not necessarily human-readable.

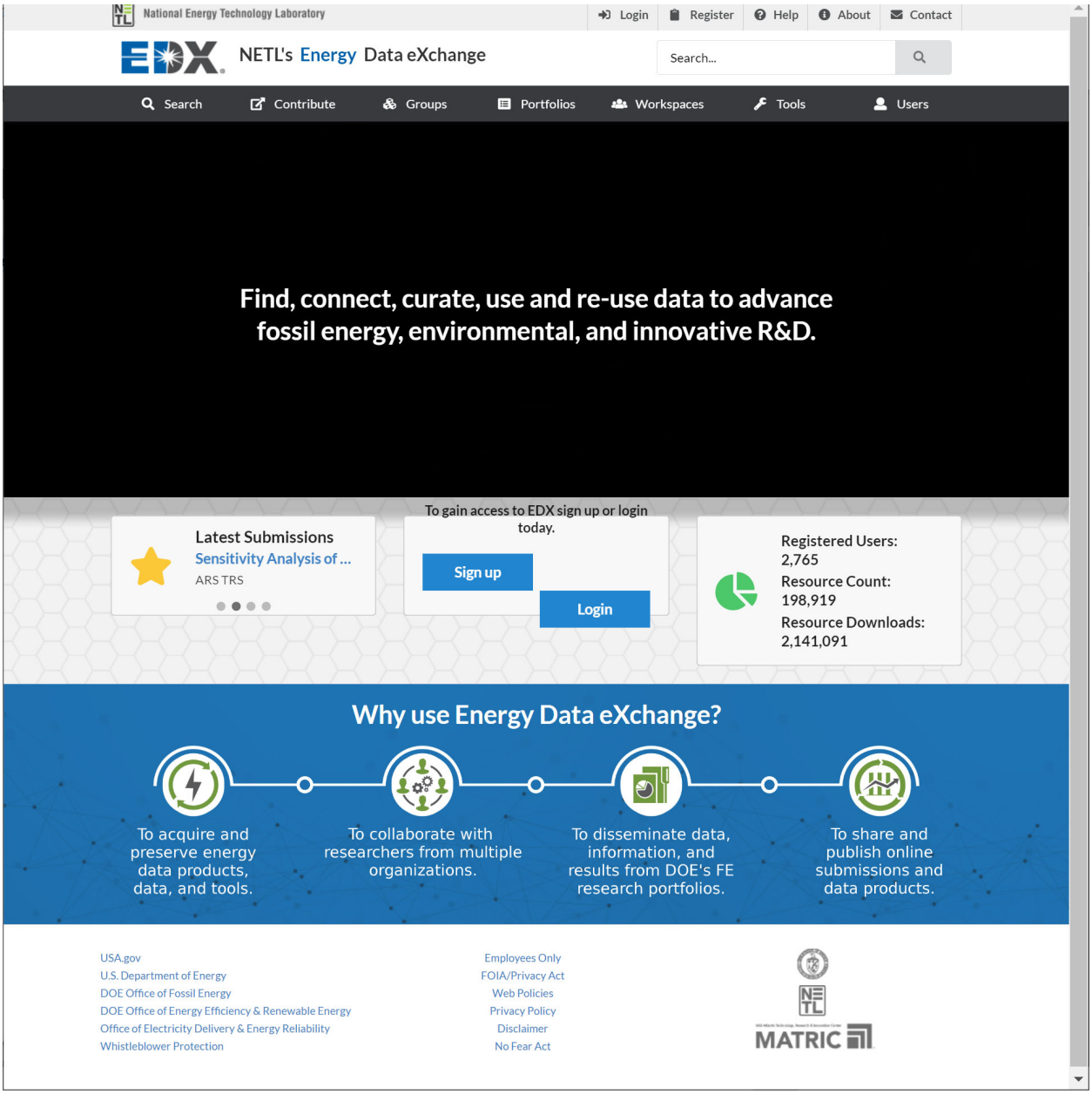

Figure 4.10. Home Page of the Energy Data eXchange (https://edx.netl.doe.gov/)

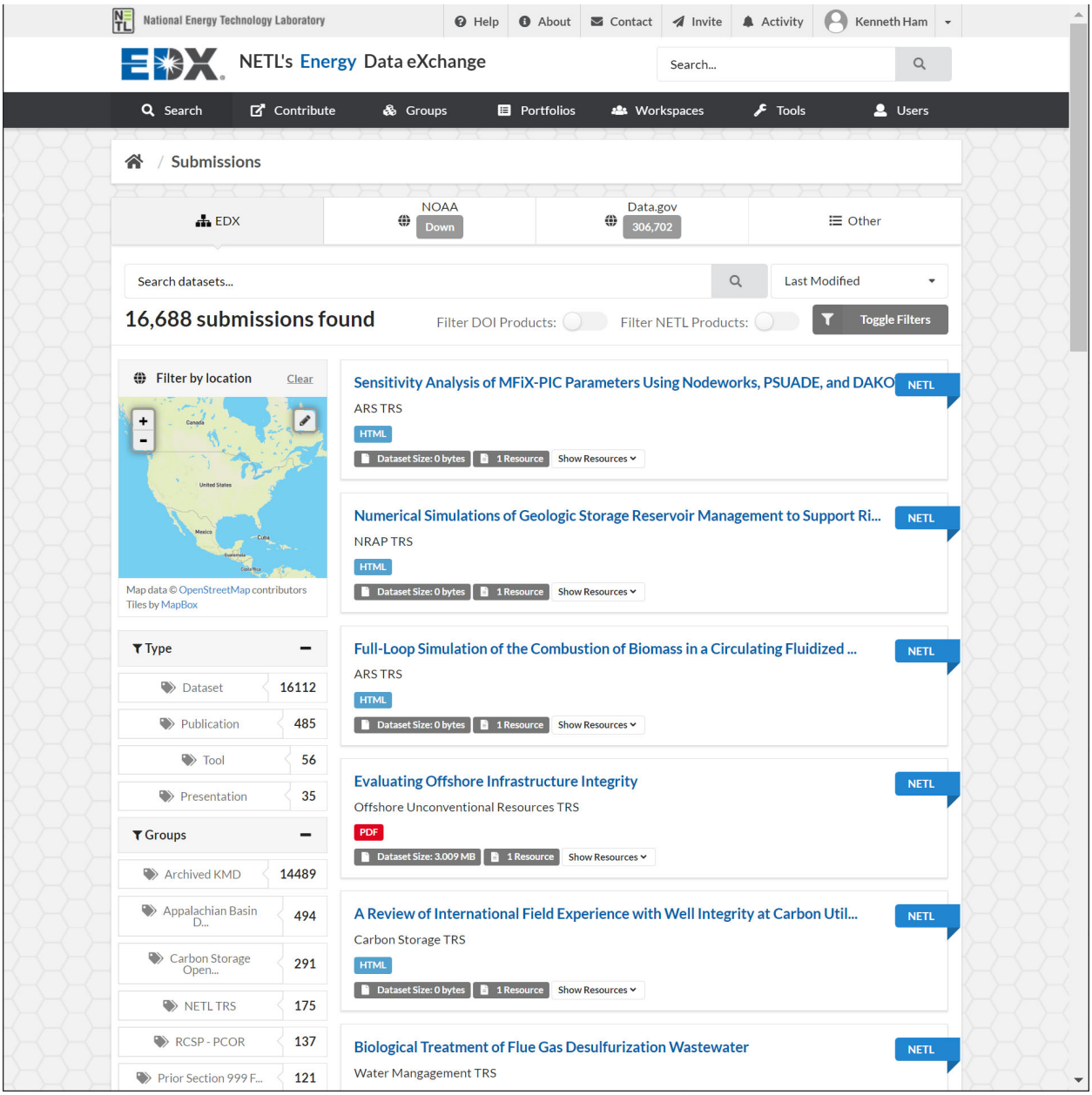

Figure 4.11. Dataset Query Page of the Energy Data eXchange (https://edx.netl.doe.gov/)

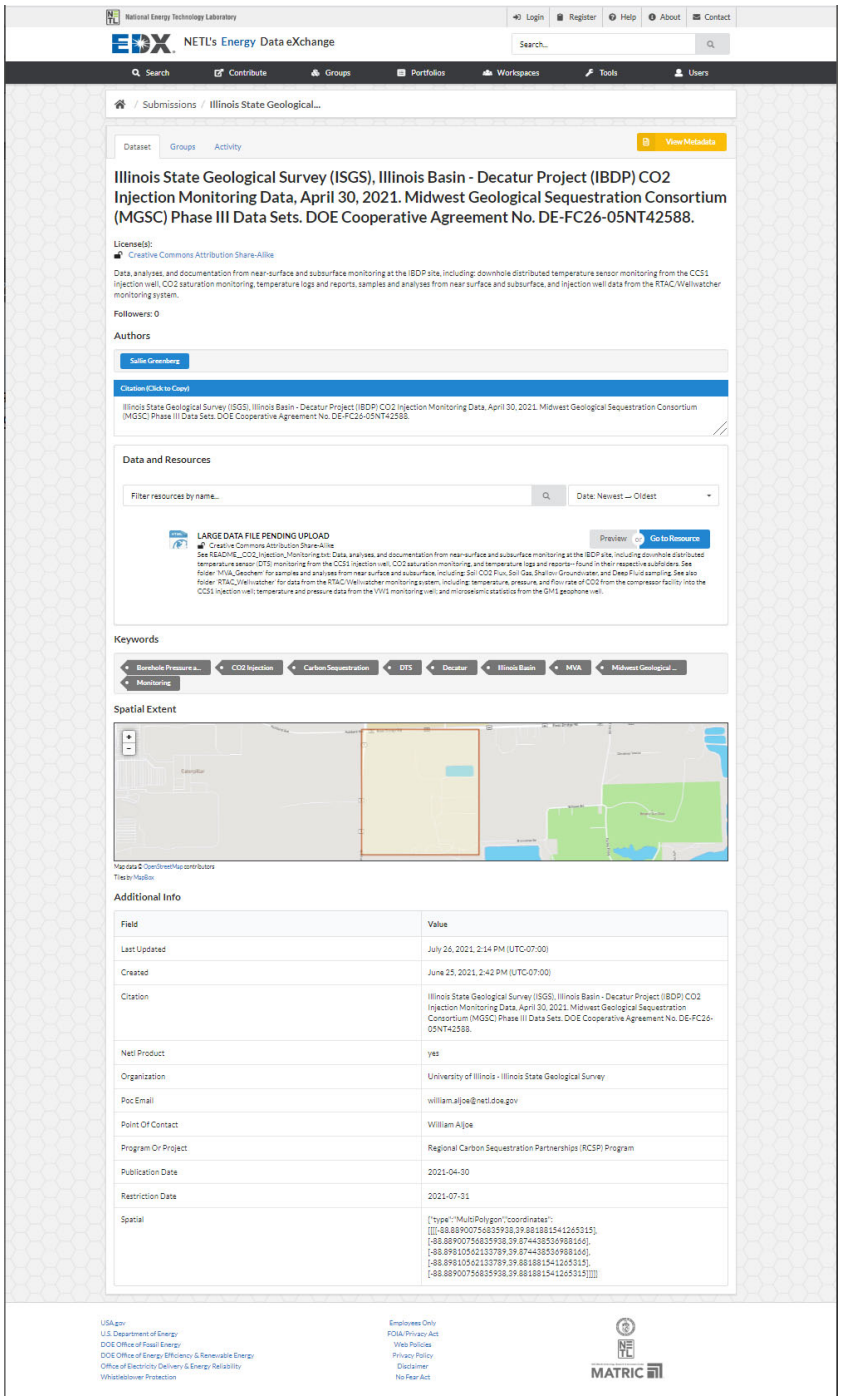

Figure 4.12. Dataset Detail Page of the Energy Data eXchange (https://edx.netl.doe.gov/dataset/illinois-state-geological-survey-isgs-illinois-basin-decatur-project-ibdp-co2-injection-monitoring)

4.4.2 Ratings: EDX

Table 4.8 lists the EDX ratings and justification information for each category of requirements/desired features.

Table 4.8. Ratings for EDX

Section	Description	Rating	Justification
3.2.1	Catalog	M	Limited customization, unclear how metadata standards and vocabularies could be enforced
3.2.2	Find	M	Difficult to search on Hanford Site location names
3.2.3	Retrieve	E	Flexible access to dataset resources
3.2.4	Large dataset storage	M	Supported, but some large datasets are pending upload
3.3.1	Hanford Site software approval	E	Not needed
3.3.2	Single sign-on	F	New sign-on required
3.3.3	Role-based access	M	Workspaces allow for private manipulation of prepublication datasets, but intent is for them to go public
3.3.4	Linking/Federation	E	Link is implemented to other major data repositories
3.3.5	Deployment	M	The data catalog is operational and maintained by others, but it would be difficult to incorporate HLAN requirements without disrupting current uses and users
	Overall rating	F	

4.5 Junar

Junar, Inc., the makers of Junar, introduce their software with “Junar data platform lets you transform your hard-to-find and useless data assets into dynamic tables, visualizations, maps, dashboards and APIs – so citizens, developers, and companies can re-use them for their interests in a simple way.”[1]

Junar is a commercial, subscription-based offering that hosts your data catalog on their cloud servers. Selected specifications are provided in Table 4.9.

[1] Junar Data Platform, https://www.junar.com/. Accessed 9/2/2021.

Table 4.9. Junar Specifications

Specification	Result
Software	Junar
Website	https://www.junar.com/
Hosting models	SaaS
Vendor organization	Junar, Inc.
Latest version	2021: Junar Version 2 2010: Company founded
Browser support	Unknown
Operating system	NA
Customization programming languages	NA
License	Commercial
Metadata	DCAT, INSPIRE
Recent open-source activity	NA

4.5.1 User Interface: Junar

The City of Palo Alto, California, employs Junar for their Open Data Portal. The landing page (Figure 4.13) displays a simple search bar, a prominent array of links to data themes, and a spotlight of a dataset on city businesses.

The City of Palo Alto Open Data Portal dataset query page (Figure 4.14) includes a search bar, a list of datasets, and the option to filter the list by category. Dataset entries include an icon that indicates the type of dataset, the title, and a brief description.

The City of Palo Alto Open Data Portal dataset detail page (Figure 4.15) includes a search bar; visualization of the data (when available); and a menu bar on the left with options to display information, refresh the view, sort, pivot, or filter rows, embed the visual, access the dataset through an API, or export/download the dataset. The information display includes basic metadata fields with clickable tags that link to lists of other datasets associated with the keyword.

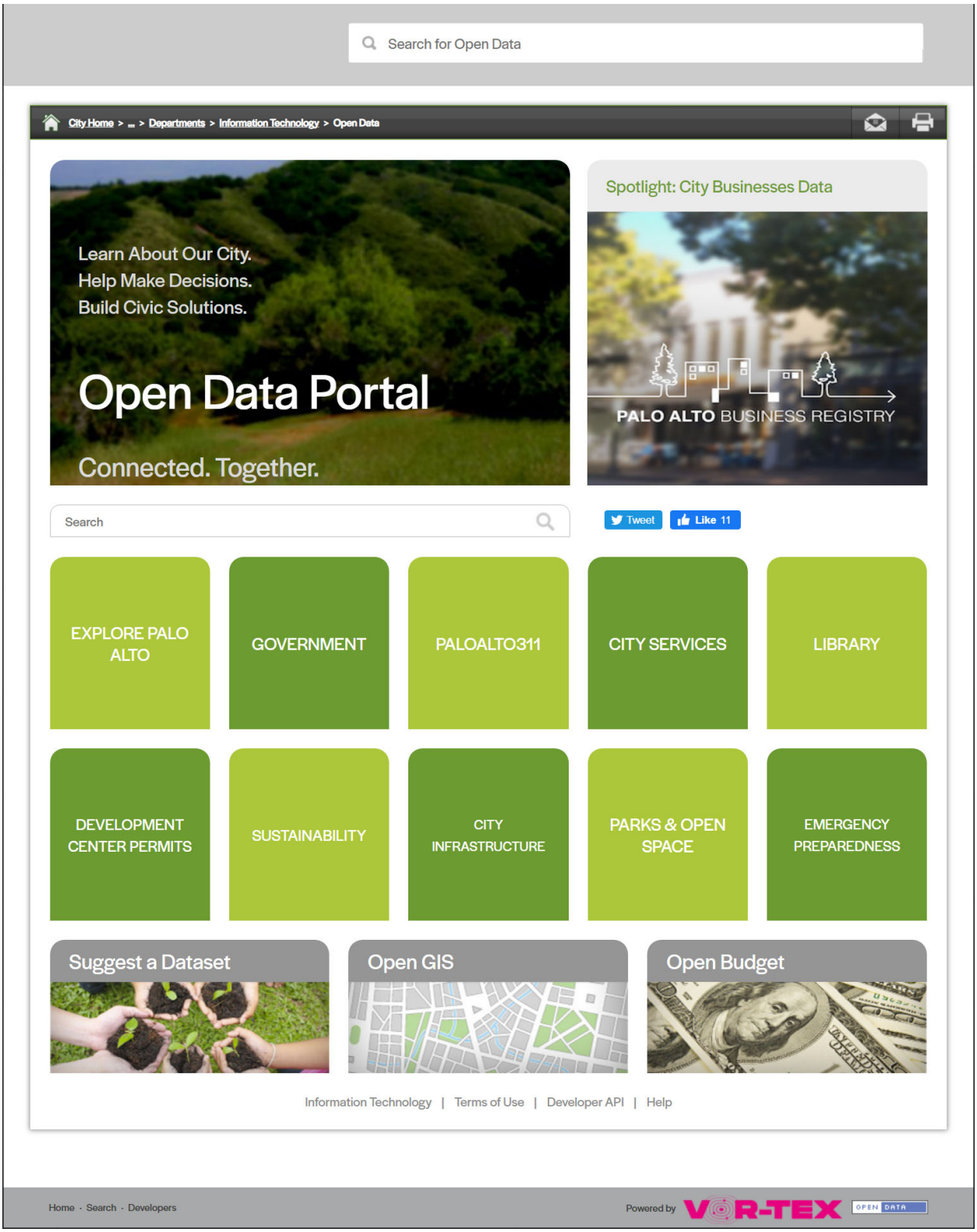

Figure 4.13. Home Page of the City of Palo Alto Open Data Portal (https://data.cityofpaloalto.org/home) Implemented Using Junar

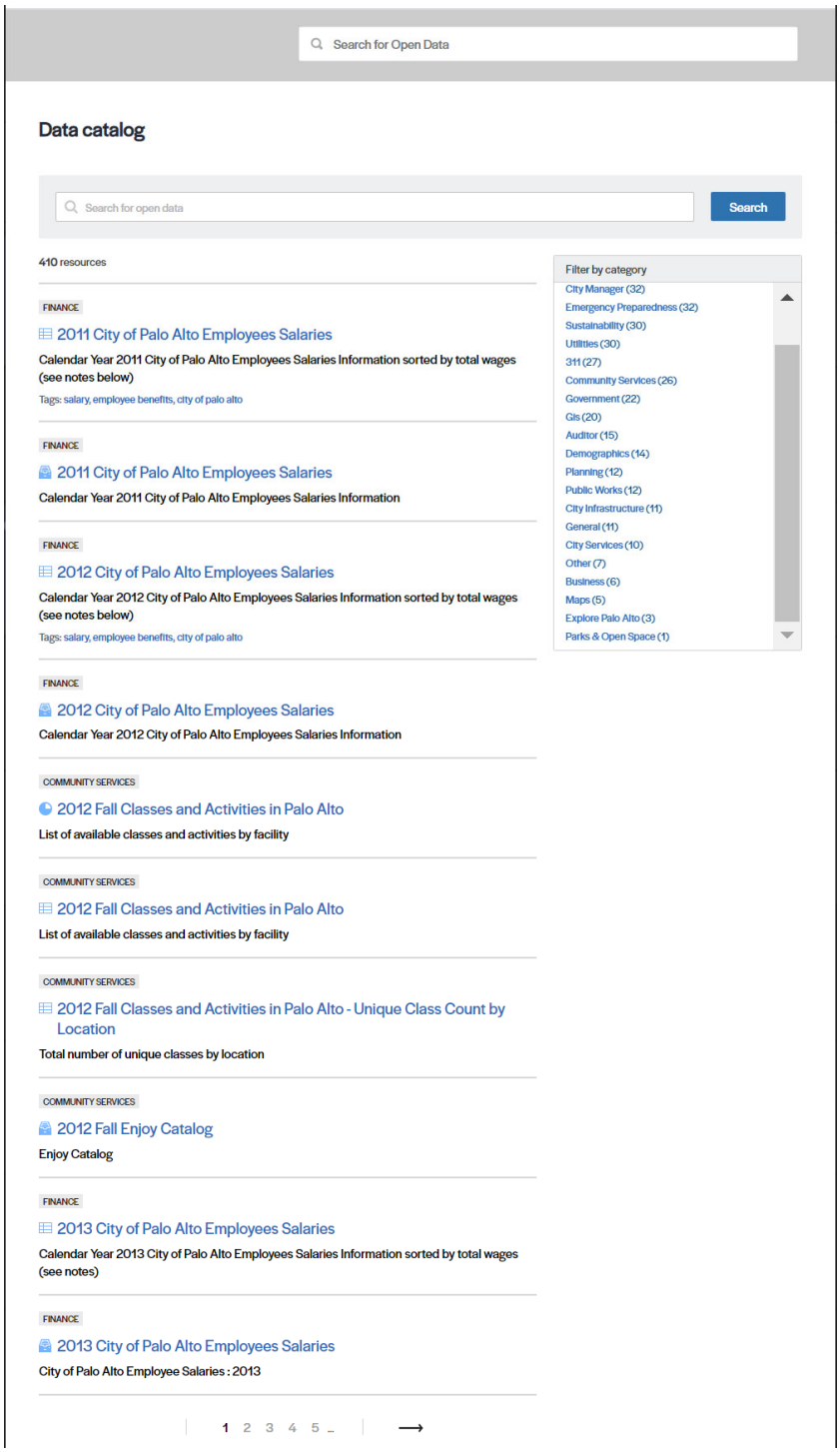

Figure 4.14. Dataset Query Page of the City of Palo Alto Open Data Portal (https://data.cityofpaloalto.org/dashboards/8863/parks-open-space/) Implemented Using Junar

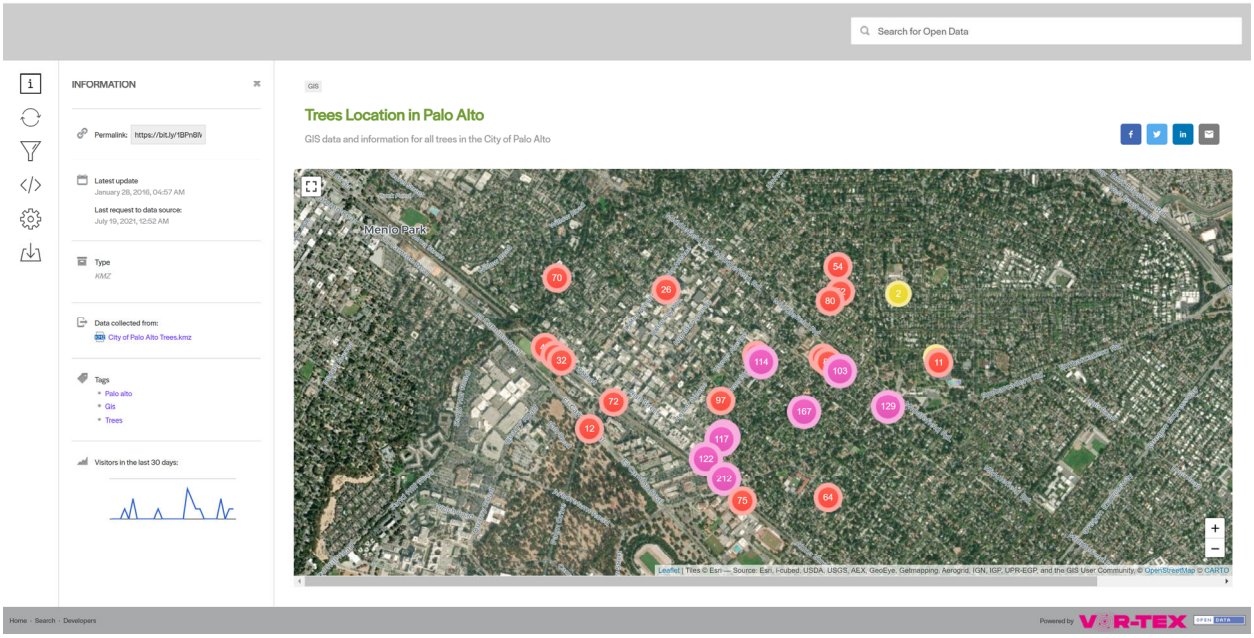

Figure 4.15. Dataset Detail Page of the City of Palo Alto Open Data Portal (https://data.cityofpaloalto.org/visualizations/8227/trees-location-in-palo-alto/) Implemented Using Junar

4.5.2 Ratings: Junar

Table 4.10 lists the Junar ratings and justification information for each category of requirements/desired features.

Table 4.10. Ratings for Junar

Section		Rating	Justification
3.2.1	Catalog	F	Metadata standard support is unclear
3.2.2	Find	E	Themes, stories, keywords
3.2.3	Retrieve	M	Table, download, API
3.2.4	Large dataset storage	F	Unknown
3.3.1	Hanford Site software approval	M	Cloud Security Alliance (CSA) self-assessment available
3.3.2	Single sign-on	F	Unknown
3.3.3	Role based access	F	Users and roles
3.3.4	Linking/Federation	E	Add-on federation feature
3.3.5	Deployment	F	SaaS
	Overall rating	F	

4.6 OpenDataSoft

The Société OPENDATASOFT, operators of the OpenDataSoft platform, introduce the software with "Opendatasoft is the data sharing platform teams use to access, reuse, and share data that grows business."[1]

The OpenDataSoft premium paid subscription allows access to all services of the OpenDataSoft Platform. A "freemium" offering provides access with fewer benefits and no guarantees of performance.

Selected specifications are provided in Table 4.11.

Table 4.11. OpenDataSoft Specifications

Specification	Result
Software	OpenDataSoft
Website	https://www.opendatasoft.com/
Hosting models	SaaS
Vendor organization	Opendatasoft, Inc.
Latest version	6 July 2021: Version 2021 week 27 2011: Company founded
Browser support	Chrome; Safari; Edge; Firefox; Internet Explorer (not recommended)
Operating system	NA
Customization programming languages	NA
License	Commercial
Metadata	DCAT, INSPIRE
Recent open-source activity	NA

4.6.1 User Interface: OpenDataSoft

The City of Vancouver, Canada, implements their open data portal on the OpenDataSoft platform, providing a good example of functionality. The Vancouver Open Data Portal home page (Figure 4.16) offers data by theme, by popularity, and by recent modification, along with some links to tools and tutorials.

The Vancouver Open Data Portal dataset query page (Figure 4.17) includes several options for sorting and filtering the list of datasets. Multiple filters can be selected to drill down to a manageable list of datasets. Dataset entries include icons to identify themes, data visualizations, and access options.

The Vancouver Open Data Portal dataset detail page (Figure 4.18) provides simple metadata fields and a schema for the dataset. A menu bar on the left enables the dataset to be filtered according to type or area. Tabs allow the user to switch among views, including information, a data table, map, plots, export, or API access.

[1] Opendatasoft - Make Your Data Bright, https://www.opendatasoft.com/. Accessed 9/2/2021.

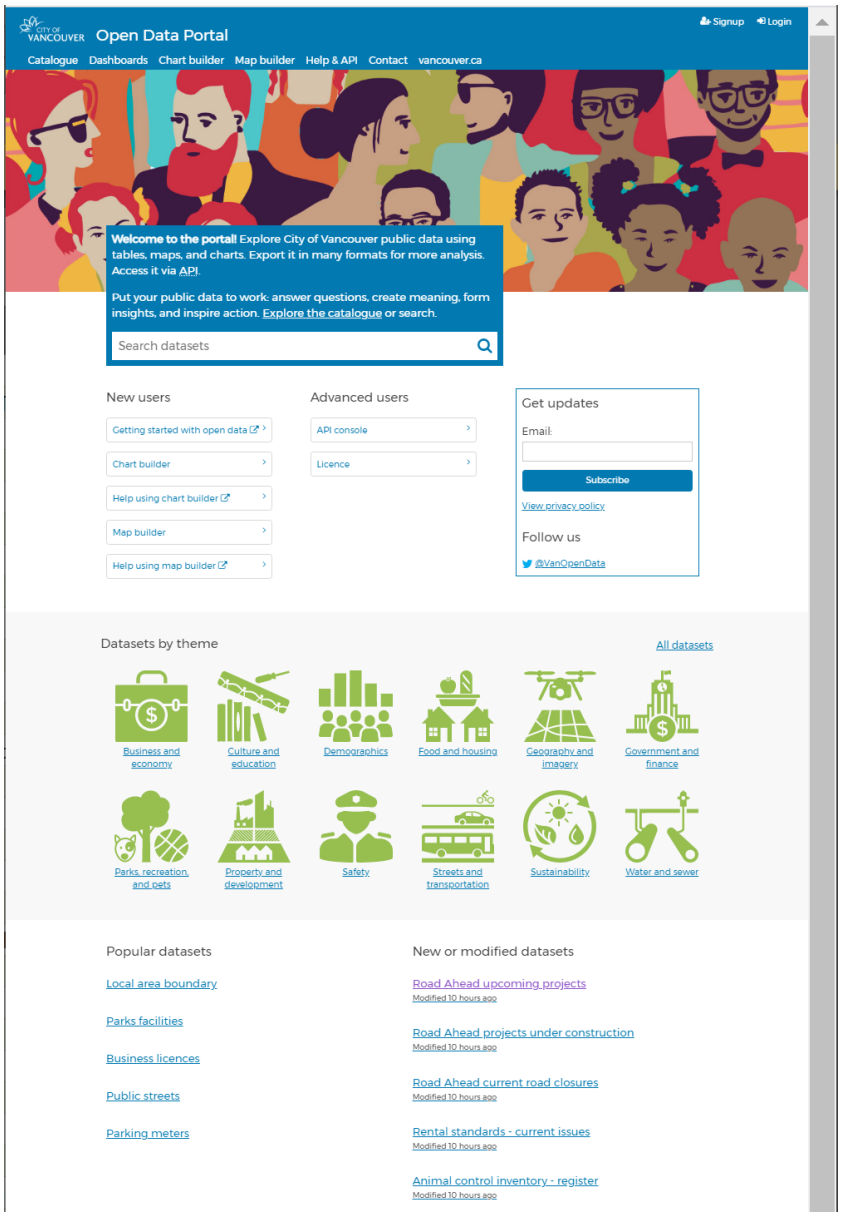

Figure 4.16. Home Page of the City of Vancouver Open Data Portal (https://opendata.vancouver.ca/pages/home/) Implemented Using OpenDataSoft

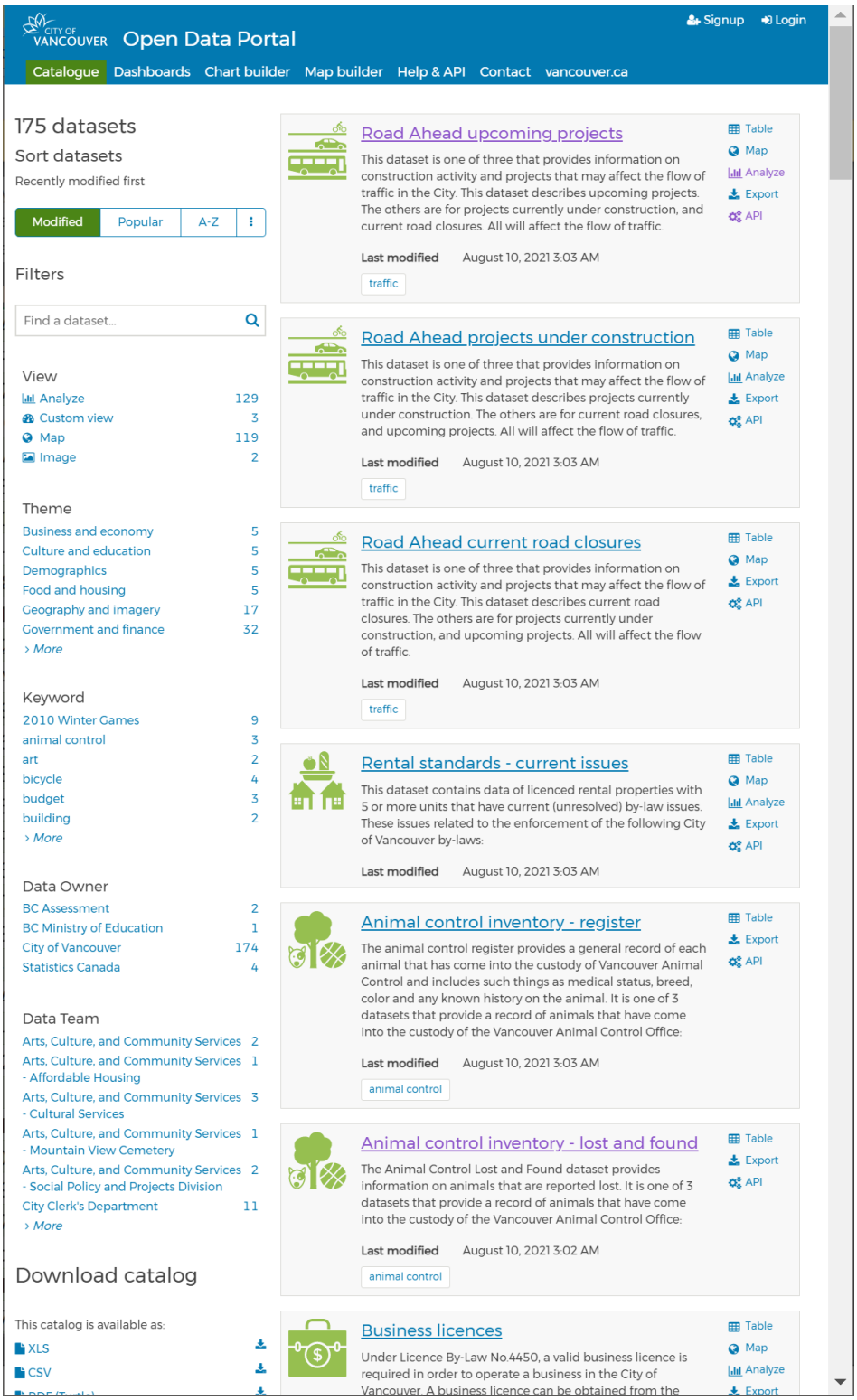

Figure 4.17. Dataset Query Page of the City of Vancouver Open Data Portal (https://opendata.vancouver.ca/explore/?disjunctive.features&disjunctive.theme&disjunctive.keyword&disjunctive.data-owner&disjunctive.data-team&sort=modified) Implemented Using OpenDataSoft

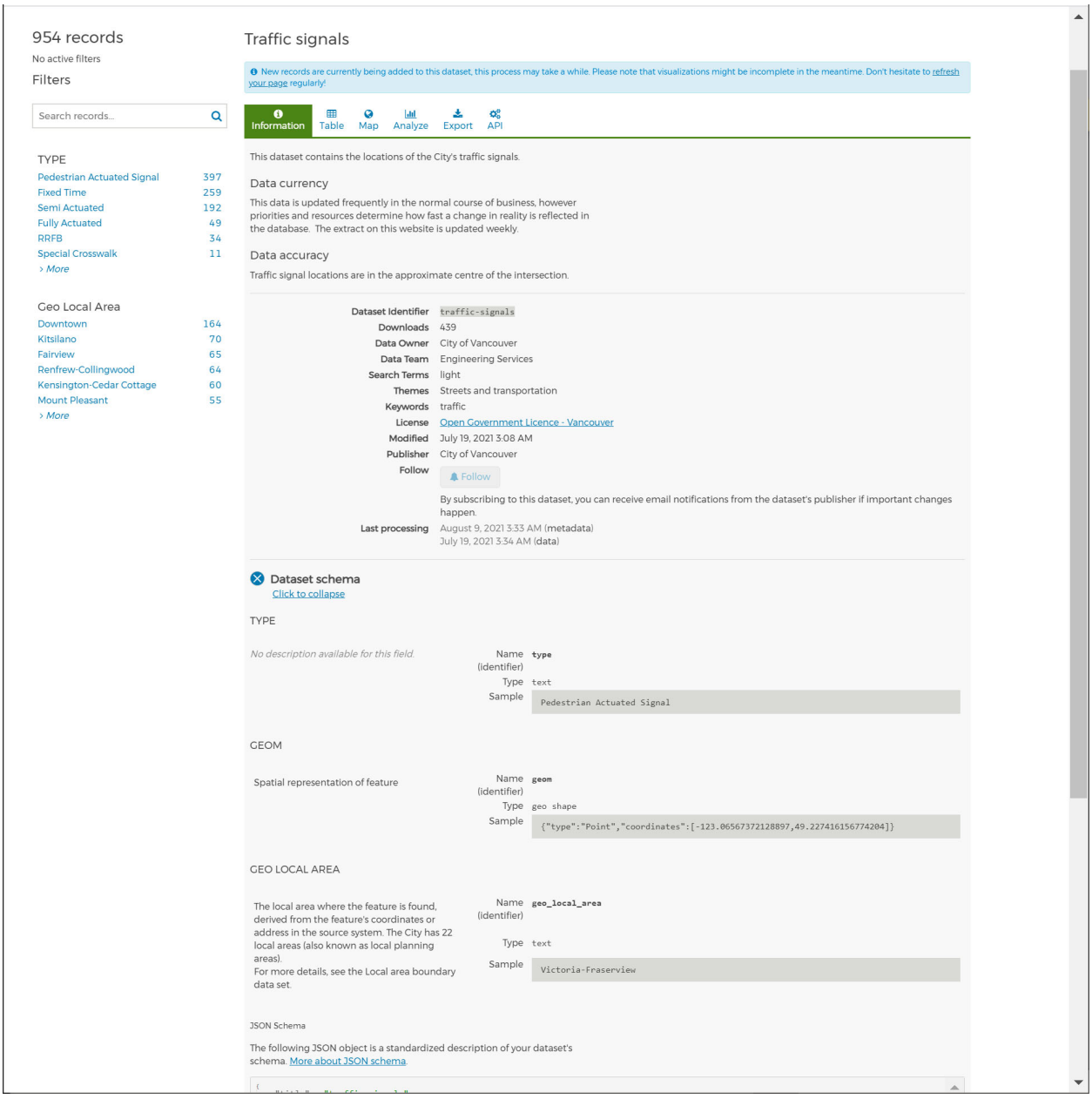

Figure 4.18. Dataset Detail Page of the City of Vancouver Open Data Portal (https://opendata.vancouver.ca/explore/dataset/traffic-signals/information/) Implemented Using OpenDataSoft

4.6.2 Ratings: OpenDataSoft

Table 4.12 lists the OpenDataSoft ratings and justification information for each category of requirements/desired features.

Table 4.12. Ratings for OpenDataSoft

Section		Rating	Justification
3.2.1	Catalog	E	Customizable application profiles implement standards
3.2.2	Find	M	Themes, keywords
3.2.3	Retrieve	E	Table, map, plot, download, API, filter
3.2.4	Large dataset storage	F	License fee increases with data storage. File size limits apply to cloud storage.
3.3.1	Hanford Site software approval	M	CSA self-assessment is available
3.3.2	Single sign-on	E	Through Security Assertion Markup Language (SAML)
3.3.3	Role-based access	M	Users, groups, and permissions
3.3.4	Linking/Federation	E	Harvesting and federation
3.3.5	Deployment	F	SaaS
	Overall rating	F	

4.7 Socrata

Tyler Technologies introduces its Socrata data platform as, “The Socrata data platform enables governments to use data as a strategic asset in the design, management, and delivery of programs. Data flows easily between staff and departments leading to more efficient programs and better decision making.”[1]

The subscription-based Socrata data platform is delivered as SaaS, with an array of optional products providing various functions. Selected specifications are provided in Table 4.13. Socrata is authorized by the Federal Risk and Authorization Management Program (FedRAMP) for moderate impact level, which means that it can handle government data where the loss of confidentiality, integrity, and availability would result in serious adverse effects on operations, assets, or individuals, but where adverse effects do not include loss of life or physical harm.

[1] Socrata - Data & Insights Solutions - Tyler Technologies, https://www.tylertech.com/products/socrata. Accessed 9/2/2021.

Table 4.13. Socrata Specifications

Specification	Result
Software	Socrata
Website	https://www.tylertech.com/products/socrata
Hosting models	SaaS
Vendor organization	Tyler Technologies or open-source option
Latest version	6 July 2021: Version 2021 week 27 2011: Company founded
Browser support	Chrome; Safari; Edge; Firefox; Internet Explorer
Operating system	NA
Customization programming languages	NA
License	Commercial or Apache Open Source
Metadata	Customizable schema
Recent open-source activity	
Contributors (1 month)	1 author
Commits (1 month)	0 commits
Open tickets (1 month)	0 active Issues

4.7.1 User Interface: Socrata

An example of the use of Socrata is the City of Austin, Texas, Open Data Portal. The home page (Figure 4.19) provides a search bar as well as an array of clickable links to data by theme.

The City of Austin, Texas, Open Data Portal dataset query page (Figure 4.20) includes a column of options to filter by category, type, department, keyword, or domain. The dataset entries include icons to indicate how the data will be presented and a list of clickable tags that provide a new list based on the tag.

The City of Austin, Texas, Open Data Portal dataset detail page (Figure 4.21) includes basic metadata fields, clickable tags, a text description of the data contents, a preview of the data, and menu options to visualize, export, or access the API.

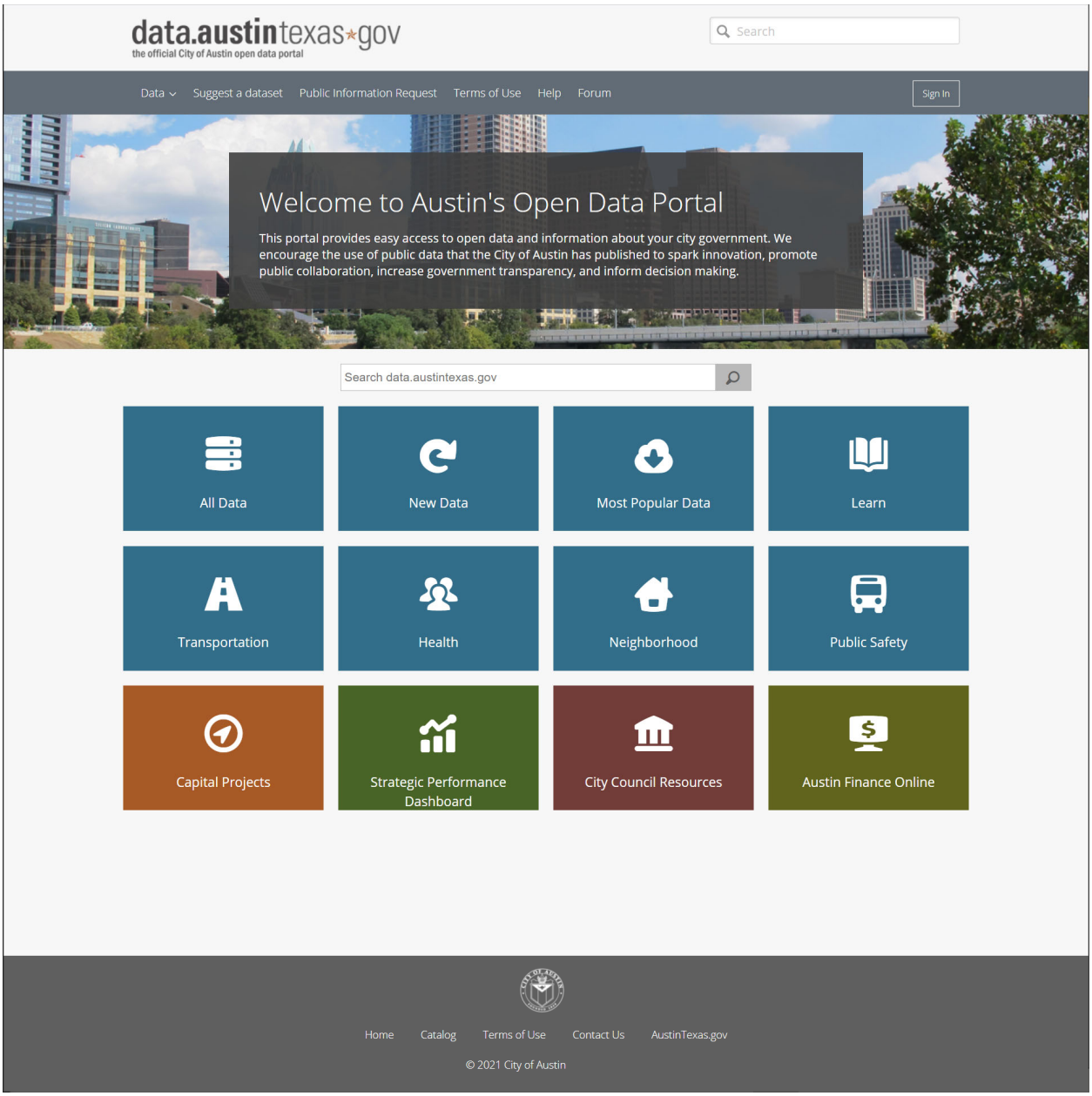

Figure 4.19. Home Page of the City of Austin, Texas, Open Data Portal (https://data.austintexas.gov/) Implemented Using Socrata

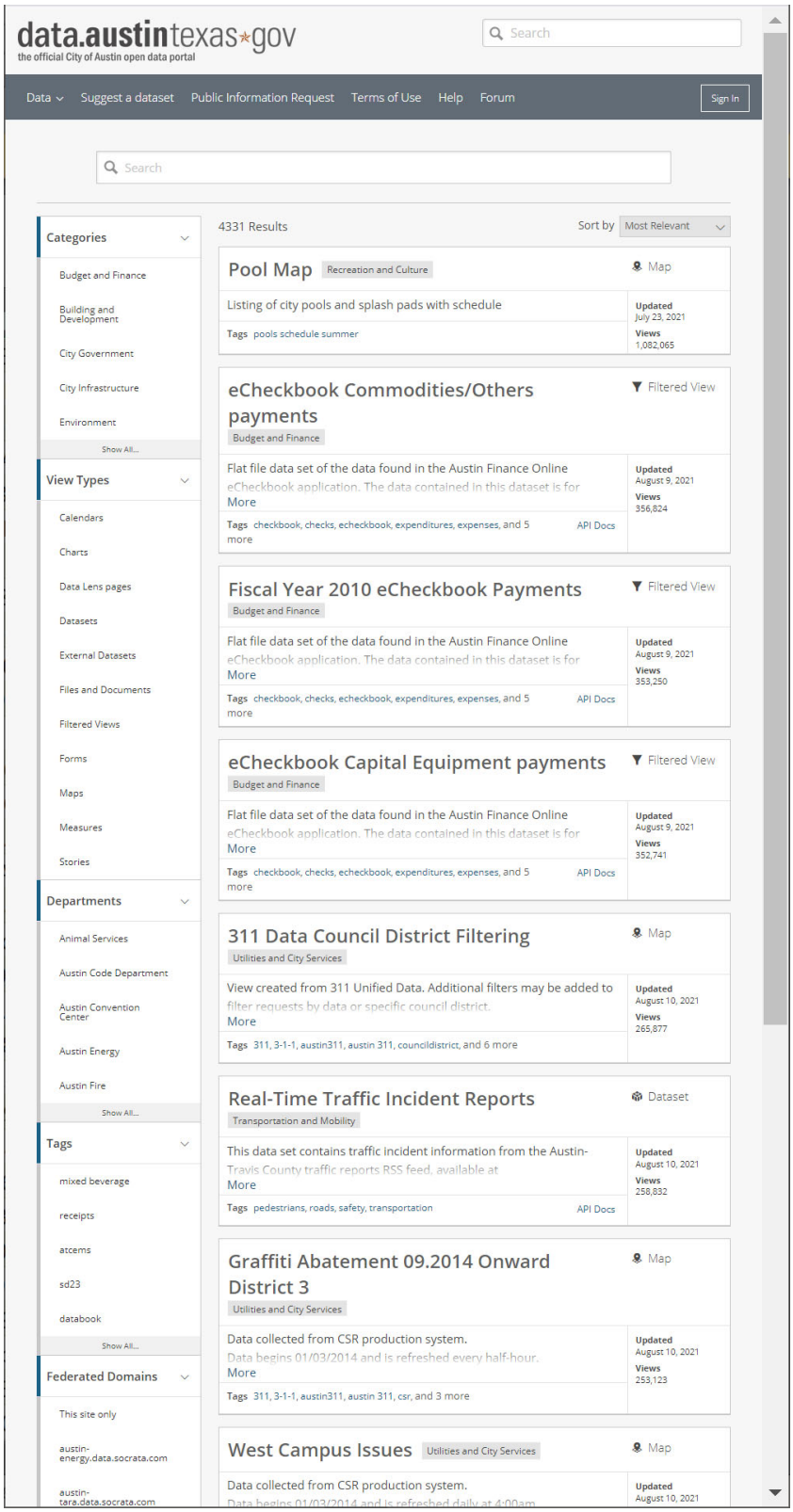

Figure 4.20. Dataset Query Page of the City of Austin, Texas, Open Data Portal (https://data.austintexas.gov/browse) Implemented Using Socrata

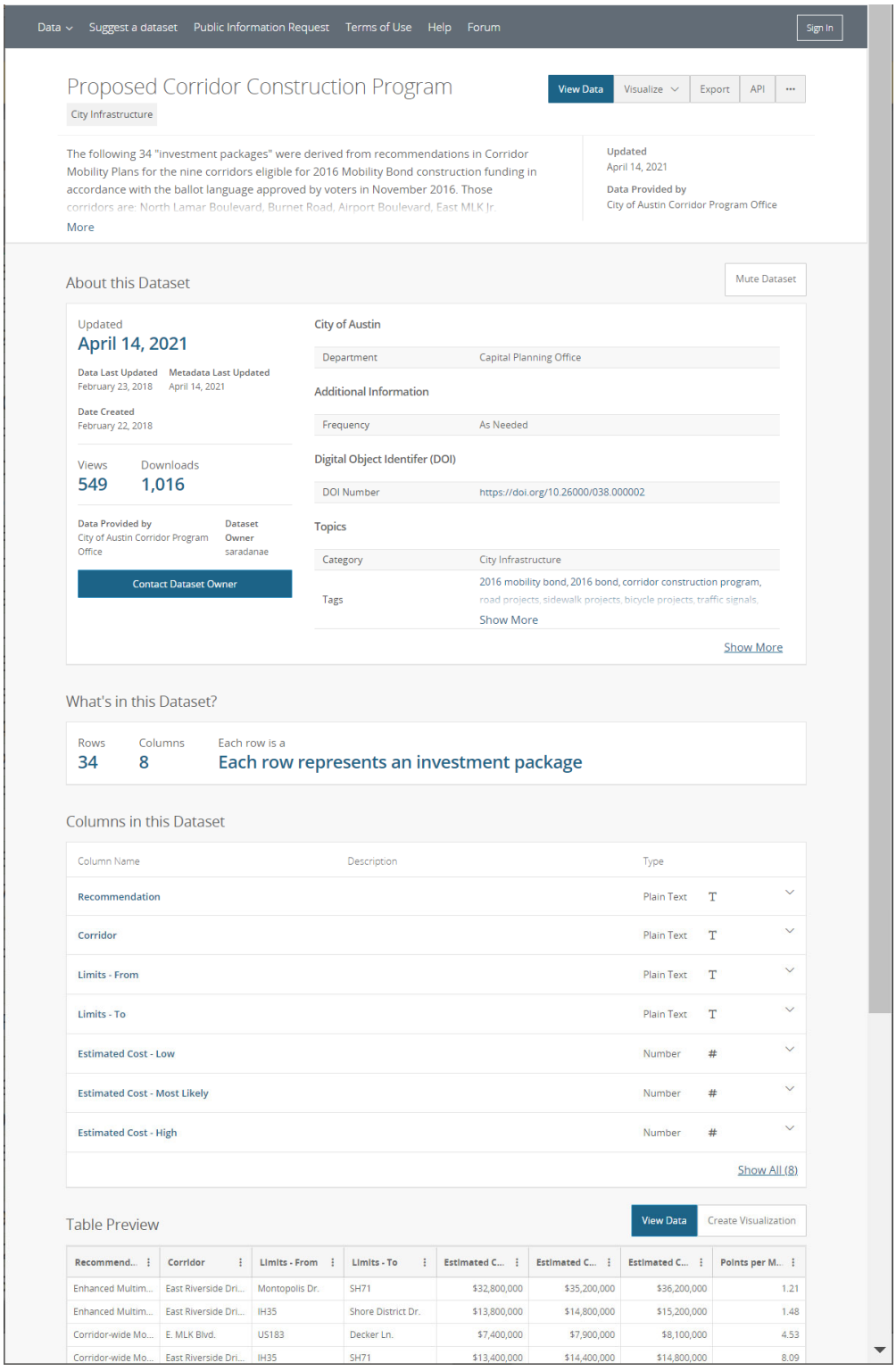

Figure 4.21. Dataset Detail Page of the City of Austin, Texas, Open Data Portal (https://data.austintexas.gov/City-Infrastructure/Proposed-Corridor-Construction-Program/psd6-eiah) Implemented Using Socrata

4.7.2 Ratings: Socrata

Table 4.14 lists the Socrata ratings and justification information for each category of requirements/desired features.

Table 4.14. Ratings for Socrata

Section		Rating	Justification
3.2.1	Catalog	E	Native tools are limited, but customization and extensions fill gaps
3.2.2	Find	E	Themes, keywords, related
3.2.3	Retrieve	E	Table, map, plot, download, API
3.2.4	Large dataset storage	E	Multiple options through extensions
3.3.1	Hanford Site software approval	E	In use by U.S. federal government (General Services Administration)
3.3.2	Single sign-on	E	Through extension
3.3.3	Role-based access	E	User types, custom roles, and privileges
3.3.4	Linking/Federation	E	Data harvesting and federation
3.3.5	Deployment	M	On-premises with new infrastructure
	Overall rating	M	

4.8 Do-It-Yourself (Custom Data Catalog)

The growing availability of open-source data portal software creates the possibility of combining modules and modifying functions to create a data repository to fit any need. CKAN and DKAN (detailed above) are examples of open-source data portal software solutions with extensive support and user communities. Other open-source solutions provide a variety of capabilities to fit, or be modified to fit, most use cases. The inherent customizability of open-source solutions means almost any need for Hanford Site environmental data could be addressed by adding or building in the needed modules. For promising tools that have yet to find mainstream deployment, the cost of implementing functions beyond those typically included would need Hanford resource support. This contrasts with the open-source tools CKAN and DKAN, which allow the same customization options but benefit from numerous developers sharing solutions with a broader user community.

A do-it-yourself solution was not included in the ratings. It would likely rate poorly as a collection of uncoordinated modules, but would rate much higher with customizations that would meet requirements and desired features. One element not considered in this evaluation is cost of implementation. If extensive customizations are needed, however, CKAN and DKAN not only offer many available extensions and have large support communities, but are provided without cost to the user.

5.0 Conclusions

Data catalog software tools are widely available and have found application in many fields. The variety of capabilities available across that range of tools makes them candidates for implementing HEIDI. This evaluation found that some software alternatives are better suited to meeting the needs of Hanford environmental data users or provide options that facilitate implementation on the HLAN.

This evaluation did not consider the costs of implementation. Promising alternatives need to be examined further with respect to the sustainability of maintaining the data catalog into the long-term stewardship period. That examination must consider whether costs and fees are in balance with ease of use and maintenance.

The propriety SaaS model of delivering a data catalog (e.g., found in software such as Junar and OpenDataSoft) favors consistency across customers at the expense of customization and configurable roles. These are important requirements for HLAN workflows. These proprietary systems may limit the size of datasets loaded to their cloud service. That limitation may be partially overcome by the ability to federate datasets hosted elsewhere, but that may limit control over metadata and custom configuration of the data vocabularies for interoperability. This was the primary driver behind the overall rating of 'F' for Junar and OpenDataSoft (Table 5.1).

Table 5.1. Ratings for Data Catalog Software Candidates

Section	Description	ArcGIS	CKAN	DKAN	EDX	Junar	OpenDataSoft	Socrata
3.2.1	Catalog	E	E	E	M	F	E	E
3.2.2	Find	E	E	E	M	E	M	E
3.2.3	Retrieve	E	E	E	E	M	E	E
3.2.4	Large dataset storage	E	E	E	M	F	F	E
3.3.1	Hanford Site software approval	E	E	E	E	M	M	E
3.3.2	Single sign-on	E	E	E	F	F	E	E
3.3.3	Role-based access	E	E	E	M	F	M	E
3.3.4	Linking/Federation	E	E	E	E	E	E	E
3.3.5	Deployment	E	M	M	M	F	F	M
	Overall rating	E	M	M	F	F	F	M

EDX is an existing U.S. government website developed by the National Energy Technology Laboratory using CKAN. As a cloud data catalog, it provides many of the same benefits as Junar or OpenDataSoft but is designed to find, connect, curate, use, and re-use data to advance fossil energy and environmental research and development. That purpose results in cataloging functions similar to those envisioned for HEIDI, but EDX users would not benefit from customizations to create Hanford-specific capabilities. The overall rating of 'F' for EDX, was primarily due to an inability to incorporate authentication from HLAN into the system.

The Socrata platform, with a similar SaaS delivery model as Junar and OpenDataSoft, provides custom options and gateways that would allow for requirements to be met and desired features to be delivered. FedRAMP approval also helped the deployment rating. These capabilities allowed Socrata to achieve an overall rating of ‘M’ along with other mature, yet flexible alternatives that provided self-hosted options, including CKAN, and DKAN. A do-it-yourself approach of combining software modules was not rated because the resources required would be impractical, given existing (open source) software capabilities (e.g., CKAN or DKAN).

ArcGIS Enterprise Sites achieved an overall rating of ‘E’, in part because of the existing implementation of ArcGIS Enterprise on the HLAN. Although this doesn’t impact the resulting end-user experience, the deployment rating benefited from fewer uncertainties about approvals, implementation, infrastructure, and integration of the data catalog into the HLAN.

The software candidates with overall ratings of ‘M’ or ‘E’ (ArcGIS, CKAN, DKAN, and Socrata) are likely to deliver the requirements and many desired features for the implementation of HEIDI. Further evaluation is needed to assess licensing terms, costs of acquisition, maintenance requirements, and sustainability through long-term stewardship at Hanford.

6.0 Quality Assurance

This work was performed in accordance with the Pacific Northwest National Laboratory Nuclear Quality Assurance Program (NQAP). The NQAP complies with DOE Order 414.1D, *Quality Assurance*. The NQAP uses NQA-1-2012, *Quality Assurance Requirements for Nuclear Facility Application,* as its consensus standard and NQA-1-2012, Subpart 4.2.1, as the basis for its graded approach to quality.

7.0 References

Atz U, E Broad, J Tennison, T Sasse, A Smith, and P Wells. 2020. *Recommendations for Open Data Portals: from setup to sustainability*. European Data Portal.

CHPRC-03503. 2017. *Hanford Environmental Data Management (HEDM) Assessment Report*. CH2M Hill Plateau Remediation Company, Richland, Washington.

DOE Order 414.1D. *Quality Assurance*. U.S. Department of Energy, Washington, D.C.

Executive Order 12906. "Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure." *Federal Register* 59(71), Wednesday, April 13, 1994.

FGDC-STD-001-1998. 1998. *Content Standard for Digital Geospatial Metadata*. Metadata Ad Hoc Working Group, Federal Geographic Data Committee, Washington, D.C.

Grzenda M and J Legierski. 2021. "Towards Increased Understanding of Open Data Use for Software Development." *Information Systems Frontiers* 23(2):495-513. https://doi.org/10.1007/s10796-019-09954-6. <Go to ISI>://WOS:000498013100001

NQA-1-2012. *Quality Assurance Requirements for Nuclear Facility Applications*. The American Society of Mechanical Engineers, New York, New York.

Schrack A. 2021. *Guide to Creating, Using, and Maintaining Open Data Portals*. Safe Software, Inc. Last Modified 4/20/2021. Accessed 7/29/2021. https://www.safe.com/blog/2021/04/guide-creating-using-maintaining-open-data-portals/.

Wilkinson MD, M Dumontier, IJ Aalbersberg, G Appleton, M Axton, A Baak, N Blomberg, J-W Boiten, LB da Silva Santos, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3(1):160018. https://doi.org/10.1038/sdata.2016.18.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov