Sophie McIntyre

LIS 545

Final Report

Dataset: ‘[Resource Mapping Data by County](https://data.tn.gov/dataset/resource-mapping-data-by-county)’ from the Tennessee Commission on Children and Youth

Repository: ‘[Child Care & Early Education Research Connections](https://www.re3data.org/repository/r3d100010330)’ from Research Connections

**Data and Metadata Profile**

**Data**

The Resource Mapping Data by County dataset was created by the Tennessee Commission on Children and Youth to document yearly statistics from programs that coordinate with the TCCY from 2015 to 2021. It can be downloaded as a single CSV file from the Tennessee Open Data Portal website and contains the name of each program, the agency it resided under, the number of children it served by county, and the amount of money it spent by county. For example, in 2015 Dolly Parton’s Imagination Library, which resides under the Governor’s Books from Birth Foundation, served 9,213 children in Rutherford County and spent $116,718 on resources and personnel for the county. Each line item also specifies whether the number of children served and the expenditure are estimates or exact values.

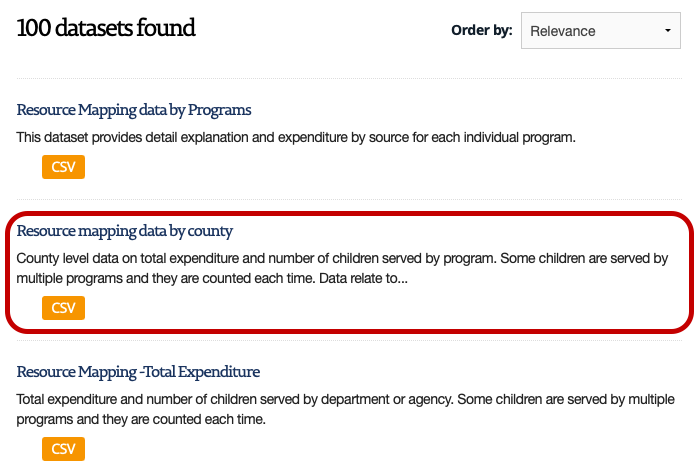
The author of the dataset is Sujit Das who is also the maintainer. Although there is no specific description or metadata that explains why the data was collated into this dataset, the TCCY is an affiliate of the Kids Count program which aims to provide legislators and program officials with data that can inform policy regarding children and families (“Kids Count”). In addition to these stakeholders, the dataset benefits the public by providing them with transparent access to how the TCCY partner programs are serving each county.

The dataset is licensed under the Creative Commons Attribution 4.0 which means users are able to share and modify the material as long as they adequately cite it (“Creative Commons License Deed”). The CSV file containing the dataset can be accessed through a web browser or through the Data Explorer widget available through the CKAN toolkit where users can sort, filter, and search the data (CKAN).  
**Metadata**

The metadata for this dataset is split into two places, one on the initial description page and one on the page that uses the Data Explorer widget. The description page metadata is brief and lists the author (Sujit Das), maintainer (Sujit Das), when it was created (October 12, 2022), and when it was last updated (October 12, 2022). The Data Explorer metadata is more detailed and lists the dataset’s license, when it was created/updated, various IDs, the file format, whether it has views, the url type, and its state. Both metadata sets contain useful information, especially the field for the Creative Commons licensing, but they don’t provide much context regarding the purpose of the data, how it was collected, what collaborators were involved, and what it contains. Since this data is used to inform not only the general public, but also policymakers, it’s imperative that the dataset has transparent metadata that can corroborate accuracy. For instance, the only creator listed was Sujit Das but there is no additional information about whether they were the only individual who worked on this project, if they worked for the TCCY, where the creator(s) got the statistics for each program, and how to contact the maintainer(s). It would also be useful to have subjects or keywords in the metadata that are free of discipline specific jargon so community members can get a general idea of what the data is showing.

The metadata is listed out directly on the web pages under an “Additional Information” header but is not publicly available in a datafile. Since it’s a public resource, the metadata should still be shown on those webpages, but there should also be either a JSON or XML file that researchers or policymakers can use. One recommendation for the metadata structure of this file is to follow the DataCite metadata schema (“Datacite Metadata schema”). Given the previous concerns on the lack of context in the original metadata, this file should primarily focus on filling out the creators, publisher, subjects, and contributors fields found in the schema.  
**Further Recommendations and Publications**

The main suggestion to improve the discoverability of the dataset in the TN Open Data Portal would be to change the title. A screenshot of how the data set appears in the browse feature of the portal can be seen below. It shows that the title, “Resource mapping data by county”, doesn’t provide much context as to what the resources are and what exactly is being mapped. The description below the title gives a better sense of what the data is showing, but it feels like both the title and description were created for internal processes of the TCCY instead of for a public repository.



By improving the title and description of the dataset, it can also improve usability for individuals who wish to incorporate the data into policies or further its scope. As previously mentioned, adding subject attributes to the dataset would also help users familiarize themselves more with the data.

There are no publications that are listed with the dataset, but upon searching with Google, a publication titled “Resource Map of Expenditures for Tennessee Children and Youth” for 2020-2021 was created by the TCCY and Tennessee State Government (Tennessee Commission on Children and Youth, 2022). It’s unclear whether the dataset was one of the primary resources for this report or if the dataset was created as a result of this report. However, it should be noted that this report by the TCCY and Tennessee State Government is created on a yearly basis so there is a good chance that the TCCY dataset was created based on these reports or was updated in parallel to these reports (Tennessee State Government). It would be beneficial to have this context explained in the metadata.

**Repository Profile**

**Repository Policies and Procedures**

According to its entry on re3data, the goal of the Child Care & Early Education (CCEE) repository is to make research and data on child care resources more available to stakeholders, like policymakers. Since one of the purposes for the resource dataset was to assist in the creation of policies regarding programs for children, I chose the CCEE repository to house this data. Additionally, the CCEE repository contains resources from Kids Count which is a program that the Tennessee Commission on Children and Youth is affiliated with (*Kids Count*, n.d.). I also found that the CCEE repository accepts state specific resources and that you can apply filters to find data from a variety of states when browsing their library (*Resources With Data From All States*, n.d.). This feature works very well with the resource mapping dataset from TCCY.

One problem that I discovered while working through this assignment is that the CCEE repository does not have a data focus, although it does contain data and is listed on the re3data website. Unfortunately, this was the only repository that I felt would be appropriate for my dataset so I wanted to still use it. I addressed this issue by looking more at the ICPSR, a consortium that I believe handles most of the data collection for the CCEE repository.

The CCEE repository has a general collection scope that is primarily based on areas of research that center on child care and early education. Although the scope can be easily located on the Research Connections website, the submission process is not as publicly accessible. The re3data metadata for the CCEE repository states that the type of data that can be uploaded is restricted and that the data upload license can be found on the Inter-university Consortium for Political and Social Research. At first, the ICPSR seemed to be unrelated to the CCEE repository, but according to a webpage from the Office of Planning, Research & Evaluation (a partner of the CCEE repository), the bibliography used by Research Connections is supported by the ICPSR (*Child Care and Early Education Research Connections*, n.d.). I’m not entirely sure if this is the workflow Research Connections is using, but I imagine that researchers can submit data to the ICPSR who publishes it to their ‘Bibliography of Data-related Literature’ which is where Research Connections pulls their bibliography from.

When looking at the ICPSR, there are no topic related limits that are explicitly stated, but I have an assumption that all submissions should be related to political and social research based on the organization’s name. As for data, the ICPSR requests that depositors submit quantitative data as SAS, SPSS, Stata, or ASCII files but it states that openICPSR can accept most file formats (*Start Deposit*, n.d.). The limits that Research Connections provide are mostly CCEE based. For instance, they collect resources that relate to child care services, policies, and workforces as well as experiences from children themselves. They do not specify a file format or data type required, but I assume it would match the ICPSR if the bibliographies are related.

Research Connections does not provide any information into what should be submitted, but the ICPSR has 3 general requirements for submission. The first requirement is to provide the data file(s) in an appropriate file format. They also provide certain criteria that the data file must meet, like assigning specific, mutually exclusive codes to each variable if the data is quantitative. The second requirement is to submit documentation files that sufficiently describe the dataset which can be submitted as an ASCII or DDI XML file or a word doc. The last requirement is to provide a study-level description that contains summary, methodology, and sample information (*Start Deposit*, n.d.).

For the ICPSR, they state that if a submitter has further questions that cannot be answered in the ICPSR FAQ page (*Frequently Asked Questions*, n.d.), then they can reach out to user support by email (icpsr-help@umich.edu). Guides are also provided during the submission process, like how to handle data confidentiality (*Data Confidentiality*, n.d.).

ICPSR does not specify a particular schema or structure but they do have an FAQ page that says the organization uses a deposit form to collect the study-level information they need (*Study-level metadata*, n.d.). Some of the fields they ask for include principal investigator, collection notes, geographic coverage, and sampling. What could be happening is that the ICPSR is taking the entries for those fields and putting them into a metadata file that follows an institutional structure.  
**Data Access Mechanisms**

Most of the resources in the CCEE repository are openly available but some texts may be blocked by a paywall. In the ICPSR, there are two types of access protocols: public use and restricted use. When downloading public use data, the user must agree to a terms of use contract that outlines principles set forth by the CoreTrustSeal that the ICPSR follows. If the data is restricted use, the user has to apply for access by filling out a Restricted Data Use Agreement.

Both the CCEE repository and ICPSR have interfaces where users can submit queries and download various files. The ICPSR has a more intricate interface and they offer more options on how to download data, such as accessing the documentation, SAS, SPSS, and R formats. The CCEE repository displays some basic metadata about a resource but it's mainly limited to something used in a citation, like creator, date, and funders. The ICPSR displays much more metadata like the methodology used, the scope of the project and subject terms. Additionally, they provide an option to export a resource’s metadata file in multiple standard formats like Dublin Core, DDI, and DATS.

From what I found, neither repository makes an explicit mention of a DIP, but based on the repository interfaces, the package includes access to the dataset and for the ICPSR, a detailed look at the dataset’s metadata along with supporting documentation.

**Additional Information**

**Recommended Data Citation**

Adhering to the data citation recommendation from DataCite [Creator (Publication Year). Title. Publisher, Identifier], the citation for this dataset would be the following:

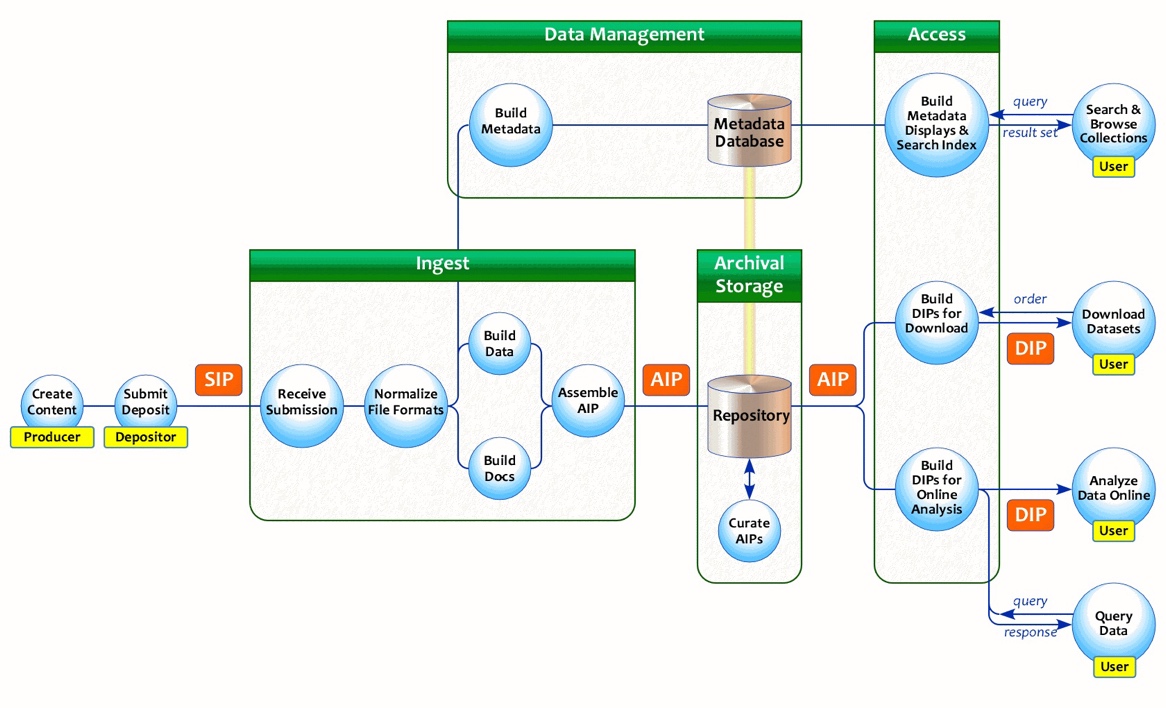
Das, S (2022). Resource mapping data by county. Tennessee State Government, <https://data.tn.gov/dataset/resource-mapping-data-by-county/resource/8d375f1d-c084-4e3e-8370-810f6893c777>

It should be noted that the persistent identifier in this case is the URL for the dataset as that was the only identifier provided by the TN data portal. Ideally, it would be better to have a DOI as the identifier and if this dataset were submitted to CCEE/ICPSR, a DOI would be generated for it.

**Long-term Preservation**

The dataset is downloadable as a CSV file which can be opened with spreadsheet software like Excel, Numbers (Mac), and Google Sheets. The CSV file format is not proprietary and while there is no immediate risk to CSV files becoming obsolete, there is always a possibility that a better format or technology will replace it. If this dataset is deposited into the CCEE/ICPSR and CSV formats were no longer supported, they would need to make a decision on whether to convert the dataset to a supported format or to use specific software that can access the original format.

The ICPSR also has a [digital preservation policy](https://www.icpsr.umich.edu/web/pages/datamanagement/preservation/policies/dpp-framework.html) that provides more detail about how data is preserved. In it, they state that they follow the OAIS model and the figure below depicts the ICPSR flowchart (*OAIS-based processes*, n.d.).



**Copyright License**

As mentioned in the Data and Metadata Profile, this dataset uses the Creative Commons licensing which makes the data publicly available and allows researchers to freely use or build on the dataset.

**Human Subject Considerations**

No human subjects were part of this dataset.

References

*Child care and early education research connections*. The Administration for Children and Families. (n.d.). Retrieved February 9, 2023, from [https://www.acf.hhs.gov/opre/project/child-care-and-early-education-research-connection s-2013-2019](https://www.acf.hhs.gov/opre/project/child-care-and-early-education-research-connection%20s-2013-2019)

CKAN. (n.d.). *Data Preview and visualization*. Data preview and visualization - CKAN 2.11.0a0 documentation. Retrieved January 29, 2023, from <https://docs.ckan.org/en/latest/maintaining/data-viewer.html#data-explorer>

*Creative Commons License Deed*. Creative Commons Attribution 4.0 International - CC BY 4.0. (n.d.). Retrieved January 29, 2023, from https://creativecommons.org/licenses/by/4.0/

*Data confidentiality*. ICPSR. (n.d.). Retrieved February 9, 2023, from <https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/>

*Datacite Metadata schema*. DataCite Schema. (n.d.). Retrieved January 29, 2023, from https://schema.datacite.org/

*Frequently asked questions for depositing data*. ICPSR. (n.d.). Retrieved February 9, 2023, from https://www.icpsr.umich.edu/web/pages/deposit/faqs.html

*Kids count*. Tennessee State Government - TN.gov. (n.d.). Retrieved January 29, 2023, from https://www.tn.gov/tccy/programs0/kc.html

*OAIS-based processes*. ICPSR. (n.d.). Retrieved February 23, 2023, from https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/oais.html

*Resources With Data From All States | Research Connections*. Child Care and Early Education Research Connections. (n.d.). Retrieved February 9, 2023, from https://www.researchconnections.org/research-library/resources-with-data-from-all-states

*Start deposit*. ICPSR. (n.d.). Retrieved February 9, 2023, from <https://www.icpsr.umich.edu/web/pages/deposit/index.html>

Tennessee Commission on Children and Youth. (n.d.). *Resource mapping reports*. Resource Mapping Reports. Retrieved January 29, 2023, from https://www.tn.gov/content/tn/tccy/programs0/resource-mapping-overview/map-reports.html

Tennessee State Government, & Tennessee Commission on Children and Youth. (2022, April). *Resource Map of Expenditures for Tennessee Children and Youth*. Retrieved January 29, 2023, from https://www.tn.gov/content/dam/tn/tccy/documents/rm/ResourceMapping\_2022.pdf

*What study-level metadata should I provide with my deposit?* ICPSR. (n.d.). Retrieved February 9, 2023, from https://www.icpsr.umich.edu/web/ICPSR/cms/3943