A Descriptive Analysis of DCAT-Compliant Data Catalogs: Characteristics and Themes Across Federal, State, County, City, and Territorial Entities in the United States

Hants Williams[†]

Department of Applied Health Informatics. Stony Brook University, Stony Brook, NY. United States

Keywords: DCAT; United States; Data Catalogs; Metadata; Government

Citation: Williams H.: A Descriptive Analysis of DCAT-Compliant Data Catalogs: Characteristics and Themes Across Federal, State, County, City, and Territorial Entities in the United States. Data Intelligence, Vol. 7, Art. No.: 2025r23, pp. 647–666, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2024.0006

ABSTRACT

The increasing availability of government data has prompted efforts to standardize data cataloging practices for enhanced accessibility and usability. The primary aim of this study is to descriptively assess data catalog and referenced dataset volume, metadata utilization, and thematic composition of United States DCAT compliant data catalogs across Federal, State, County, City, and Territory entities.

Data collection involved compiling a list of relevant government agencies and data resources to identify data catalogs. DCAT compliance was then assessed, and metadata from compliant catalogs was extracted. Thematic mapping utilized Python packages RegEx and FuzzyWuzzy to categorize themes into eight standard categories. A combination of descriptive statistics and 1-way ANOVA tests were conducted to analyze dataset volume, metadata utilization, and reported themes.

Of the 305 data catalogs identified, 259 were found to be DCAT compliant. Federal entities exhibited the highest DCAT compliance rates (92.3%), followed by County (88.1%), City (86.9%), and State (77.0%), while Territory (0%) had no compliant data catalogs. Descriptive analysis revealed that federal DCAT compliant data catalogs (n = 59) had the highest average number of data assets across their data catalogs (μ = 1,133.2) with the predominant themes being transportation at 21.2% (n = 14,785) and geospatial at 15.4% (n = 10,761). While county data catalogs (n = 52) had the lowest average (μ = 232.6) with the most referenced themes being geospatial at 77.6% (n = 8450) and finance at 2.4% (n = 270).

[†] Corresponding author: Hants Williams (E-mail: hantsawilliams@gmail.com; ORCID: 0000-0003-1447-2327).

After applying thematic mapping to eight standard categories, the three most dominant themes across all entities were transportation at 38.1% (n = 16,504), natural resources with 19.6% (n = 8,501), and health and safety with 14.7% (n = 6,367).

These findings underscore the widespread adoption of the DCAT standard across government entities, with notable gaps at the territorial level. Federal and state entities exhibited the highest data catalog and dataset volumes, while metadata utilization remained relatively consistent across all entity levels. The thematic analysis highlights the importance of standardization efforts to enhance thematic consistency and facilitate effective data interpretation. Further collaboration and investment are warranted to address gaps in catalog coverage and establish standardized data cataloging practices to maximize the accessibility and usability of these data catalogs along with their referenced datasets.

1. INTRODUCTION

The Open Government Data Act (OGDA) of 2019 marked a significant milestone in the advancement of data accessibility and government transparency within the United States [1]. The OGDA mandates federal agencies must publish their data in machine-readable formats with a standardized approach: Data Catalog Vocabulary (DCAT) [2]. DCAT is a Resource Description Framework (RDF) vocabulary designed to facilitate interoperability between data catalogs published on the Web [3]. It defines a schema and provides examples for its use, enabling a data publisher to describe datasets and data services in a digital catalog using a standard model and vocabulary. The application of DCAT by United States agencies is intended to facilitate faster discovery, easier access, and utilization of the referenced datasets contained within the data catalog [4] by public and private entities.

DCAT compliance is crucial for several reasons. First, it enhances the findability, accessibility, interoperability, and reusability (FAIR) of datasets, which are key principles in the management and dissemination of scientific data [5]. By adhering to these principles, DCAT-compliant data catalogs support broader efforts in open data and government transparency, ensuring that datasets are more easily discoverable, accessible to a wider audience, and can be integrated and reused across different platforms and applications. This compliance is integral to fostering an open data ecosystem where data from various sources can be seamlessly combined to drive innovation, support data-driven decision-making, and promote accountability.

In its first five years (2019 to present) the DCAT standard has not only been adopted by federal entities, but also by state, county, and city governments that are not required by the OGDA to share their data. This movement of transparency and accountability to standardize the practice of data dissemination has subsequently created a growing set of invaluable resources for public and private stakeholders, of which include researchers, policymakers, and the public [6, 7]. These and other stakeholders can then leverage these catalogs to support data-driven decision-making and policy development across a wide array of topics ranging from public health, transportation, judicial, to agriculture and military [8, 9].

Despite significant strides in data publication and accessibility, there remains a notable scarcity of research and publicly available reports that examine similarities and differences in metadata components of DCAT-compliant data catalogs produced by government entities within the United States. The United States General Service Administration (GSA) released a report in 2021 that reviewed four federal agencies, and of those four found that none were fully DCAT compliant [10]. In addition, the GSA manages the data.gov platform that serves as a centralized repository of data catalogs and data assets across various levels of government, but this does not provide any in-depth analyses of referenced data catalogs or the included metadata they contain [11].

Partially based on the 2021 GSA report, and on our own anecdotal first-hand experience of accessing data across federal, state, county, and city data catalogs, we believe major differences exist not only in the volume and variety of datasets referenced by these data catalogs at the entity level (e.g., federal versus city), but that discrepancies may also exist in the classification of metadata across data catalogs. Our preliminary observations suggest a lack of consistency in how data is classified, commonly labeled as a 'theme' attribute, between different government entities for the same types of data. As an example, the classification of 'health' themed data can be referred to or described in several different ways by a federal, state, or city level entity. At one entity level or agency it may be categorized under "public safety", while at other entity levels or agencies it may be labeled as a "public health" or a "medical" theme. These inconsistencies pose a challenge to data discovery and access by stakeholders, potentially impeding the objectives of the OGDA to foster government transparency, accountability, and the ability to conduct larger more systematic reviews across entity levels on themes of interest.

Therefore, this paper's primary objective is to conduct a retrospective study of DCAT-compliant data catalogs across the United States to generate preliminary observations through descriptive analyses of refenced metadata at federal, state, county, city, and territorial levels; and subsequently offer preliminary insights and recommendations that may be used to enhance the standardization, accessibility, and utility of these data catalogs in the United States.

2. METHODS

2.1 Data Collection

The initial phase of data collection focused on federal agencies; a process initiated by compiling a list of relevant agencies from the Office of Personnel Management (OPM) website [12]. This list served as the foundation for identifying federal data catalogs, with the goal of ensuring representation from a diverse array of agencies across various domains and functions. Each agency's website was then visited to locate their data catalog. In instances where the agency did not appear to have a data catalog referenced or linked on their main page or search tool, a targeted Google search strategy was employed. By querying the agency name (e.g., Department of Justice) with the term "data catalog" in Google we attempted to double check that the data catalog was not accidently missed. The goal for the federal entity level was to identify a minimum of 50 data catalogs to achieve adequate representation.

After each federal data catalog was identified we determined if the catalog utilized the DCAT (Data Catalog Vocabulary) standard [2]. To determine if the DCAT standard was used we examined the metadata of each data catalog's data dictionary. DCAT compliant catalogs feature a JSON file that will conform to a prescribed structure, therefore catalogs were deemed DCAT if they contained the key-value JSON pairs of "@type": "dcat:Catalog" and "conformsTo": "https://project-open-data.cio.gov/v1.1/schema" within their metadata [13]. If a JSON file contained these two key attributes we labeled the data catalog as 'DCAT compliant'. This process ensured that only datasets adhering to the DCAT standard were included in the analysis, enhancing the reliability and comparability of the collected data.

It is important to note that some United States government efforts have taken a more granular approach to assessing 'full DCAT compliance', such as the GSA's DCAT validator that assesses if 12 mandatory attributes are present, and the JSON key:value structured is followed [14]. Because our primary objective is to accurately describe the metadata detailed within the data catalogs, rather than emphasizing strict adherence to compliance metrics, we decided to take a more generalized approach. Non-DCAT data catalogs were excluded from the analysis due to their non-standardized structure, which can pose challenges for automated processing and comparative analysis. The frequency of Non-DCAT compliant catalogs was documented in a CSV file to provide transparency and context for the Non-DCAT compliant datasets.

For DCAT compliant data catalogs, the JSON file containing the data catalogs metadata was then saved to the non-relational database MongoDB [15]. Non-relational databases like MongoDB offer flexibility in schema design and do not enforce fixed column or variable structures like traditional relational databases, allowing for dynamic adaptation to the diverse metadata schemas encountered during the data collection process [16]. This flexibility is essential given the inherent variability observed in the structure and content of data catalogs, even within those that follow the prescribed DCAT structure and attributes.

After saving DCAT compliant metadata from federal agencies into MongoDB, we replicated the same process for data catalog identification across state, county, city, and territorial levels. To ensure an adequate volume and variety of data catalogs across these entity levels we first utilized the official search platform for United States datasets: data.gov [17]. The goal for this phase was to capture all 50 states and territories, in addition to a minimum of 50 counties and 50 cities to provide comprehensive representation across different geographical regions.

Since there are more than 3,000 counties and 19,000 incorporated cities located within the United States, the most recent census data from the census bureau was utilized to rank-order counties and cities from most to least populous [18]. We purposefully prioritized more populous counties and cities over less populous, based on the assumption that locales with greater population densities will have the financial and personal resources required to provide a data catalog service.

Similar to our described approach for the federal level, if one of the 50 states or a well-populated county or city was not referenced by data.gov, we took the name of the entity (e.g., "State of Alaska" or "Cook County" or "New York City") along with the term "data catalog" to Google's search engine. After

a data catalog was identified we assessed the metadata confirmed to the DCAT standard using the same process as described for the federal agencies. Compliant datasets were then saved to MongoDB, and non-compliant DCAT data catalogs were documented in a CSV file.

2.2 Theme Mapping

Within most DCAT compliant metadata JSON files a 'theme' attribute is used to describe the type of data contained within an individual dataset. The DCAT-US v1.1 standard specifies that the theme variable represents the "main thematic category of the dataset" but is not a required field, leaving room for interpretation and variation across different catalog entries [4]. Given the absence of a terminology standard such as NIST or ANSI to guide responses to the theme variable, it was anticipated that there will be considerable diversity in theme descriptions across datasets. To address this challenge while acknowledging the potential for uncovering valuable high-level insights and general trends through a simple mapping process, we implemented a three-fold approach to map non-standardized theme responses to standardized ones:

2.2.1 Step 1

Python 3.11 was used to query the MongoDB database and perform frequency counts on the most documented "raw" themes across the datasets. This step provided counts of the most and least relevant themes present within the collected data. At this stage no modifications or transformations were performed on theme responses to retain the original intent of each theme as described by the data catalogs maintainers.

2.2.2 Step 2

To effectively analyze and categorize the diverse array of themes present within the collected datasets, we then established a set of thematic categories. We utilized an observational approach that drew on the thematic categories or data groups typically described on the landing pages of federal, state, city, and county data catalogs [19, 20], in addition to the most frequent themes discovered by our descriptive analysis described in 2.2.1. By leveraging these two approaches we developed a preliminary set of eight themes to be used to bucket the broad spectrum of topics represented across all entities and referenced datasets. The eight constructed themes are as follows:

1. *Agriculture and Food*: Datasets pertaining to agricultural practices, food production, and related industries.
2. *Education*: Datasets related to educational institutions, programs, and resources.
3. *Business, Economic, and Financial*: Datasets concerning economic indicators, business activities, and financial transactions.
4. *Health and Safety*: Datasets focusing on public health, medical information, and safety measures.
5. *Infrastructure*: Datasets related to physical infrastructure, such as transportation networks, utilities, and facilities.

6. *Natural Resources, Energy, and the Environment*: Datasets concerning natural resource management, energy production, and environmental conservation.
7. *Public Service, Politics, and Governance:* Datasets pertaining to government services, political processes, and administrative functions.
8. *Transportation*: Datasets related to transportation systems, networks, and services.

2.2.3 Step 3

To conduct the preliminary mapping exercise and recode raw themes into our standardized themes, the Python packages RegEx and FuzzyWuzzy were utilized. RegEx, or Regular Expressions, enables the development of rules and patterns to identify and categorize themes based on predefined criteria [21]. For example, when categorizing themes related to "Agriculture and Food," our RegEx patterns searched for keywords such as "agriculture," "food," "farm," "farming," and "crops," among others. These patterns were curated to encompass a broad range of terms commonly associated with agricultural and food-related datasets.

In addition to RegEx we employed FuzzyWuzzy's fuzzy matching (FM) algorithms to address discrepancies and variations in theme descriptions that may not match exactly our predefined criteria [22]. This approach allowed us to identify and align similar or related themes, even in cases where exact matches were not present, based on Levenshtein Distance [23]. Appendix A contains the code used for theme mapping functions.

The utilization of both RegEx and FuzzyWuzzy provides distinct advantages in the mapping process. RegEx provides a structured and deterministic approach with precise pattern matching based on predefined criteria. FuzzyWuzzy's FM approach provides a more adaptability and less pre-defined approach, enabling the identification of similar or related themes even in cases of spelling variations or partial matches. By combining both approaches we leverage the strengths of each method to result in a more comprehensive and potentially robust theme mapping process that accounts for both exact and approximate matches.

2.3 Analysis Plan

Our analytical plan encompassed a range of descriptive statistics and inferential techniques to describe the characteristics of collected metadata from the data catalogs. Python 3.11 was used to conduct all descriptive and inferential tests.

First we generated summary statistics for the number of unique data catalogs by type and the volume of datasets referenced by each entity, and then evaluated DCAT compliance (True/False) for each data catalog. This involved providing means, standard deviations, and interquartile ranges for continuous numerical values, while value frequency counts to summarize compliance characteristics across each entity's data catalogs (federal, state, county, city, and territory).

Next we quantified the unique and recurring metadata variables, in particular the 'theme' variable used to describe data sets between each data catalog. We implemented two separate one-way ANOVA's to

test for significant differences in 1) the volume of referenced datasets between each entity (federal, state, county, city, territory) and 2) metadata variable frequency between entity data catalogs. Post-hoc tests were then conducted to further explore significant findings at the .05 level.

Lastly, theme frequencies were assessed before and after applying Regex and Fuzzywuzzy transformations described in 2.2.3. First, we generated value counts of the 'raw' themes within each dataset and across all entities. Subsequently the frequencies were reassessed post-transformation to evaluate the effectiveness of the mapping process. This analysis enabled the identification of any shifts or alterations in theme frequencies, offering initial insights into the impact of our process for theme standardization on the dataset's thematic composition.

3. RESULTS

3.1 Compliance

The results of the data collection process identified a total of 305 data catalogs, of which 259 were identified as DCAT compliant and 46 as Non-DCAT compliant. Compliance levels varied across entity types, with Federal being the most compliant (92.3%), followed by County (88.1%), City (86.9%), State (77.0%), and then Territory (0.0%) with no DCAT compliant data catalogs.

Of the 259 DCAT compliant data catalogs, metadata was successfully extracted across federal, state, county, and city entities for a total sum of 142,913 referenced data assets. Like compliance, the total count of data assets listed across all data catalogs varied by entity level with Federal catalogs listing the highest number of data assets (n = 66,859), followed by State (n = 38,589), City (n = 25,365), and County (n = 12,100).

3.2 Descriptives: Entity Level

Descriptive statistics computed for DCAT compliant data catalogs revealed an overall mean of 578.60 (sd = 1,376.7) data assets per catalog. Further analysis by entity level indicated variations in the mean number of data assets, with Federal catalogs having the highest average number of data sets per data catalog and County catalogs with the lowest (Table 1). The NASA data catalog had the highest volume of data assets for the Federal level (n = 22,261), Utah's state data catalog for the state level (n = 9,689), Lake County Illinois for county level (n = 1,517), and Baltimore city at the city level (n = 3,431).

A one-way ANOVA test was conducted to examine if the differences in data volume between entity levels were significant using entity level as the independent variable, which yielded a statistically significant result ($F(4,242) = 4.2$, $p < 0.006$). The Tukey Honestly Significant Difference (HSD) test was used to compare mean differences between different pairs of entity groups at a significance level of 0.05 (Table 2). The results revealed significant mean differences between the City and Federal groups (mean difference = 838.8, $p < 0.05$) and between the County and Federal groups (mean difference = 900.5, $p < 0.05$). However, no significant mean differences were observed between the City and County groups,

the City and State groups, the County and State groups, or the Federal and State groups ($p > 0.05$ for all comparisons). These findings suggest that there are significant differences in data volume between certain pairs of entity groups, particularly between County and Federal entities, while others do not exhibit significant differences in data volume.

Table 1. DCAT Compliant Data Catalogs and Datasets by Entity Level.

	Unique Entities	Data Sets within Data Catalogs							
	Count*	Mean**	Median**	Std**	Min**	25%**	50%**	75%**	Max**
Federal	59	1133.2	132	3079.1	1	33.5	132	1012.5	22261
State	47	831.3	402	1532.6	15	152.5	402	994.5	9689
County	52	232.6	132.5	286.2	14	51.75	132.5	335.5	1517
City	98	294.3	118	609.0	1	54.5	118	235.75	3431
Territory	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

* Represents unique number of data catalogs at the entity level (e.g., 59 unique federal agencies); ** Represents mean, median, std, min, 25%, 50%, 75%, or max number of data sets within each row, representing a unique entity level.

Table 2. HSD* Results of Datasets by Entity.

Group 1	Group 2	Mean Difference	Adjusted p-value	Lower CI	Upper CI	Reject Null
City	County	–61.6	0.9	–800.2	676.8	FALSE
City	Federal	838.8	0.0	129.4	1548.1	TRUE
City	State	536.9	0.2	–226.8	1300.6	FALSE
County	Federal	900.5	0.0	81.7	1719.	TRUE
County	State	598.6	0.2	–267.7	1464.9	FALSE
Federal	State	–301.9	0.7	–1143.5	539.7	FALSE

* Tukey's Honest Significant Difference (HSD).

3.3 Descriptives: Data Catalog Level

At the data catalog level there was an average of 14.4 (*sd* = 2.4) metadata variables (e.g., attributes) per catalog, with a median metadata count of 15 indicating a relatively consistent distribution of metadata variables across the catalogs (Table 3). The range of metadata attribute counts spanned from a maximum of 30 (Federal - Department of Commerce) to a minimum of 6 (City and County - Philadelphia) highlighting the diversity in the depth and complexity of metadata descriptions among the catalogs.

Across all entity levels in the 256 DCAT compliant data catalogs, the only metadata attributes that were referenced by 100% of all 142,913 data assets were 'title', 'description', and 'contactPoint'.

Table 3. Metadata Variables by Entity.

	Count*	Mean	Median	Std	Min	25%	50%	75%	Max
Federal	59	16.2	15	4.5	10	13	15	18	30
State	47	13.9	15	1.3	10	13	15	15	15
County	52	14.0	15	2.0	6	14.5	15	15	15
City	98	13.7	15	1.9	6	13	15	15	20
Territory	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

* Represents unique number of data catalogs at the entity level (e.g., 59 unique federal agencies).

An ANOVA test for metadata level features, using a model that included catalog level as the independent variable, yielded a significant result ($F(4,251) = 12.10$, $p < 0.001$). The results of the TSD test at a level of 0.05 significance indicated significant mean differences (Table 4) between the City and Federal groups (mean difference = 2.5432, $p < 0.05$), the County and Federal groups (mean difference = 2.2304, $p < 0.05$), and the Federal and State groups (mean difference = –2.3520, $p < 0.05$). However, no significant mean differences were observed between the City and County groups, the City and State groups, or the County and State groups ($p > 0.05$ for all comparisons). These findings suggest that there are significant differences in metadata volume between certain pairs of entity groups, particularly between City and Federal entities, County and Federal entities, and Federal and State entities, while others do not exhibit significant differences in metadata volume.

Table 4. HSD* Results of Datasets by Entity.

Group 1	Group 2	Mean Difference	Adjusted p-value	Lower CI	Upper CI	Reject Null
City	County	0.3	0.9	–0.8	1.5	FALSE
City	Federal	2.5	0	1.3	3.7	TRUE
City	State	0.1	0.9	–1.0	1.4	FALSE
County	Federal	2.2	0.0	0.8	3.5	TRUE
County	State	–0.1	0.9	–1.5	1.2	FALSE
Federal	State	–2.3	0.0	–3.7	-0.9	TRUE

* Tukey's Honest Significant Difference (HSD).

3.4 Themes

Despite being a non-mandatory field, the "theme" attribute was utilized in 132,047 (92.3%) of the 142,913 datasets across the 256 catalogs, with 2,695 unique themes identified. Among the top 15 themes, "geospatial" emerged as the most prevalent theme, with a count of 40,244 (30.4%) indicating a significant focus on spatial data across various datasets (Table 5). Other prominent themes included "Transportation" (n = 16,014), "Earth Science" (n = 10,297), "Natural Resources" (n = 3,119), and "Education" (n = 2,734), highlighting the breadth of subject areas covered within the datasets.

Table 5. Top 15 Themes across all Entity Levels.

Rank	Theme	Percent (Count)
1	Geospatial	30.4% (40244)
2	Transportation	12.1% (16014)
3	Earth Science	7.7% (10297)
4	Natural Resources	2.3% (3119)
5	Education	2.0% (2734)
6	Water	2.0% (2722)
7	Health	1.7% (2300)
8	Public Safety	0.9% (1231)
9	City Government	0.8% (1104)
10	Health and Human Services	0.7% (943)
11	Finance	0.6% (803)
12	Agriculture	0.5% (780)
13	Energy & Environment	0.5% (771)
14	Housing & Properties	0.5% (740)
15	Government	0.5% (717)

Analysis of themes by entity level further elucidated differences in thematic distribution (Table 6). The predominant themes at the federal level were "Transportation" (n = 14,785) and "geospatial" (n = 10,761), while state level themes first emphasized "geospatial" (n = 10,784) followed by "Natural Resources" (n = 3,119). Similarly, county datasets were led by "geospatial" (n = 8,450) then "Finance" (n = 270), and city datasets with "geospatial" first (n = 10,249) and "Education" (n = 1,377) second.

Table 6. Top 10 Themes by Entity Levels.

Rank	Type	Theme	Percent (Count)
1	Federal	Transportation	21.2% (14785)
2	Federal	Geospatial	15.4% (10761)
3	Federal	Earth Science	14.8% (10297)
4	Federal	Research and Statistics	0.9% (680)
5	Federal	Agriculture	0.9% (667)
6	Federal	Operational Data	0.8% (598)
7	Federal	Basic Statistics	0.8% (598)
8	Federal	Use	0.8% (598)
9	Federal	NNDSS	0.8% (590)
10	Federal	Roadways and Bridges	0.7% (492)
1	State	Geospatial	35.5% (10784)
2	State	Natural Resources	10.2% (3119)
3	State	Water	8.9% (2716)
4	State	Health	6.1% (1868)
5	State	Education	4.3% (1315)
6	State	Health and Human Services	3.0% (929)
7	State	Energy & Environment	2.5% (770)
8	State	Transportation	1.9% (576)
9	State	Economy and Demographics	1.4% (425)
10	State	Public Safety	1.2% (390)
1	County	Geospatial	77.6% (8450)
2	County	Finance	2.4% (270)
3	County	GIS/Maps	1.8% (205)
4	County	Finance & Administration	1.6% (177)
5	County	Government	1.4% (154)
6	County	Services	1.3% (148)
7	County	Health	1.3% (142)

Table 6. *Continued.*

Rank	Type	Theme	Percent (Count)
8	County	Voting & Elections	1.2% (131)
9	County	Public Safety	0.9% (106)
10	County	GIS	0.8% (91)
1	City	Geospatial	48.0% (10249)
2	City	Education	6.4% (1377)
3	City	City Government	5.1% (1104)
4	City	Housing & Properties	3.4% (740)
5	City	Public Safety	3.4% (732)
6	City	Transportation	2.8% (608)
7	City	Finance	2.2% (477)
8	City	Environment	1.4% (303)
9	City	Geospatial	1.3% (287)
10	City	Government	0.9% (203)

Following the descriptive analysis of raw themes, the themes were mapped to one of eight standardized theme categories with Regex and FuzzyWuzzy. Out of the 132,047 themes identified across the 142,913 data assets, 43,257 themes (32.7%) were successfully mapped to one of the eight new standardized categories. Table 7 highlights the ranking of theme prevalence, with Transportation emerging as the most prominent category with 16,504 occurrences (38.1%), followed by Natural Resources, Energy, and the Environment (19.6%), Health and Safety (14.7%), Public Service, Politics, and Governance (8.7%), and Education (8.3%).

Table 8 illustrates the distribution of mapped themes by the four entity levels. Across all state, county, and city data catalogs the Public Service, Politics, and Governance theme appeared consistently, ranging from 10.6% (state level) to 25.02% (county level). Similarly, Health and Safety themes are prominently featured across City, County, Federal, and State datasets suggesting a universal emphasis on public well-being and safety measures. However, there are notable differences in the occurrence of certain themes among the entities. Transportation emerged as the dominant theme only at the federal entity level (78.8%). Similarly, the theme of Agriculture and Food was only prevalent at the federal entity (6.00%) compared to State, City, and County where it accounted for less than 1% for each level.

Table 7. Mapped Themes.

New Theme	Percent (Count)
Transportation	38.1% (16504)
Natural Resources, Energy, and the Environment	19.6% (8501)
Health and Safety	14.7% (6367)
Public Service, Politics, and Governance	8.7% (3764)
Education	8.3% (3593)
Business, Economic, and Financial	6.3% (2708)
Agriculture and Food	3.0% (1302)
Infrastructure	1.1% (518)

Table 8. Mapped Themes by Entity Level.

Type	New Theme	Percent (Count)
City	Public Service, Politics, and Governance	21.9% (1465)
City	Education	20.8% (1394)
City	Health and Safety	17.7% (1187)
City	Business, Economic, and Financial	14.5% (968)
City	Transportation	11.4% (767)
City	Infrastructure	6.4% (430)
City	Natural Resources, Energy, and the Environment	6.3% (421)
City	Agriculture and Food	0.6% (41)
County	Health and Safety	32.8% (399)
County	Business, Economic, and Financial	26.2% (319)
County	Public Service, Politics, and Governance	25.0% (304)
County	Natural Resources, Energy, and the Environment	7.0% (86)
County	Transportation	5.5% (67)
County	Education	3.1% (38)
County	Infrastructure	0.1% (2)

Table 8. *Continued.*

Type	New Theme	Percent (Count)
Federal	Transportation	78.8% (15040)
Federal	Health and Safety	6.3% (1206)
Federal	Agriculture and Food	6.0% (1146)
Federal	Natural Resources, Energy, and the Environment	3.2% (628)
Federal	Education	2.4% (476)
Federal	Business, Economic, and Financial	1.4% (280)
Federal	Public Service, Politics, and Governance	1.3% (257)
Federal	Infrastructure	0.2% (53)
State	Natural Resources, Energy, and the Environment	45.2% (7366)
State	Health and Safety	21.9% (3575)
State	Public Service, Politics, and Governance	10.6% (1738)
State	Education	10.3% (1685)
State	Business, Economic, and Financial	7.0% (1141)
State	Transportation	3.8% (630)
State	Agriculture and Food	0.7% (115)
State	Infrastructure	0.2% (33)

4. DISCUSSION

Our analysis revealed a notable utilization of data catalogs across federal, state, county, and city entities in the United States (n = 305). However, a non-trivial amount of Non-DCAT compliant (n = 46) and data catalogs not referenced by data.gov (n = 173) were identified. By implementing a multifaceted search methodology to capture data catalogs, as opposed to relying on a single search approach such as the GSA's data.gov platform which only referenced 132 data catalogs in a .csv file at the time of this study [2], we discovered an additional 173 data catalogs. This underscores a gap in coverage in data.gov as of February 2024, and emphasizes the critical importance of having a comprehensive search methodology to generate a list of data catalogs.

In our study, 84% (259 out of 305) of the included data catalogs were DCAT compliant. This compares favorably with the findings from the 2023 Open Data Maturity Report for Europe, where 89% (24 out of 35) of the participating countries reported DCAT compliance [24]. Our study aligns with the European

report in showing high penetration of DCAT compliance for data catalogs. However, we are not aware of other studies that compare the utilization of alternative standards such as CKAN, Socrata, Inspire, or Open Geospatial Consortium (OGC) in relation to DCAT. The remaining 10-15% of catalogs that do not adopt DCAT standards warrant further investigation. Understanding whether the reasons for non-adoption are financial, technical, or due to other barriers is crucial for addressing these gaps. Identifying and mitigating these challenges will be essential to achieving broader compliance and maximizing the benefits of standardized data cataloging practices.

An intriguing observation emerged regarding the absence of structured DCAT-compliant data at the territorial level. None of the five U.S. territories currently maintain DCAT-compliant catalogs. A recent 2024 U.S. Government Accountability Office (GAO) report acknowledged a general gap in data collection and statistical reporting within the U.S. territories [25], aligning with the findings reported here. The absence of DCAT-compliant data catalogs in these territories may be attributed to a mix of several factors.

First, funding constraints are likely a significant issue for U.S. territories as they often lack the financial resources necessary for developing and maintaining technology systems due to their smaller economies and reliance on federal funding. For example, the U.S. territories have received more than $32 billion in COVID-19 relief funds through over 100 federal programs [26].

Additionally, a shortage of personnel that are technology literate in data management, metadata standards, and information technology may be impeding DCAT utilization in the U.S. territories. We are not aware of any studies or research that have examined the potential impact of technology literacy on DCAT adoption. This represents an intriguing area for future investigation, as understanding the role of technology literacy could inform strategies to enhance the implementation and effectiveness of DCAT-compliant data catalogs. Furthermore, investigating the implications of technology literacy on user design and user experience, particularly considering the unique cultural and geographical contexts of the U.S. territories, is essential. These factors may compound low literacy levels, significantly impacting how users interact with data catalogs. Identifying how these varying levels of technology literacy affect user interactions could reveal unseen barriers and inform the development of more user-friendly designs and provide valuable insights for future planning and capacity-building efforts.

Lastly, there may exist a general lack of awareness or understanding of the importance and benefits of DCAT-compliant data catalogs, leading to lower prioritization and slower adoption of these modern data management practices by local governments and institutions. Communication gaps and differing local priorities may result in federal initiatives not being appropriately communicated or disseminated at the territorial level.

To address this gap, we recommend that data officers at federal or state levels assist the territories in strategizing how to develop and deploy their own catalogs. Collaboration and knowledge-sharing from federal and state entities to territories could help leverage existing resources and expertise to establish standardized data cataloging practices at the territorial level. Additionally, targeted funding initiatives and

capacity-building efforts could provide vital support to territories in overcoming obstacles and ensuring the availability of comprehensive and accessible data resources for their constituents.

Dataset volume was greatest within data catalogs at the federal and state levels, with federal datasets significantly surpassing city and county volumes, but aligning closely with state figures. Despite this delta, metadata attribute utilization remained relatively consistent across entity levels with means ranging from 13.7 (sd = 1.9) to 16.2 (sd = 4.5). This suggests a growing standardized approach to metadata attribute utilization across federal, state, county, and city levels despite a small variance.

Regarding our thematic analysis, the prominent raw themes that emerged were geospatial, transportation, and earth science ranking among the top categories. Notably, the prevalence of the earth science theme is likely attributed to the disproportionately high volume of datasets from NASA (n = 22,261). While our subsequent thematic mapping process revealed a convergence towards the themes of 1) transportation, 2) natural resources, energy, and the environment, and 3) health and safety.

We believe that based on our initial raw thematic analysis and initial theme mapping, that the ability for a stakeholder to perform any thematic trend analysis across "local" county and or city level's will be unique challenge not only due to the heterogeneity in data cataloging practices and thematic representation as show in Tables 5-8, but also due to the sheer volume of counties and cities in the United States. This variability of mapped and unmapped themes underscores the need for future standardization efforts across federal, state, county, and city entities to enhance thematic consistency and facilitate effective data interpretation and utilization.

While our manual, rule-based mapping with RegEx and FuzzyWuzzy has proven valuable as an initial exploratory approach, the role of more sophisticated approaches like natural language processing (NLP) cannot be overlooked and should be investigated in the future [27]. We believe that either apriori or posteriori application of NLP could streamline the thematic analysis process and contribute to more accurate and standardized categorization of datasets beyond the approach taken here. But regardless of the future approach selected and tested, this thematic challenge will require a collaborative effort across entities to establish a common interoperable set of attribute-level response types that is additive to the existing DCAT standard.

Another significant finding of this paper is that despite the adoption of the DCAT standard and the use of FAIR principles by U.S. entities, there remains significant variability in the specific concepts and themes represented within the data. This inconsistency necessitates additional analytical cleansing, such as the techniques we employed in this paper using FuzzyWuzzy and RegEx. While DCAT provides standardized fields, the lack of a uniform vocabulary complicates data integration and analysis. For U.S. entities there is a need for adoption of existing standardized vocabularies into DCAT, such as North American Industry Classification System (NAICS) codes from the Census Bureau [28] or Business Activity Codes [29] from the Internal Revenue Service (IRS), as potential examples. Utilizing such standardized vocabularies would enhance clarity and consistency in data representation, reducing the need for extensive data cleaning and improving the overall utility of

the data catalogs. This approach would help ensure that the data is not only well-structured but also easily interpretable and comparable across different datasets and agencies.

There are several limitations of our study. First, while we identified datasets that were DCAT compliant our approach treated all compliant datasets equally without considering potential variations in the level or degree of compliance. Some datasets may adhere more closely to the DCAT standard than others, which could influence or moderate the frequency of themes or other descriptive information represented within our preliminary analysis. Second, our method of deriving the eight mapped themes relied on the combination of observational approach, subject matter expertise, and the use of RegEx and FuzzyWuzzy algorithms. While these techniques succeeded in helping us to develop an initial process for categorization that we believe is justifiable for a preliminary or exploratory analysis, our approach is subject to human interpretation and lacks a degree of objectivity. Future research could explore more objective methods for theme identification and classification to enhance the reliability and validity of the reported findings.

5. CONCLUSION

Our analysis of United States government data catalogs underscores the widespread adoption of the DCAT standard across federal, state, county, and city levels, while revealing a significant gap at the territorial level. The absence of structured DCAT-compliant data catalogs for U.S. territories highlights the imperative for future collaboration and investment in establishing compliant catalogs. Federal and state entities displayed higher dataset volumes and metadata consistency compared to city data catalogs. And while our thematic analysis revealed a diverse range of dominant categories that included transportation, natural resources, and health and safety, the variability of themes by entity level emphasizes the importance and need for standardization efforts. Overall, our findings reflect a growing trend in DCAT adoption for cataloging across government levels, while emphasizing the ongoing need for innovation and collaboration to standardize response attributes to maximize interoperability and utility of these data catalogs.

AUTHOR CONTRIBUTIONS

HW was the sole author of this manuscript and was responsible for all aspects of the research and publication, including conceptualization, data collection, analysis, and writing.

SUPPLEMENTARY MATERIALS

GitHub Repository: https://github.com/hantswilliams/publication_dcat_descriptives

REFERENCES

[1] H.R.4174-Foundations for Evidence-Based Policymaking Act of 2018. Available at: https://www.congress.gov/bill/115th-congress/house-bill/4174. Accessed 15 January 2024

[2] United States: Open Government. Available at: https://data.gov/open-gov/. Accessed 15 January 2024

[3] W3: Data Catalog Vocabulary (DCAT)–Version 2. Available at: https://www.w3.org/TR/vocab-dcat/. Accessed 31 May 2024

[4] DCAT-US Schema v1.1 (Project Open Data Metadata Schema). Available at: https://resources.data.gov/resources/dcat-us/#introduction. Accessed 15 January 2024

[5] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C.T., Goble, C., Guizzardi, G., Hansen, K.K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R.W.W., Imming, M., Jeffery, K.G., Kaliyaperumal, R., Kersloot, M.G., Kirkpatrick, C.R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.A., Bonino da Silva Santos, L.O., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M.D., Willighagen, E.L., Wittenburg, P., Roos, M., Mons, B., Schultes, E.: FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, Volume 2, Issue 1-2, 2020, Pages 10–29. https://doi.org/10.1162/dint_r_00024

[6] City Governments Making Public Data Easier To Get: 90 Municipal Open Data Portals. Available at: https://www.forbes.com/sites/metabrown/2018/04/29/city-governments-making-public-data-easier-to-get-90-municipal-open-data-portals. Accessed 15 January 2024

[7] National League of Cities and the Center for City Solutions and Applied Research: City Open Data Policies. Available at: https://www.nlc.org/wp-content/uploads/2016/12/CSAR-Open-Data-Report-FINAL.pdf. Accessed 15 January 2024

[8] Wilson, B., Cong, C.: Beyond the supply side: Use and impact of municipal open data in the U.S. Telematics and Informatics, Volume 58, 2021, Article 101526. ISSN 0736-5853. https://doi.org/10.1016/j.tele.2020.101526

[9] Mergel, I., Kleibrink, A., Sörvik, J.: Open data outcomes: U.S. cities between product and process innovation. Government Information Quarterly, Volume 35, Issue 4, 2018, Pages 622–632. ISSN 0740-624X. https://doi.org/10.1016/j.giq.2018.09.004

[10] Open Data: Additional Action Required for Full Public Access (GAO-22-104574). Available at: https://www.gao.gov/products/gao-22-104574. Accessed 15 January 2024

[11] The Home of the U.S. Government's Open Data. Available at: https://data.gov/. Accessed 15 January 2024

[12] U.S. Office of Personnel Management: Federal Agencies List. Available at: https://www.opm.gov/about-us/open-government/Data/Apps/Agencies/. Accessed 15 January 2024

[13] Centers for Disease Control and Prevention (CDC) Data Catalog JSON File. Available at: https://data.cdc.gov/data.json. Accessed 15 January 2024

[14] The United States DCAT Validator. Available at: https://catalog.data.gov/dcat-us/validator. Accessed 15 January 2024

[15] MongoDB. Available at: http://mongodb.com/. Accessed 15 January 2024

[16] MongoDB: What is a Non-Relational Databases? Available at: https://www.mongodb.com/databases/non-relational. Accessed 15 January 2024

[17] The Home of the U.S. Government's Open Data. Available at: https://data.gov/. Accessed 15 January 2024

[18] U.S. Census Bureau Dashboard. Available at: https://data.census.gov/profile?q=United%20States&g=010XX00US. Accessed 15 January 2024

[19] California State Open Data Portal. Available at: https://data.ca.gov/. Accessed 15 January 2024
[20] New York State Open Data Portal. Available at: https://data.ny.gov/. Accessed 15 January 2024
[21] Python Software Foundation: Regular expression operations documentation. Available at: https://docs.python.org/3/library/re.html. Accessed 15 January 2024
[22] FuzzyWuzzy Documentation. Available at: https://github.com/seatgeek/thefuzz. Accessed 15 January 2024
[23] Redis: What is Fuzzy Matching? Available at: https://redis.com/blog/what-is-fuzzy-matching/. Accessed 15 January 2024
[24] European Commission: 2023 Open Data Maturity Report. Martin Page PhD, Hajduk, E., Lincklaen Arriëns, E.N., Cecconi, G., Brinkhuis, S., December 2023. Available at: https://data.europa.eu/sites/default/files/odm2023_report.pdf. Accessed 15 May 2024
[25] U.S. Government Accountability Office: U.S. Territories: Coordinated Federal Approach Needed to Better Address Data Gaps. GAO-24-106574, May 09, 2024. Available at: https://www.gao.gov/products/gao-24-106574. Accessed 31 May 2024
[26] U.S. Department of the Interior: Federal Assistance to the U.S. Territories and Freely Associated States during the Coronavirus Disease 2019 (COVID-19) Pandemic. Available at: https://www.doi.gov/oia/covid19. Accessed 31 May 2024
[27] Zaki, M., Namireddy, S.R., Pittie, T., Bihani, V., Keshri, S.R., Venugopal, V., Gosvami, N.N., Jayadeva, Krishnan, N.M.A.: Natural language processing-guided meta-analysis and structure factor database extraction from glass literature. Journal of Non-Crystalline Solids: X, Volume 15, 2022, Article 100103. ISSN 2590-1591
[28] U.S. Census Bureau: North American Industry Classification System. Available at: https://www.census.gov/naics/. Accessed 31 May 2024
[29] U.S. Internal Revenue Service: Business Activity Codes. Available at: https://www.irs.gov/pub/irs-soi/18pf_business_codes.pdf. Accessed 31 May 2024

AUTHOR BIOGRAPHY

Dr. Hants Williams is a licensed public health nurse, published scientist, and data specialist. As a full-stack healthcare professional, he combines expertise in both frontline care and data operations. His skills enable him to translate complex analytical insights between IT professionals, care providers, and executives, allowing him to excel with emerging technologies in new care settings. Dr. Williams uses his clinical and research background to assess the feasibility and efficacy of novel digital therapeutics and data-driven care pathways. His passion and curiosity drive business innovations to enhance care delivery and health outcomes.
ORCID: 0000-0003-1447-2327