



Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D8.1 Urban Health Data - Guidelines and Recommendations

Work Package	WP08 – Urban Health
--------------	---------------------

Lead Author (Org)	Ana Ortigoza (Drexel University)
Contributing Author(s) (Org)	Moore K (Drexel University), Lazo- Elizondo M (Drexel University), Quitsberg A (Drexel university), Li R (Drexel University)
Due Date	31.05.2023
Date	30.05.2023
Version	1 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.7887523

Dissemination Level

<input checked="" type="checkbox"/>	PU: Public
<input type="checkbox"/>	PP: Restricted to other programme participants (including the Commission)
<input type="checkbox"/>	RE: Restricted to a group specified by the consortium (including the Commission)
<input type="checkbox"/>	CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	11.05.2023	Ortigoza Ana	Draft for internal review
0.2	27.05.2023	Ortigoza Ana	Review comments resolved
0.3	27.05.2023	Simon Hodson, Laura Molloy	Proofread and copy edit
1.0	30.05.2023	Laura Molloy	Final for upload

Disclaimer

WorldFAIR has received funding from the European Commission’s WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

Abbreviations and Acronyms

CARE	Collective benefit, Authority to control, Responsibility, and Ethics
CORDIS	Community Research and Development Information Service
DDI	Data Documentation Initiative
DMC	Data and Methods Core
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
FIP	FAIR Implementation Profile
HEIs	Higher Education Institutions
NHS	National Health Surveys
SALURBAL	Salud Urbana en America Latina (Urban Health in Latin America)

Executive Summary

This report provides a summary of actions and findings of the Urban Health Mapping and Assessment (Task 8.1) for WorldFAIR Work Package 08. Firstly, we assessed the implementation of FAIR principles within the Urban Health field, through two case studies: 1) the elaboration of a web data platform for the SALURBAL project (Urban Health in Latin America); and 2) the elaboration of a FAIR Implementation Profile (FIP) for the Urban Health discipline in general. Secondly, we focused on the data collection and harmonisation process of health survey data that was led by the SALURBAL team and allowed the elaboration of consensus on terminologies and procedures that facilitates the use of survey health data in cities for research and action.

The FAIR Implementation Profile for the SALURBAL case study contributed to the renovation process that the data system was undergoing, offering valuable guidance on good practices currently possible for making data FAIR. The elaboration and documentation of standard procedures used in data and metadata identification for the SALURBAL web platform not only contributed to their findability ('F') but also the access and reuse of the data ('A', 'R'). For the Urban Health discipline, the FAIR Implementation Profile shed light on the lack of a common repository for urban health data, and showed that most urban health data should be encoded following DDI standards. It also showed the inconsistent process of identifying data and metadata used in urban health. Lessons learned during the process support recommendations towards 1. the promotion of a deeper understanding of the FAIR principles among urban health researchers and practitioners; and 2. the systematisation of a data management plan during the design and initial steps of a research project that can guide the implementation of FAIR principles across different domains and working groups. The results of the SALURBAL (WorldFAIR WP08) FIP were used to create a web resource (A FAIR Primer) which contextualises the activity, provides information about the FAIR principles, relays some of the FIP's findings and provides guidance for making SALURBAL data more FAIR. This resource is included here as Appendix A.

Regarding health survey data, we found challenges in the harmonisation process that may be difficult for the use and interoperability of data between countries and within countries across time such as 1. disagreement in the definition of risk factors; 2. lack of consistency in categories or measurement units used for an indicator; 3. discrepancy in scales and questionnaires used for retrieving information about similar health behaviours or health outcomes. We leveraged the SALURBAL experience during this harmonisation process to propose a guideline for future harmonisation in health survey data in which the main recommendations are focused on the need to 1. generate consensus on definition and measurements in health data; and 2. revisit questionnaires and scales commonly used for some health behaviours and establish commitment on common uses.

The development of these deliverables for Task 8.1 made visible the gaps and needs of FAIR implementation in the urban health research community. Consequently, we will design and develop dissemination and training materials that can support and guide research and practitioners in urban health, which is part of Task 8.2, planned for the second year of the project.

Table of contents

Executive Summary	4
1. Introduction	7
1.1. The complexity and relevance of urban health data	7
1.2. The challenges of FAIR and CARE principles in Urban Health Data	7
1.3. The SALURBAL project	9
2. Scope of WorldFAIR Work Package 8	10
2.1. Task 8.1	10
2.2. Deliverable 8.1: Urban Health Data - Guidelines and Recommendations	10
3. Assessment of FAIR principles in Urban Health	11
3.1. SALURBAL data platform	11
3.2. SALURBAL FAIR Implementation Profile	13
4. Health survey harmonisation	14
4.1. Sources of data for health survey harmonisation	16
4.2. Strategies used for harmonisation	18
4.3. Harmonisation results	20
4.4. Challenges for survey harmonisation	22
5. Conclusions for Deliverable 8.1	23
5.1. Overall conclusions	23
5.2. Recommendations for FAIR implementation in Urban Health Data	24
5.3. Recommendations for Health Survey Harmonization in Urban Health	25
6. Next steps	25
7. References	26
8. Appendix A: FAIR Implementation Profile and FAIR Primer for the SALURBAL data platform	28
8.1. SALURBAL - FAIR Principles	28
8.2. Findable	33
8.3. Accessible	55
8.4. Interoperable	56
8.5. Reusable	57
9. Appendix B: Summary of guidelines and recommendations for FAIR practices in health survey data harmonisation	1

1. Introduction

1.1. The complexity and relevance of urban health data

Cities are considered the primary contributors to global environmental change and human development, being at the centre of leading mitigation and adaptation strategies that could promote human health along with environmental sustainability (Valhov, 2007). Urban health is considered a domain of multiple disciplines involved in understanding how the urban environment contributes to shaping population health, and the social inequalities related to health disparities among city dwellers. Urban health requires the interconnection of knowledges and practices from different disciplines to depict the complexity of urban systems.

Many stakeholders from different disciplines and levels of governance are involved in the creation, maintenance and dissemination of data used for urban health research and practice. Knowledge sharing, capacity building, and trust-building have been identified as three critical values to foster effective collaboration across and beyond disciplines and to pursue valuable solutions for complex problems in cities (Pineo, 2020). All these capacities are closely connected to the FAIR (findable, accessible, interoperable, and reusable) and CARE (Collective benefit, Authority to control, Responsibility, and Ethics) data principles that have yet to be comprehensively implemented in urban health. The better understanding of FAIR and CARE data principles can then foster and improve the above mentioned capacities towards Open Science (Wilkinson 2016; Russo Carroll 2021).

1.2. The challenges of FAIR and CARE principles in Urban Health Data

Understanding the processes involved in the generation of data (often referred to as the data lifecycle) as well as the different technical terminologies or operational languages used in data, are among the capacities necessary to develop within and beyond urban health for the correct interpretation and interaction with data in research and their translation into evidence for scientific

findings and action. These capacities are also an essential part of the creation of accountability and in bridging the diversity of backgrounds, knowledge, and experience between actors from different sectors in order to promote trust-building among them (Pineo, 2020).

Some of the challenges related to knowledge sharing are:

- *Data acquisition and use of data from different sources that are not always publicly available.* Although public health or environmental data is usually collected by official organisations, in many cases this information is not accessible to users without previous request and authorisation, involving long and time-consuming processes. In other cases, data on transportation, car use, or other consumer-related data in cities are collected and owned by private corporations, which limits their open and free access and use.
- *Availability of data at spatial scales and levels of aggregation necessary for studying problems in cities.* Many health survey data are sampled and collected following jurisdictional boundaries that sometimes are neither aligned with the geographic limits of cities nor possible to be disaggregated to lower levels than could allow the creation and compilation of within-city data.
- *Integration of data from different sources using different levels of spatial reference depending on the definition of urban setting used.* Sometimes health data is governed by administrative jurisdictions that are not coincident with the data that pertains to cities (either the administrative units composing cities or the geographic limits of cities).
- *Disparities in quality and completeness of data over time and across geographies.*
- *Adequate standardisation processes to make data comparable (across cities within and between countries and regions).*

It is necessary to raise awareness about these challenges within the urban health community to begin to incorporate the FAIR and CARE principles in urban health research, policy, and practice.

1.3. The SALURBAL project

The SALURBAL project (Urban Health for Latin American Cities) is a five-year project based at the Urban Health Collaborative, Drexel University and with partners throughout Latin America and in the United States. The project investigates how urban environments and urban policies impact the health of residents from almost 370 cities in 11 Latin American countries (Diez Roux, 2019). To pursue this goal, the SALURBAL project has 1) systematised a process for city definition and operationalization that integrates multiple ways in which a city can be delimited; 2) created a data structure that allowed the incorporation of data from different sources, making it shareable across several cores and disciplines; and 3) developed procedures and standards that systematically documented issues related to data access, quality, and completeness during the process of data harmonisation (Diez Roux, 2019).

Throughout the course of the project, the SALURBAL team has acquired expertise in:

- identifying and defining the geographical units and subunits that constitute the universe of cities included in the study;
- collecting, processing, and harmonising health data (from national vital statistics registries and national health surveys);
- creating and updating data on the physical and social environment in cities;
- integrating all available information within a multilevel data structure that allows definition and measurement of constructs at different geographical levels and over time.

This expertise involves practical examples that can support the implementation of the FAIR and CARE principles, particularly in relation to data and metadata identification, collection, processing, and management, and contribute to promoting best practices in data sharing and use in urban health. The project can provide valuable lessons in the process of creating similar data platforms for urban health research and practice, particularly in highly urban or rapidly urbanising regions of the Global South, where integration of data is still an incipient process (Gatzweiler, 2021).

Moreover, lessons learned from the SALURBAL experience on data compilation, processing, and integration could be translated to urban data systems beyond research and academia, supporting the establishment of guidelines and recommendations in the development and implementation of information systems oriented to climate crisis and further epidemic preparedness in urban settings (Liu, 2021).

2. Scope of WorldFAIR Work Package 8

WorldFAIR Work Package 8 aims to provide expertise gained on city-definition and data integration to support the implementation of FAIR and CARE data principles through the elaboration of consensus on terminologies, procedures, and good practices recommendations that facilitates the use of health data in cities for research and action.

2.1. Task 8.1

Task 8.1 is to undertake an 'Urban Health Mapping and Assessment'. This task supports the implementation of FAIR principles within the urban health field, through the elaboration of FAIR Implementation Profiles for the SALURBAL project and for the urban health discipline in general. We particularly focused on the elaboration of consensus on terminologies and procedures that facilitates the use of survey health data in cities for research and action.

2.2. Deliverable 8.1: Urban Health Data - Guidelines and Recommendations

As the deliverable from Task 8.1 we propose the documentation of cross-domain challenges in the harmonisation of National Health Surveys (NHS) across Latin American countries that led to the establishment of a set of guidelines and recommendations for FAIR practices in data harmonisation, use and sharing. We consider that this deliverable can guide key stakeholders carrying out National Health Surveys in future data collection efforts and may facilitate the use of survey health data for urban health research and practice in Latin American countries.

As part of Task 8.1 we also documented the experience of implementing FAIR data principles in the development of a data platform that will be used as a data repository and tool for urban health research in Latin American cities.

3. Assessment of FAIR principles in Urban Health

3.1. SALURBAL data platform

The motivation for the data platform¹ was to make the SALURBAL data, metadata, and project outputs reusable by other researchers in the field and the public as part of a final product to be delivered at the culmination of the project.

Throughout the SALURBAL project, the process of data collection and harmonisation within the different data working groups (urban built and natural environment, social environment, health data) led to successful results in terms of data compilation and internal use, although there was no explicit structure established across working groups in terms of data format, variable naming conventions, codebook format, codebook content, file structures. Therefore, before sharing SALURBAL data to be reused, the existing SALURBAL infrastructure went under a process of ‘renovation’ in order to make it findable and accessible by machines. To address this, we first designed a project-specific informatics standard (variable naming, data format, metadata format) and then compiled SALURBAL data into a single database².

Finally we created a user-friendly web interface to access/download SALURBAL data and metadata. The process of designing and developing the interface involved various skills and much expertise, from data scientists and engineers to user experience researchers. We also brought together people with expertise in communication and design to leverage their visual editing skills and people with

¹ SALURBAL data platform available at <https://data.lacurbanhealth.org/>

² SALURBAL restructuring process step by step available at <https://drexel-uhc.github.io/salurbal-project-dashboard/>

backgrounds in policy and community engagement to contribute with their knowledge of the audience in the development of an interface that could meet the needs of potential users besides academic researchers. The data platform has recently been launched for its initial use while other areas are still in progress (SALURBAL FAIR Implementation Profile, 2022). This platform presents a section where users can explore the data and metadata that SALURBAL compiled and made available; a list of the metadata available for each variable can be found here.³ Users can select from this section the data that they are interested in and compile it into a ‘cart’. Note that while the entirety of SALURBAL’s metadata is available on the web platform, we control which variables/data are available for download. Completely open access data will be available to any user—even those that are not logged in—but to adhere to the legal restrictions of our data sources, some data will be locked behind an authentication and authorization mechanism. For example, some data is limited to SALURBAL Data Methods Resource Core Members; they will need to login and only then can they add these to the cart. Another limitation, is that while SALURBAL data exists at multiple geographic level (city, neighbourhoods and small areas), small area estimates are not currently supported on the web platform due to methodological limitations: we do not want to make these available to anyone as they may require more complex statistical methodology to account for estimation uncertainty. Once data are selected, the users can go to the cart and send the request for dataset creation. For a typical data request—for example 10 variables that cover city level data for 11 countries for a span of 20 years—the portal would take less than a minute to process and send to the user via email. The user will receive an email with the data in .csv format as well as a codebook in .csv format. The data is organised and described so as to be easily accessible by most statistical software. The metadata is formatted as a table containing column descriptions for each column in the data table. Another part of the platform that is still in progress will allow users to create live

³ Drexel UHC GitHub, FAIR Renovation Codebook https://github.com/Drexel-UHC/salurbal-fair-renovations/blob/main/documents/templates/_codebook_fields.csv

charts, figures, and tables using the SALURBAL data embedded in the platform (SALURBAL data web platform, 2023).

In conclusion, the SALURBAL web platform was designed to increase the exposure of our data and importantly to be flexible enough to accommodate the data driven interactive web content we plan to develop in order to engage policy and community audiences. Once the renovation of the data infrastructure into a machine-actionable/FAIR-state has been completed, we plan on submitting our data and metadata in DDI format to ICPSR to ensure long-term stewardship and persistence.

3.2. SALURBAL FAIR Implementation Profile

The elaboration of FAIR Implementation Profiles (FIPs) based on the SALURBAL platform experience and on the overall urban health discipline helped us to visualise the current gaps and needs relating to the FAIR data principles for urban health research.

For the urban health discipline, the FIP shed light on the lack of a common repository for urban health data. Although a great amount of data used in urban health research can be also found in repositories traditionally used in social research such as the one developed by the Inter-university Consortium for Political and Social Research (ICPSR), most of the data regarding the built and natural environment in cities is spread across different platforms using bespoke codebooks. This poses a challenge in terms of interoperability and reusability. To address this, we recommend that most urban health data should be encoded following DDI standards.

The FIP also showed the inconsistent process of identifying data and metadata used in urban health. For example, we found that most digital objects related to spatial visualisations of health data lack systematic identifier processes. For example, sometimes maps or spatial images related to cities were not considered digital objects and did not have DOIs, or other widely used system-independent PID, assigned which inhibits Findability, unambiguous identification, consistent citation and falls short of the most basic FAIR principle.

For the SALURBAL project, the creation of a data platform involved revisiting the FAIR principles and looking for FAIR Enabling Resources (FERs) within the urban health discipline that helped restructure the data and metadata records for the project, with a standardisation of identifiers (variable and dataset names), and datasets and data protocol formats that consequently enabled the reorganisation of the data structure to increase and improve findability, interoperability, and reuse of the data through machine-actionable mechanisms. In summary, the elaboration of a FIP for the SALURBAL data platform contributed to support the renovation process that the data system was undergoing, offering valuable guidance on the best practices possible in making FAIR data. The results of the SALURBAL (WorldFAIR WP08) FIP were used to create a web resource (A FAIR Primer) which contextualises the activity, provides information about the FAIR principles, relays some of the FIP's findings and provides guidance for making SALURBAL data more FAIR. This resource is included here as Appendix A.

4. Health survey harmonisation

To allow for the investigation of questions on urban health in Latin American cities, SALURBAL collected, integrated, and linked data across domains (health, built and natural environment, social environment) and levels (countries, cities, sub-cities, and “neighbourhoods”) (Figure 1, SALURBAL Data Brief 1 - 2018). Within the SALURBAL project, the Data and Methods Core (DMC) provides overall data and analytical support to the project. This core is responsible for rigorously protecting data privacy, which is especially complex due to the geographically specific nature of the data SALURBAL works with (Diez Roux, 2019).

The SALURBAL DMC prioritised the use of existing data resources, which offered great cost-efficiency but at the same time required large retrospective harmonisation efforts to ensure equivalence in the measures of variables to minimise bias and noise in the data. The SALURBAL DMC has compiled and harmonised data on key health and health-related domains: individual socio-demographic characteristics, and health survey data (including adults and children). (Moore, work

in progress)⁴. As a result of the harmonisation process of health surveys, it was possible to link SALURBAL health data with data measures from the built, social, and natural environment at different levels: Level 1 (L1) measures representing city features, Level 2 (L2) measures representing sub-city (smaller administrative areas composing the cities), and Level 3 (L3) measures representing the neighbourhoods (smallest units for which data is available) (Figure 1).

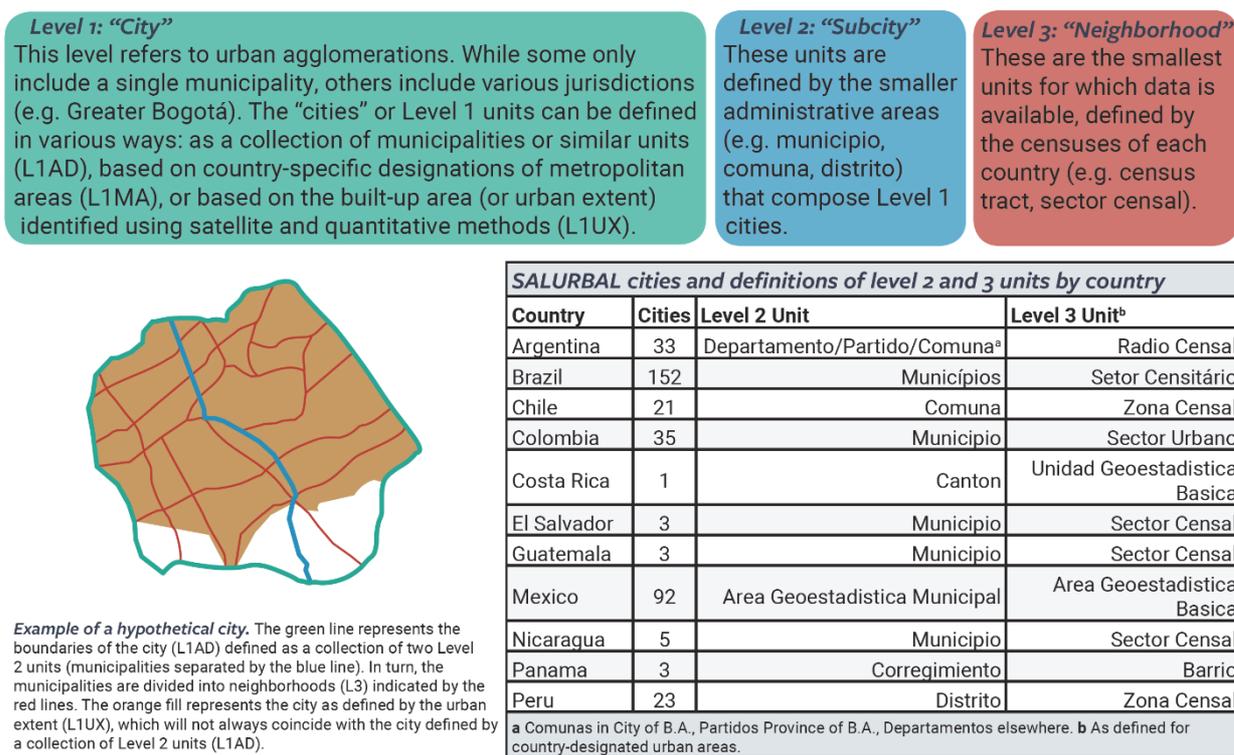


Figure 1. Definition of cities using different administrative levels (SALURBAL Data Brief 1, 2018)

Survey data were gathered from National Bureaux of Statistics or other relevant government ministries or agencies responsible for the survey use. Data from health surveys is de-identified data, and is publicly available. National Bureaux of Statistics systematically collect these data and release

⁴ Harmonised datasets and its metadata available upon request at <https://data.lacurbanhealth.org/>

them in public platforms once the data are revised and consolidated, along with metadata documents (manual of procedures for the questionnaire interview, methodological protocol, codebooks).

The initial focus of SALURBAL was on non-communicable disease risk factors (e.g., anthropometry, diabetes, hypertension, tobacco use, alcohol use, diet, physical activity). Other key domains included sociodemographic (age, sex, educational attainment, and other SES measures), mental health, and health care access. Additional domains have been added in response to interest from co-investigators (e.g., experiences of violence). Survey questionnaires are usually released to one member of the household who answers about socioeconomic and demographic characteristics of the whole group living in the household. Regarding anthropometric measures, some surveys are based only on self-reported health information, while others include objective measurements of height, weight, and blood pressure. For the harmonisation of variables in relation to which there were different data collection methods (e.g. self-reported vs objective measures) we kept track of the origin of each data and ran a quality assessment to ensure consistency in response patterns across surveys and over time.

4.1. Sources of data for health survey harmonisation

The SALURBAL health survey data includes 27 surveys from 11 countries (Table 1). Survey years ranged from 2000 to 2018, and the number of surveys per country was 1-5. The total number of adults (aged 18-100) and children (aged 0-17) included in the surveys was 246,738 and 69,126, respectively. In 5 countries there are multiple repeat surveys over time. Data available include individual-level demographic and socio-economic characteristics, alcohol/tobacco use, anthropometry, diet/physical activity, diabetes, hypertension, mental health, self-reported health, for adults. For children, the dataset includes household-level demographic and socio-economic characteristics, and anthropometry.

Table 1: Summary of health surveys included in SALURBAL data resource. By country name (alphabetical order) and survey year.

Country	Survey name	Age	Child data	Survey year	SALURBAL Sample size
Argentina	Encuesta Nacional de Factores de Riesgo, ENFR (National Risk Factors Survey)	≥18 years	No	2005	25,753
				2009	16,218
				2013	21,451
Brasil	Pesquisa Nacional de Saúde, PNS (National Health Survey)	≥18 years	No	2013	29,353 in L2s 40,703 in L1ADs
				2019	33,515 in L2s 46,767 in L1ADs
Chile	Encuesta Nacional de Salud, ENS (National Health Survey)	≥15 years	No	2003	2,032
				2010	3,140
				2017	3,805
	Encuesta Longitudinal de Primera Infancia (ELPI) (Longitudinal Survey of Early Childhood)	1-12 years	Yes	2017-2018	6,723
Colombia	Encuesta Nacional de Salud, ENS (National Health Survey)	18-69 years	No	2007	43,182
	Encuesta Nacional de la Situación Nutricional en Colombia, ENSIN (National Nutritional Situation in Colombia)	0-69 years	Yes	2005	42,336 adults (18-69) 23,794 children (<18)
				2010	55,863 adults (18-69) 30,278 children (<18)
				2015	36,593 adults (18-69) 17,104 children (<18)

Costa Rica	Encuesta Multinacional de Diabetes mellitus y Factores de Riesgo, CAMDI (Multinational Survey of Diabetes Mellitus & Risk Factors, Central American Diabetes Initiative)	≥20 years	No	2005	1,427
Guatemala	CAMDI (See Costa Rica)	≥20 years	No	2002-2003	1,397
	Demographic and Health Survey (DHS)	Females 18-49 years Children <5 years	Yes	2014-2015	2,730 adult females (18-49) 983 children (<5)
Nicaragua	CAMDI (See Costa Rica)	≥20 years	No	2003	1,397
Mexico	Encuesta Nacional de Salud, ENSA (National Health Survey)	Adults ≥18 Children <5 years	Yes	2000	29,733 adults (≥18) 23,966 children (<5)
	Encuesta Nacional de Salud y Nutricion, ENSANUT (National Survey for Health and Nutrition)	All ages	Yes	2006	31,532 adults (≥18) 19,431 children (<18)
				2012	26,335 adults (≥18) 25,014 children (<18)
				2016	14,618 adults (≥18) 3,274 children (<18)
2018	27,118 adults (≥18) 7,538 children (<18)				
Panama	Encuesta Nacional de Salud y Calidad de Vida ENSCAVI (National Survey of Health and Quality of Life)	≥18 years	No	2007	11,394
Peru	Encuesta Nacional de Demografia y Salud, ENDES (National Survey of Demographics and Health)	Adults ≥15 years Children <5 years	Yes	2016	11,929 adults (≥18) 8,547 children (<5)
El Salvador	CAMDI (see Costa Rica)	≥20 years	No	2004	1,872
	Encuesta Nacional de Salud Familiar (National Family Health Survey)	Females 18-49 years Children <5 years	Yes	2008	4,297 adult females (18-49) 1,290 children (<5)
	Encuesta Nacional de Enfermedades Cronicas no transmisibles en Poblacion Adulta de El Salvador ENECA (National Survey of Non-communicable Chronic Diseases in the Adult Population of El Salvador)	≥20 years	No	2014-2015	1,546

4.2. Strategies used for harmonisation

The health survey data resource production was led by SALURBAL DMC with the support of SALURBAL members from country teams. For the creation of this resource, SALURBAL followed four guiding principles:

1. Use existing national health survey data administered by agencies within each country.
2. Use only surveys that could be linked to country geographical administrative IDs, corresponding to SALURBAL sub-city level (L2) (e.g., availability of geographic identifiers in survey datasets, publicly or through request to the agency implementing the survey).
3. Prioritise surveys with information on non-communicable health behaviours and risk factors.
4. Use harmonisation approaches that are rigorous but flexible to accommodate differences across surveys.

We developed a flexible retrospective harmonisation process that included:

1. Identifying and collating survey questions and responses by domain, with attention to response patterns in the questionnaires asked on the survey and respondent universe.
2. Reviewing surveys conducted by other institutions such as the Centers for Disease Control and Prevention or the World Health Organization for standard variable definitions as well as harmonisation approaches proposed by other projects.
3. Proposing harmonised variable definitions and response categories with attention to differences in wording across countries.
4. Applying the harmonisation and revising the protocol as needed, based on descriptive statistics of initial harmonised variables.

Harmonisation approaches included:

1. Creation of multiple versions due to country differences that did not allow a single harmonised variable (e.g., diabetes, hypertension).

2. Unit conversion (e.g., height, weight).
3. Collapsing categories (e.g., education, self-rated health).

4.3. Harmonisation results

The total number for survey variables harmonised to date include 109 for adults and 29 for children. Table 2 summarises the data availability in National Health Surveys (NHS) after the harmonisation process.

For NHS with information from the adult population, the variables that were more feasible to harmonise across countries were the ones related to physical activity and depressive symptoms (14 and 23 harmonised variables, respectively). This is mostly because these domains present standardised questionnaires that are replicated almost identically in all the countries. Conversely, questions soliciting self-reported health outcomes, such as diabetes diagnosis or access to health care, are less likely to be similar across countries, and therefore the number of variables that could be harmonised were very low (Table 2).

For NHS with information from the child population, the number of domains available for comparison across countries was much lower than for adults. For children, the only health-related domain that was possible to harmonise was that of anthropometric measures, for which the harmonisation process mostly consisted of the conversion of different measurement units used for weight and height.

Overall, we observe that some questions related to diagnosis of non-communicable diseases (e.g., diabetes or hypertension) as well as self-reported health, access to health care, and perception of violence are poorly developed in survey questionnaires. However, as a result of the harmonisation process, the percentage of surveys with at least one variable that is comparable across different countries was high even for domains with few questions available, such as diabetes, hypertension or pregnancy diagnosis.

Table 2. Summary of harmonised survey core data available in SALURBAL by domain
For detailed characteristics of variables please see supplemental table 2.

Domain	Number of core harmonised variables*	% of surveys with all harmonised variables in the domain	% of NHS with at least one harmonised variable in the domain
Adults			
Demographics	6	64%	100%
Socioeconomic status	15	67%	100%
Alcohol	6	33%	81%
Tobacco Use	8	26%	92%
Anthropometry	6	54%	96%
Diet	9	50%	75%
Physical Activity	14	34%	62%
Diabetes	3	73%	81%
Hypertension	7	64%	81%
Depressive symptoms	23	12%	50%
Self-reported health	3	22%	65%
Health care	4	41%	46%
Pregnancy	2	46%	62%
Violence	3	13%	19%
Children			
Demographics	5	68%	100%
Socioeconomic status	16	85%	100%
Anthropometry	8	99%	100%

4.4. Challenges for survey harmonisation

During the harmonisation process we identified different challenges that made the comparison of survey data difficult not only across countries but also within the same survey over time.⁵

1) Disagreement in the definition of risk factors. For example, in some countries the question that retrieved diagnosis of diabetes (*'Has a physician ever diagnosed you with diabetes or "high sugar" in your blood?'*) included a separate question for gestational diabetes for women (*'Did this occur while you were pregnant?'*) while other countries did not have this consideration. The comparison of the data retrieved by these questions is difficult due to discrepancies in the respondent universe.

2) Lack of consistency in categories or measurement units used for an indicator. For example, many surveys included the same wording for questions about self-rated health (*'In general, would you say your health is...'*) but the categories used in the answer option differed (*"Excellent," "Very Good," "Good," "Fair," or "Poor"* vs *"Very Good," "Good," "Fair," "Poor," or "Very Poor."*) which made the harmonisation of responses difficult as some categories are not necessarily equivalent or could not evidently be combined with or collapsed into other ones.

3) Discrepancy in scales and questionnaires used for retrieving information about similar health behaviours or health outcomes. Many surveys used different sets of questionnaires for measuring levels of physical activity (International Physical Activity Questionnaire - IPAQ vs Global Physical Activity Questionnaire - GPAQ) or fruit and vegetable consumption (frequency questionnaires based on servings vs frequency questionnaires based on portions) that led to a different assessment of the outcome.

⁵ The guidelines and recommendations for FAIR practices in health survey data harmonisation that are discussed in this document will be further refined and presented in an academic article. Further detail is presented in this Appendix B.

5. Conclusions for Deliverable 8.1

5.1. Overall conclusions

The development of Deliverable 8.1 contributed to revisiting the strengths and limitations that the SALURBAL study has with respect to FAIR data implementation. The lessons learned throughout this process serve also as a valuable case study for raising awareness of the existing gaps in FAIR data among the urban health community and particularly in relation to the accessibility, interoperability, and (re-)use of key sources of data such as health surveys.

The elaboration and documentation of standard procedures used in data and metadata identification for the SALURBAL web platform not only contributed to the improvement of its findability ('F') but also of the accessibility and reusability of the data ('A', 'R').

Actions developed by SALURBAL on city-definition and data integration are practical examples of interoperability ('I'). Finally, the collection, processing, and harmonisation of health surveys contributed to the interoperability and reuse ('I', 'R') of data.

A better understanding of the limitations and challenges faced during this process has been key for envisioning practical recommendations for improving FAIR data within the SALURBAL project that are expandable (and possibly generalizable) to the urban health community.

The timeline for this process includes the following: March 2022–March 2023 focused on renovating the existing SALURBAL data resource to a machine actionable structure and development of the full-stack web portal. In March 2023, we conducted an internal release targeted at a subset of the SALURBAL community where we sought feedback. We received positive feedback for our website's user experience but a lot of negative feedback in regards to the sparsity of metadata. This was expected as the renovation process was ongoing and had not yet finished. We plan on wrapping up the FAIR renovations and hope to have our entire resource in a machine actionable format for a pre-release targeted at the entire SALURBAL group in mid-June 2023. After another round of feedback

during the summer and the addition of a policy impact section we hope to have a public launch in September of 2023.

The availability of a unique and harmonised health survey dataset allowed researchers to compare health outcomes in a wide range of cities across Latin American countries. Up to the present there more than 30 original research articles have been produced using SALURBAL data. Without these harmonised data, such a significant contribution would not have been possible. Most researchers found the data and metadata produced understandable and ready to use. Barriers regarding findability and interoperability of this data have improved after the implementation of the data platform. The SALURBAL team is in the process of developing a satisfaction survey to receive more feedback from data providers and users regarding the FAIR features of the data platform.

5.2. Recommendations for FAIR implementation in Urban Health Data

1. Promote a deeper understanding of the FAIR principles among the urban health researchers and practitioners, and data professionals. We noticed that not only among researchers but also among people working with data in urban health (such as data managers, data system personnel) the knowledge of FAIR principles is fragmented and heterogeneous as a consequence of the lack of formal training in the definition and exemplification of the FAIR principles. We consider that this step is key for setting the basis for better FAIR practices in the future.
2. Explore adoption of DDI standards. We found that most urban health data should be encoded following DDI standards. This will be done in order to submit the SALURBAL data to ICPSR.
3. Systematization of a data management plan during the design and initial steps of a research project that can guide the implementation of FAIR principles across the different domains and working groups.

5.3. Recommendations for Health Survey Harmonization in Urban Health

1. Generate consensus on definition and measurement of key indicators for health data included in surveys.
2. Review of questionnaires and scales commonly used for some health behaviours and outcomes, and create common agreement on which would be more appropriate to use in the context of urban health in Latin America.

6. Next steps

Following the identification and highlighting of the gaps and needs of FAIR implementation in the urban health community, a consequent action will be the design and development of dissemination and training materials that can support and guide research and practitioners in urban health. Further work is required to provide guidance on Interoperability and Reusability, as the SALURBAL FIP identified a number of gaps, e.g. in relation to the use of FAIR vocabularies (I2) and the provision of provenance and processing metadata (R1.2). Some of this can be addressed by engagement with the DDI standards, good practices in related communities and the WorldFAIR CDIF recommendations.

Beginning in 2023, all NIH- funded grant applications or renewals that generate scientific data must include a robust and detailed plan for how researchers will manage and share data during the entire funded period (Data Management and Sharing Plan, DMSP) (NIH, 2023). We see this context as very propitious for introducing and promoting formal training on FAIR and CARE principles in urban health. Therefore, as part of Task 8.2 we are envisioning a training course that includes FAIR and CARE definition and practical examples as well as a hand- on guidance on how to create a data management plan through the FAIR lens.

7. References

- Diez Roux A, Slesinski C et al. (2019) A Novel International Partnership for Actionable Evidence on Urban Health in Latin America: LAC-Urban Health and SALURBAL. *Global Challenges* 2019, 3, 1800013. <https://doi.org/10.1002/gch2.201800013>
- Gatzweiler F. (2021) Data for Intelligent Urban Systems / Data-Knowledge-Action Systems. [Under review]
- Liu J, Gatzweiler F et al. (2021) Co-creating FAIR data for healthy and resilient urban development. [Under review]
- Moore K, Kari Lazo – Elizondo M, Ortigoza A, Quitsberg A, Sanchez B, Acharya B, et al. Data Resource Profile: Harmonised health survey data for across 11 countries in Latin America: the SALURBAL project. [Under Review]
- NIH Data Management and Sharing Policy 2023 [Last date accessed 1/13/2023] <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>
- Pineo H, Audia C et al. (2021) Building a Methodological Foundation for Impactful Urban Planetary Health Science. *J Urban Health* 2021, 98:442–452. <https://doi.org/10.1007/s11524-020-00463-5>
- Quistberg D , Diez Roux A et al. (2019) Building a Data Platform for Cross-Country Urban Health Studies: the SALURBAL Study. *J Urban Health* 2019- 96:311–337. <https://doi.org/10.1007/s11524-018-00326-0>
- Russo Carroll S et al. (2021) Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Nature -Scientific Data*. 2021 8:108. <https://doi.org/10.1038/s41597-021-00892-0>
- SALURBAL, Data Brief 1 ‘ Data in the SALURBAL Project’ , 2018. Available at <https://drexel.edu/lac/data-evidence/data/>
- SALURBAL Data Web Platform, 2023. Available at <https://data.lacurbanhealth.org/> [Accessed on 04/29/2023]
- SALURBAL FAIR Implementation Profile, 2022. Available at <https://drexel-uhc.github.io/salurbal-project-dashboard/> [Accessed 04/29/2023]
- Vlahov, D., Freudenberg, N., Proietti, F. et al. Urban as a Determinant of Health. *J Urban Health* 84 (Suppl 1), 16–26 (2007). <https://doi.org/10.1007/s11524-007-9169-3>
- Journal of Urban Health: Bulletin of the New York Academy of Medicine*, Vol. 84, No. 1 [doi:10.1007/s11524-007-9169-3](https://doi.org/10.1007/s11524-007-9169-3)*2007 The New York Academy of Medicine Urban as a Determinant



DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION

of HealthDavid Vlahov, Nicholas Freudenberg, Fernando Proietti,Danielle Ompad, Andrew Quinn, Vijay Nandi,and Sandro Galea

Wilkinson M et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Nature- Scientific Data. 2016 -3:160018 <https://doi.org/10.1038/sdata.2016.18>



8. Appendix A: FAIR Implementation Profile and FAIR Primer for the SALURBAL data platform

The results of the SALURBAL (WorldFAIR WP08) FIP were used to create a web resource (A FAIR Primer) which contextualises the activity, provides information about the FAIR principles, relays some of the FIP's findings and provides guidance for making SALURBAL data more FAIR. Further work is required to provide guidance on Interoperability and Reusability, as the SALURBAL FIP identified a number of gaps, e.g. in relation to the use of FAIR vocabularies (I2) and the the provision of provenance and processing metadata (R1.2). Some of this can be addressed by engagement with the DDI standards, good practices in related communities and the WorldFAIR CDIF recommendations.

This material is a printable version of the site, you can find more information at <https://drexel-uhc.github.io/salurbal-project-dashboard/>

8.1. SALURBAL - FAIR Principles

A primer on FAIR principles and the SALURBAL project.

The why

The problem with the current data ecosystem is that the outputs of millions of dollars of public research funding are stored within each recipient or research group's system, without any standards for sharing their outputs or data generated. This issue is prevalent in health informatics as well where a large commercial EMR market provides a wide selection of disparate informatics platforms that don't work well with each other.

This means there is a **large pool of data that cannot be found, accessed, interoperated with each other and thus unable to be reused**. This problem is prevalent within institutions as well as on a industry/national/global level (COVID-19 is a prime example of how important and beneficial quick and reliable data sharing and integration is to epidemiological surveillance and evidence based policy decisions).

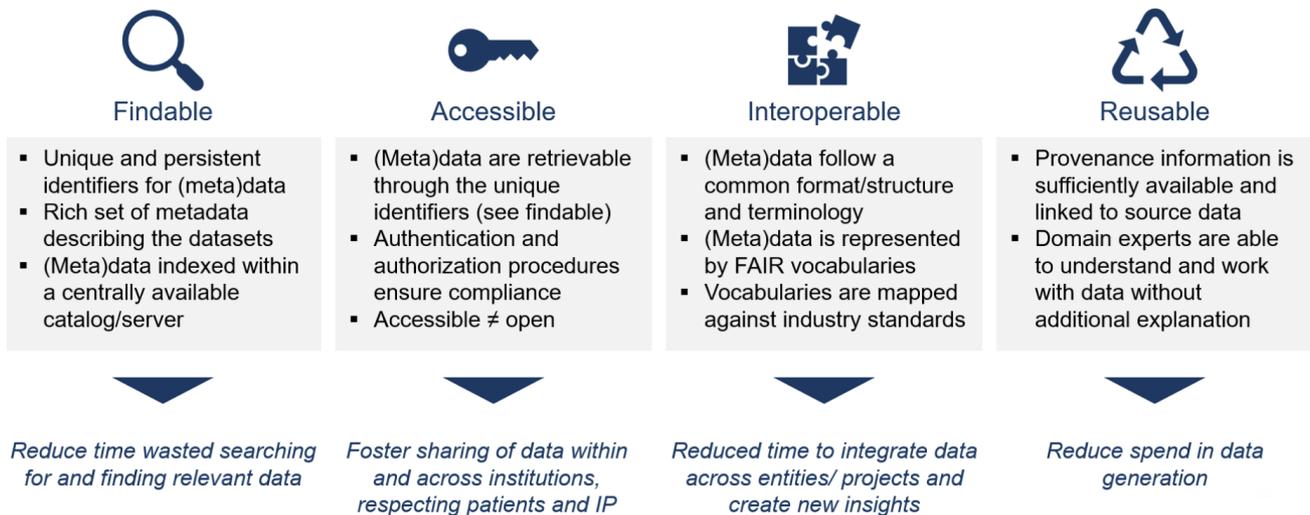
FAIR Principles

Largely to address these data challenges, a collective of pharma, academia, government and publishing representatives published the **FAIR data principles were published in 2016** ([Wilkinson et al. 2016](#)) to provide guiding principles for improved Findability, Accessibility, Interoperability, and Reuse of digital assets.

The principles emphasise **machine-actionability** (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans

increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

FAIR data principles and direct value generated by FAIR data are displayed below ([Wise et al. 2019](#)).



FAIR ≠ Open data

Open data (everyone share your data) does not mean the data is FAIR (easily reusable). The open data movement ([Murray-Rust 2008](#)) predates FAIR data but it quickly became apparent that there was not much value to just exposing our data to the world if other not could find, access, interoperate or reuse.

FAIR data (easily findable/reusable) does not equal open data. The FAIR data principles are a set of best practices for data management and stewardship; **the Accessibility principle states that FAIR projects should respect any ethical legal or contractual restrictions** ([Wilkinson et al. 2016](#)). Authentication and authorization mechanisms should be in place to prevent access to sensitive/protected data. **However, if possible, metadata or information about the data should be made available so others can find the data and go through proper channels for access.** Even if your data is not open, it is still best practice to enforce FAIR principles to not only maximise the value of your data within the project but also to provide details.

Projects should aim to make their data FAIR (regardless of whether it is open) in order to maximise the value of their data assets.

How to be FAIR

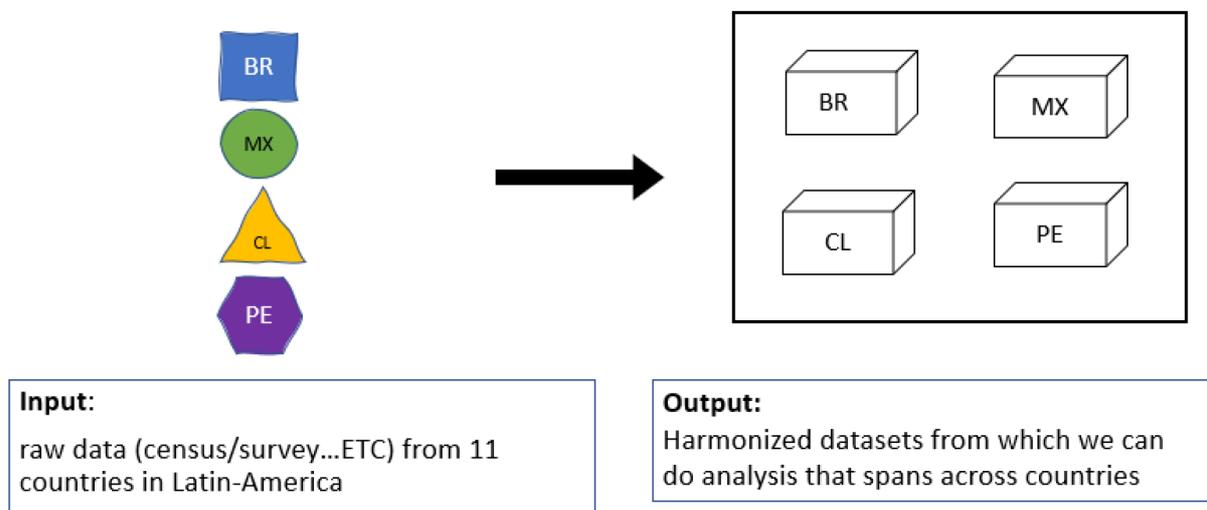
Since the introduction of the FAIR data principles in 2016, there is still **no global multi-disciplinary accepted implementation for the FAIR principles**. However there have been FAIR implementations executed at several levels from within a project, to within a field/industry, to national or even global efforts. Current effort led by the **Go FAIR initiative seeks to profile the diverse FAIR implementation ecosystem in order to document convergent implementation methods** ([Schultes et al. 2020](#)).

What does this mean for SALURBAL?

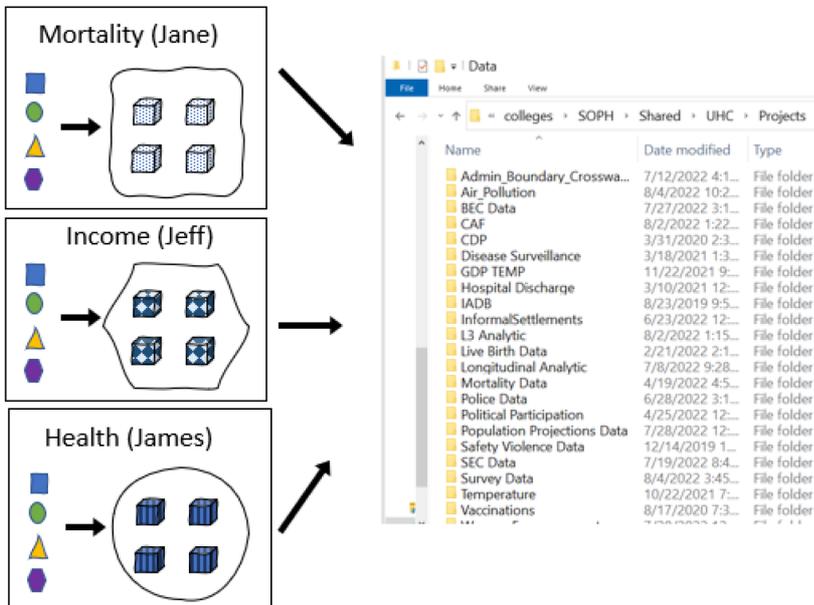
The first step is to achieve FAIR/machine-actionable data within our own project (SALURBAL). Then we can utilise appropriate FAIR data services (e.g. ICPSR) to share our resources with the larger global FAIR data ecosystem.

Here we first evaluate the baseline SALURBAL infrastructure in terms of some FAIR principles, before talking about our efforts to improve the FAIRness of our project.

SALURBAL: goal

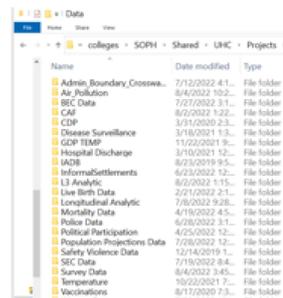


SALURBAL: what initially happened



- Groups work independently
- Lack of strict standards across groups: variable naming conventions, codebook content/format
- Then outputs were stored in separate folders

#1 Low Findability



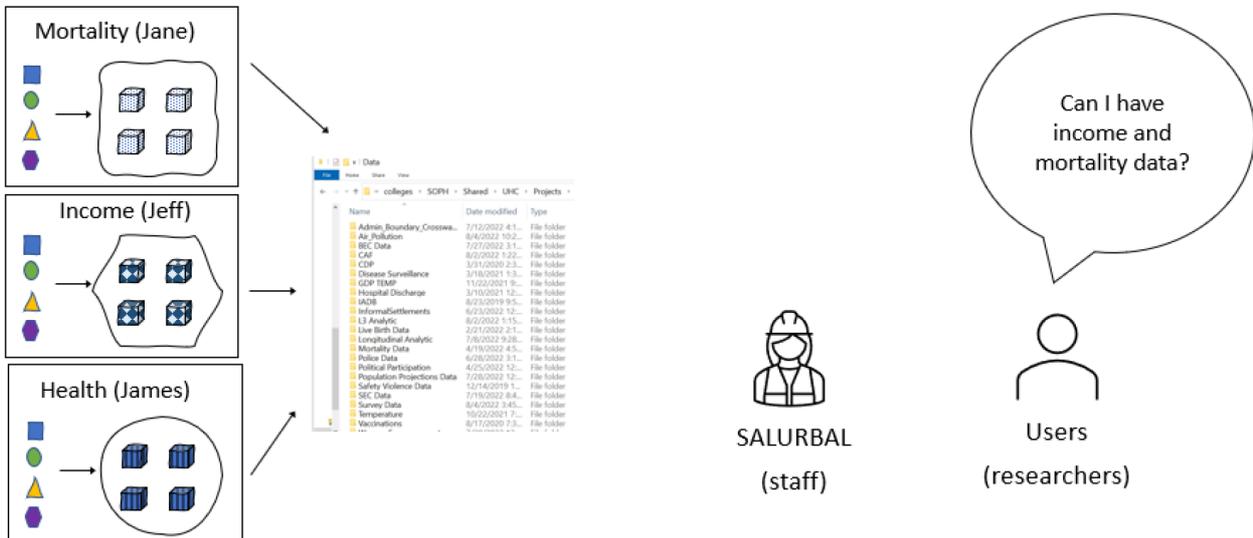
What data is available?

Do I have to click every folder individually and check every data file?

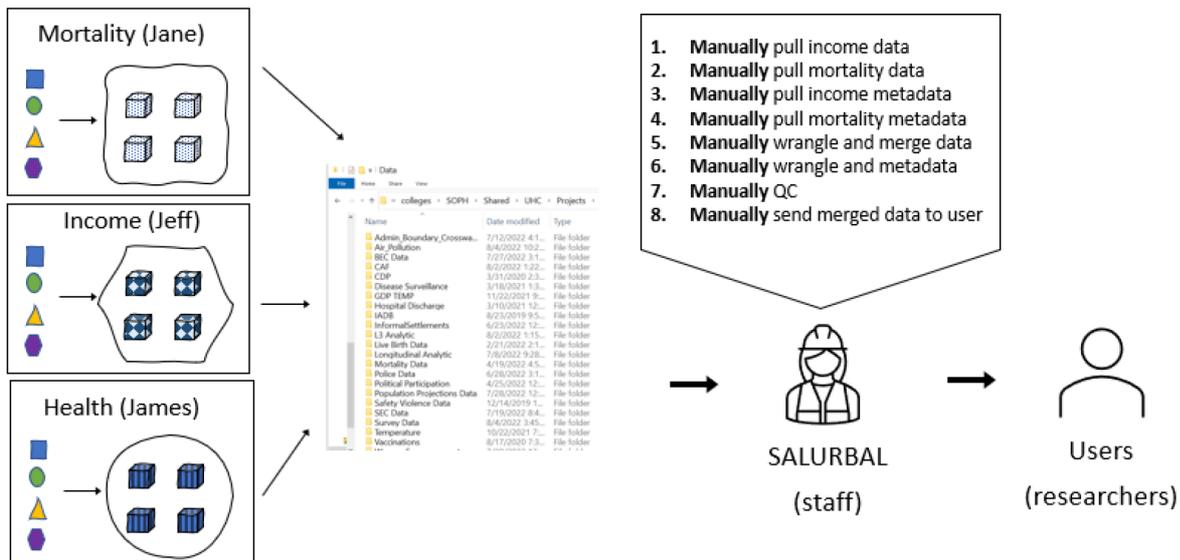


Users
(researchers)

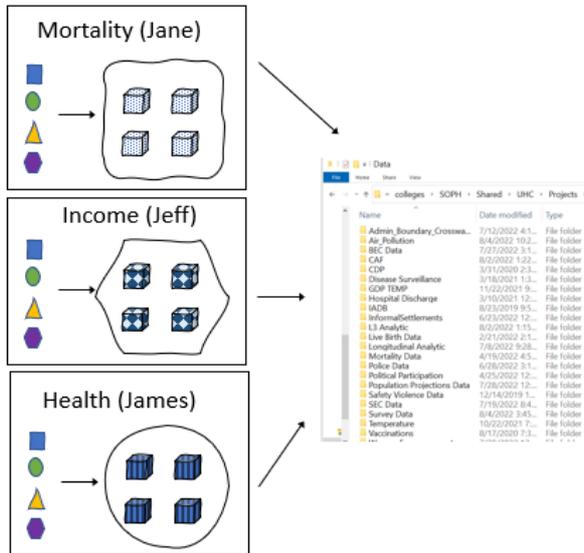
#2 Access is human intensive



#2 Access is human intensive



☹️ #3 No interoperability (big picture)



If data was in one data structure... I could:

- Automate enforcement of FAIR standards
- Conduct systematic QC checks
- Automate data distribution
- Produce parameterized reports
- ETC


SALURBAL
(staff)

8.2. Findable

Findable: Metadata and data should be findable for both humans and computers.

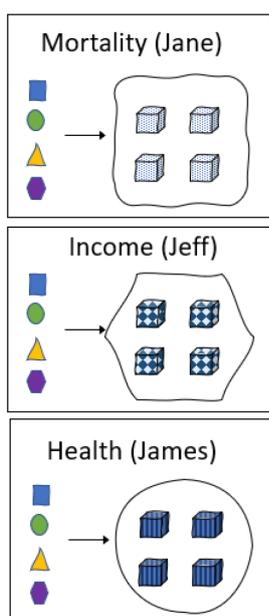
Findable is broken down into four principles

Principle	Verbatim	Layman's Terms (ELI5)
F1	Globally unique and persistent Identifiers	Variables need unique ids
F2	Data are described with rich metadata (defined by R1 below)	Comprehensive codebooks that cover community needs
F3	Metadata clearly and explicitly include the identifier of the data they describe	Codebooks can be linked to data
F4	(Meta)data are registered or indexed in a searchable resource	User interface to search through data/codebooks

Actions taken to improve SALURBAL Findability

The following sections will detail how we implemented each Findability principle F1-F4.

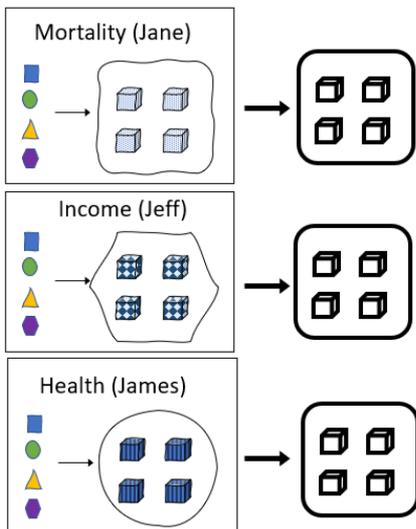
Baseline infrastructure



⚠ Findable - Challenges

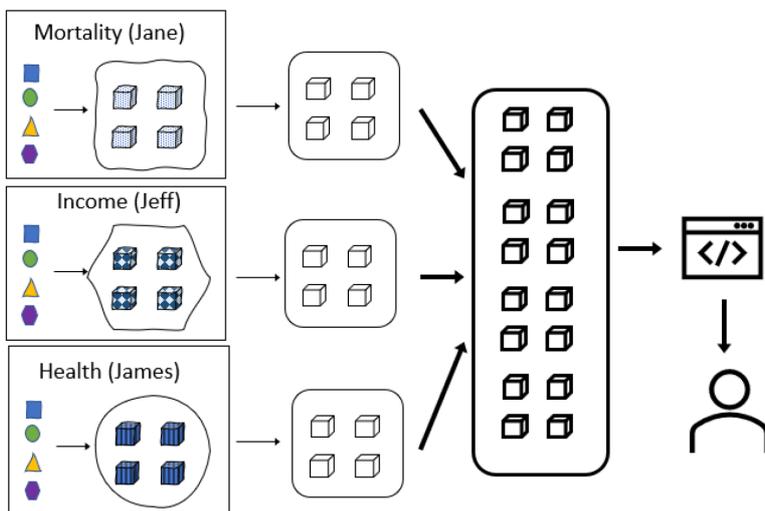
- lack of key research oriented metadata (units, interpretations, ETC)
- lack of project management metadata (variable domain/subdomain, censor status)
- metadata may be in machine unfriendly formats (pdf, word, doc)
- complex metadata that vary within variable by country, year or strata (mortality censorship by certain countries)
- No established way to link complex metadata to data

Step 1: Standardization (F1, F2, F3)



- enforce strict variable naming convention (F1.)
- create more comprehensive codebooks (F2.)
- link codebooks to datasets that capture by country, by year, by strata complexities (F3.)

Step 2: User Interface (F4)



- Compile standardized data into a single database
- Create a user interface for users to interact with our SALURBAL database (F4)

Improve Findability - action items

- (F1) enforce unique variable level identifiers
- (F2) create comprehensive codebooks which contain additional research focused and project related metadata
- (F3) link complex metadata to data via identifiers, country and year
- (F4) web application to search and access standardize data/codebooks

F1. Globally unique and persistent identifiers

ELI5: Variables need unique ids

Within project id - var_name

The SALURBAL database is a collection of data items, each item being an individual variable. So within the scope of the project, our primary concern is that **each variable has a unique identifier which we term var_name**. For example, the data item of the variable *SALURBAL Life Expectancy* is assigned a var_name of *LEAEA*. Within SALURBAL no other variable/data-item has this identifier.

var_name Rules

1. **var_name** is a string containing only letters and numbers and does not have spaces or special characters.
2. **var_name** is a variable level identifier that should not contain strata information. For example SECLABPARTM and SECLABPARTF are invalid because they indicate that the variable SECLABPART is for sex (male) and sex (female) strata; the correct var_name in this case is just SECLABPART. **var_name** is strictly for the variable and strata is captured in supplementary identifiers detailed in [F3](#)

Outside project id - DOI

DOI at collection or item level?

From a global perspective, DOIs are a common way to uniquely and persistently identify digital assets. After we have established a certain level of FAIRness we can upload our data to a FAIR data repository (such as ICPSR) and they will mint a DOI for our data collection. Then we can append our within-project identifier `var_name` to the collection/SALURBAL-level DOI to allow item level identification. For example:

- Data Asset: SALURBAL Life expectancy data
- Within project unique identifier: LEAEA
- SALURBAL project identifier (e.g. DOI): 0.1000/ICPSR/xyz123
- Globally unique and persistent Identifier: 0.1000/ICPSR/xyz123/LEAEA

F2. Data are described with rich metadata (defined by R1 below)

ELI5: Comprehensive codebooks that cover community needs.

Machines are great for computing but can't extrapolate certain things based on context. Consequently, how informative our data portal is will depend on how comprehensive and machine-actionable our data/codebooks are. Below we document standards for SALURBAL data/codebooks that improve FAIRness of our project and make a much more comprehensive amount of context accessible to the data portal.

Data

Legacy SALURBAL data tables structure were in general pretty FAIR. The only major change in the renovated data table structure is that we enforce strict rules for our within project variable identifiers. `var_name` is our workhorse identifier which links things at the variable-level - it should not contain strata information. In some cases metadata is available within variable by country details or data strata and we use `iso2` or `strata_id` to do data-metadata linkage. More details can be found on the [F3 principle page](#).

The first tab below gives [details on what fields/columns](#) should be present in renovated SALURBAL data tables and the second shows an [example data table](#). The new data columns/fields can be grouped into the following categories

- **Identifiers**: columns responsible for linkage of data and metadata. `var_name` is our workhorse identifier which links things at the variable-level - it should not contain strata information
- **Data**: data related fields including SALID, year of data, geographic level and year.
- **Internal**: internal project related metadata (intermediate strata details and file directories).



- [Data Fields \(details\)](#)
- [Data \(example\)](#)

field
description
coding
var_name
(IDENTIFIERS) SALURBAL wide unique variable identifier Character- all upper case. No spaces or special characters.
dataset_id
(IDENTIFIERS) dataset id character. Operationalized from file name. No spaces; use underdash to separate words.
iso2
(IDENTIFIERS) iso2 of the country character. Iso2.
geo
(DATA) Geographic unit of data character. Currently only L1AD.
salid1
(DATA) SALID1 character. six digit SALID1
value
(DATA) Data value character. We can parse things later.



year

(DATA) Year for which data value represents

Character four character - YYYY

Codebooks

Machines are great for computing but are quite dumb ... in other words they can't extrapolate certain things that humans can based on context. So making the web application to interface with the SALURBAL data we found that there were fundamental metadata (data about the data) - which may be possible for SALURBAL staff to extrapolate - missing (either explicitly missing from codebooks, or in machine unreadable formats).

Below is a more comprehensive codebook structure to address these gaps. The first tab below give [details on what fields/columns](#) should be present in renovated SALURBAL codebooks and the second shows an [example codebook](#). The new codebooks columns/fields can be grouped in to the following categories:

- **Identifiers:** columns responsible for linkage of data and metadata. `var_name` is our workhorse identifier which links things at the variable-level - it should not contain strata information
- **Categorization:** columns responsible for grouping variables into user friendly domains and subdomains.
- **Details:** research related variables details, this will be useful for users who want to reuse our data/codebooks.
- **Internal:** internal project related metadata (file directories and access status).

The new codebooks columns/fields can be grouped in to the following categories

- **Identifiers:** columns responsible for linkage of data and metadata
- **Categorization:** columns responsible for categorising variables into domain or subdomain
- **Details:** research related variables details, this will be useful for users who want to reuse our data/codebooks.
- **Internal:** internal project related metadata

- [Codebook fields \(details\)](#)
- [Codebook \(example\)](#)

field

description

coding

var_name

(IDENTIFIERS) SALURBAL wide unique variable identifier

Character- all upper case. No spaces or special characters.

dataset_id

(IDENTIFIERS) dataset id

character. Dataset id is operationalized by the original file name. For example, for the dataset from APSL1AD_02152022.csv the dataset id should be APS. See examples in the next tab. Formatting: No spaces; use underdash to separate words.

iso2

(IDENTIFIERS) Optional identifier only used for metadata that varies by country. iso2

Character- all lower case. For metadata that differ by country please enter a row for each country or groups of countries. You can write a single iso2 or a string of iso2 separated by commas.

strata_id

(IDENTIFIERS) Optional identifier only used for metadata that varies by strata. Stratification identifier.

Character - concatenation of strata_1_name, strata_1_value, strata_2_name, strata_2_value separated by underdashes. Not that if any strata field is empty then do not concatenate anything or underdash

year

(IDENTIFIERS) Optional identifier only used for metadata that varies by year. Year for which data value represents

Character four character - YYYY

domain

(CATEGORIZATION) Domain which variable belongs to. Highest level of variable categorization.

Character. Should fall into the DMC curated list of acceptable domains.

subdomain

(CATEGORIZATION) Subdomain which variable belongs to. Lower level of variable categorization.

Character.

var_label

(DETAILS) Short human readable variable label.

Character. Aim to be less than 30 characters. Hard limit at 75 characters.

var_def

(DETAILS) Details definition of what the variable is about. If categorical include coding here.

Character. No length limit.

value_type

(DETAILS) What type of data is the value?

Must be one of the following: - double (continuous values) - discrete (integer values) - categorical (non-numeric groups or categories)

units

(DETAILS) This is the short label to be appended to the data value. It will be used for annotating text or visualisations with a unit label (e.g. cases per 100k).

Characters

coding

(DETAILS) This is an optional internal field that describes in details the measurement

Characters

strata_description

(DETAILS) This should describe in detail what strata are available for this variable. Please include details about each strata if applicable.

Character. Try to be as detailed as possible. If a variable is not stratified the value should be NA.

source

(DETAILS) Data source

Character **Note if this varies by country please leave as NA in the variable level codebook and attach this field in the supplementary by iso2-yr codebook.

dataset_notes

(DETAILS) Any additional information can be added here. For example some dataset or file specific notes could be added here.

Character

limitations

(DETAILS) Place to describe any limitations for this variable.

acknowledgements

(DETAILS) Any acknowledgements for this variable

file_data

(INTERNAL) File name of the data file used for this dataset

Character.

file_codebook

(INTERNAL) File name of the codebook file used for this dataset

Character.

longitudinal

(INTERNAL) Is this variable qualified for longitudinal analysis or visualisations?

Binary: 1=yes or 2=no **Note if this varies by country please leave as NA in the variable level codebook and attach this field in the supplementary by iso2-yr codebook.

public

(INTERNAL) Is it okay to make the data and metadata for this variable publicly available for visualisation and download?

Binary: 1=yes or 2=no **Note if this varies by country please leave as NA in the variable level codebook and attach this field in the supplementary by iso2-yr codebook.

censor

(INTERNAL) If the data is not public. Can we show the metadata (e.g. show the data is available to be searched but not visualised or downloaded). This way others can request data if needed.

Binary: 1=yes or 2=no **Note if this varies by country please leave as NA in the variable level codebook and attach this field in the supplementary by iso2-yr codebook.

licence

(INTERNAL) What is the licence for this variable

Categorical. Options include "CC BY" or TBD by Kari and Goro.

fair

(INTERNAL) Has this metadata for this variable been updated for FAIR principles. If renovated as per new standards then yes. Otherwise no.

Binary: 1=yes or 2=no

TLDR (Too Long Didn't Read)

In trying to make a FAIR data portal we found two major challenges: 1) our existing codebooks were not accessible or comprehensive enough to support creating a FAIR data portal 2) no existing way to link complex metadata (by strata, country) to data. This page documents a proposed data and codebooks standard that will guide the FAIR renovation of existing datasets and serve as templates for future datasets.

- a set of identifier fields (`var_name`, `strata_fields`) to link metadata and data at a variable level while accounting for complex by strata/country/year metadata
- More comprehensive codebooks to explicitly codify important metadata (identifiers, strata details, categories, internal info, research details)

F3. Metadata clearly and explicitly include the identifier of the data they describe.

ELI5: Metadata is linkable to data

Description

The SALURBAL database is a collection of data variables; each variable has a unique identifier `var_name` (F1). If life were simple, metadata would match one to one with each variable and we could do linkage with just `var_name`. However, the pairing between variable and individual metadata

fields are not always one to one. SALURBAL metadata/data linkage scenarios are listed below based on prevalence.

1. **simple:** (Very common) Metadata links 1:1 to data at the variable level via `var_name` (e.g. domain, subdomain .. ETC). This linkage specific codebook would be called `codebook.csv`
2. **by_country** (Very common) This may be a common complexity where metadata differs by variable + country and needs to be linked by `var_name` and `iso2` (e.g. data source or censor status). This linkage specific codebook would be called `codebook_by_iso2.csv`
3. **by_year** (Uncommon?) metadata differs by variable + country and needs to be linked by `var_name` and `year` (e.g. data source or censor status). This linkage specific codebook would be called `codebook_by_year.csv`
4. **by_strata** (Rare) This case is rare but should be noted. Here metadata differs by variable + strata thus needs to be linked by `var_name` and `strata_id` (e.g. var_def or interpretation). This linkage specific codebook would be called `codebook_by_strata.csv`

The direct consequence of having multiple linkages between data and metadata is that for each dataset we need to **1) evaluate what type of linkage works best for each metadata field** then **2) operationalize separate linkage to specific codebooks for each of those linkages**. We will discuss each of these two steps further below.

1. Evaluate metadata linkage

The first step is to evaluate what type of linkage works best for each metadata field. The interactive table represents how you should fill out for the dataset you are trying to process. Moreover you can download [salurbal_codebook_evaluation.csv](#) which is a csv template for the required metadata fields, which shows by default all metadata that have simple linkage; use this as a starting point to evaluate the metadata linkage for your dataset.

Guidelines:

- this **linkage categorization for each field is mutually exclusive** (only one category per field). For now we assume linkage complexity exists at one level (by a single identifier). Let's try this for now and deal with more complex nomenclature later.
- [Salurbal codebook evaluation.csv](#) is a **template** containing a table of required metadata and possible linkage types.
- the template provided assumes everything is simple (which most of the time it is). Please **go through each field and assign a linkage type** by asking your team 'is this field going to vary by?' then update the template based on your reply.

field

description

simple

by_country

by_year

by_strata

domain

(CATEGORIZATION) Domain which variable belongs to. Highest level of variable categorization.

subdomain

(CATEGORIZATION) Subdomain which variable belongs to. Lower level of variable categorization.

var_label

(DETAILS) Short human readable variable label.

var_def

(DETAILS) Detailed definition of what the variable is about. If categorical include coding here.

value_type

(DETAILS) What type of data is the value?

units

(DETAILS) This is the short label to be appended to the data value. It will be used for annotating text or visualisations with a unit label (e.g. cases per 100k).

coding

(DETAILS) This is an optional internal field that describes in detail the measurement

strata_description

(DETAILS) This should describe in detail what strata are available for this variable. Please include details about each strata if applicable.

source

(DETAILS) Data source

dataset_notes

(DETAILS) Any additional information can be added here. For example some dataset or file specific notes could be added here.

limitations

(DETAILS) Place to describe any limitations for this variable.

acknowledgements

(DETAILS) Any acknowledgements for this variable

file_data

(INTERNAL) File name of the data file used for this dataset

file_codebook

(INTERNAL) File name of the codebook file used for this dataset

longitudinal

(INTERNAL) Is this variable qualified for longitudinal analysis or visualisations?

public

(INTERNAL) Is it okay to make the data and metadata for this variable publicly available for visualisation and download?

sensor

(INTERNAL) If the data is not public: can we show the metadata (e.g. show the data is available to be searched but not visualised or downloaded)? This way others can request data if needed.

licence

(INTERNAL) What is the licence for this variable?

fair

(INTERNAL) Has this metadata for this variable been updated for FAIR principles? If renovated as per new standards then yes. Otherwise no.

2. Operationalize linkage specific codebooks

Why linkage of specific codebooks rather than a fully merged codebook?

After we have evaluated the metadata linkage for a dataset, we will know which codebook and codebook variations to prepare. For each dataset we could potentially have up to four:

1. **codebook_simple.csv**: (Very common) will link to the data via only a single identifier `var_name` and contain all the metadata fields that were categorised as ‘simple’.
2. **codebook_by_country.csv** (Very common) will link to the data via `var_name` and `iso2`; it will contain all the metadata fields that were categorised as ‘by_country’
3. **codebook_by_year.csv** (Uncommon?) will link to the data via `var_name` and `year`; it will contain all the metadata fields that were categorised as ‘by_country’
4. **codebook_by_strata.csv** (Rare) will link to the data via `var_name` and `strata_id`; it will contain all the metadata fields that were categorised as ‘by_country’. Metadata links to data via

Integration into data pipeline

Simple linkage:

data.csv		+	codebook_simple.csv		=	Completed Merge		
var_name	data		var_name	var_def		var_name	data	var_def
APSPM25MEAN	1		APSPM25MEAN	Mean PM2.5		APSPM25MEAN	1	Mean PM2.5
APSNOXMEAN	2		APSNOXMEAN	Mean NOX		APSNOXMEAN	2	Mean NOX
APSNO2MEAN	3		APSNO2MEAN	Mean NO2		APSNO2MEAN	3	Mean NO2

By country

data.csv			+	codebook_by_iso2.csv			=	Completed Merge			
var_name	iso2	data		var_name	iso2	censor		var_name	iso2	data	censor
LIVERCANCER	AR	2.41		LIVERCANCER	AR	1		LIVERCANCER	AR	2.41	1
LIVERCANCER	BR	1.2		LIVERCANCER	BR	0		LIVERCANCER	BR	1.2	0
OBESITY	AR	12.5		OBESITY	AR	1		OBESITY	AR	12.5	1
OBESITY	BR	16.5		OBESITY	BR	1		OBESITY	BR	16.5	1

By Year

data.csv			+	codebook_by_year.csv			=	Completed Merge			
var_name	year	data		var_name	year	source		var_name	year	data	source
LIVERCANCER	2000	2.41		LIVERCANCER	2000	Health survey 2000		LIVERCANCER	2000	2.41	Health survey 2000
LIVERCANCER	2010	1.2		LIVERCANCER	2010	Mortality records 2010		LIVERCANCER	2010	1.2	Mortality records 2010
OBESITY	2000	12.5		OBESITY	2000	Health survey 2000		OBESITY	2000	12.5	Health survey 2000
OBESITY	2010	16.5		OBESITY	2010	Mortality records 2010		OBESITY	2010	16.5	Mortality records 2010

By Strata

data.csv			+	codebook_by_year.csv			=	Completed Merge			
var_name	strata_id	data		var_name	strata_id	var_def		var_name	strata_id	data	var_def
LIVERCANCER	sex_male	1.5		LIVERCANCER	sex_male	... for males.		LIVERCANCER	sex_male	2.41	... for males.
LIVERCANCER	sex_female	1.2		LIVERCANCER	sex_female	... for females.		LIVERCANCER	sex_female	1.2	... for females.
OBESITY	sex_male	12.5		OBESITY	sex_male	... for males.		OBESITY	sex_male	12.5	... for males.
OBESITY	sex_female	16.5		OBESITY	sex_female	... for females.		OBESITY	sex_female	16.5	... for females.

F4. (Meta)data are in a searchable resource

(Meta)data are registered or indexed in a searchable resource.

Takeaway

We built a bespoke web application to index and make our data searchable. We leverage several fundamental technologies.

- Blob storage (Azure): highly scalable, performant, and affordable storage solution. We organised files via our project's global identifiers `var_name`; consequently, our Blob storage container functioned as an API for our front-end application.
- Front-end application (Next.js): Next.js is a full stack JavaScript web development framework which can be built via static site generation (SSG) or work with server-side rendering (SSR). We chose to deploy our web application as a statically generated site because the content of our data portal would largely be static other than between large data updates. SSG at build-time allows us to deploy a static site at a fraction of the cost of a web application that requires a server.

- Serverless architecture for computation (Azure Functions). We also wanted to allow users to compile data and download a custom data extract. While the actual selection UI can be built with JavaScript and run client-side, the final data merges, file preparation and email notification could be problematic both resource and security-wise to be run client-side. We leverage serverless computing to handle these more complex or computationally intensive actions. Compared to Cloud 1.0 computing, serverless computing takes care of the server administration and allows us to develop bespoke solutions that are comparably performant, much more affordable and easier to maintain.

Goal and Tool Alignment

Why not use ICPSR or CKAN to FAIR principles for SALURBAL?

It is key to recognize that CKAN or ICPSR are solutions to global / multi-national / multi-disciplinary FAIR data problems. FAIR data implementation for our own organisation needs only to serve our community of researchers - think 100 epi/public health researchers rather than millions of researchers of all fields. Given the dramatic difference in the scale of the problem, it is logical that the solution could be and likely should be different.

Baggage of traditional portals

They are by design **extremely generalisable** - to be widely or generally applicable to all types of projects, data or organisations and allow **large scale open data contribution** - mechanisms to allow authentication, authorization, and uploading from thousands of contributors. These selling points, which make them good platforms for serving a broad community of hundreds of thousands of researchers and organisations from all types of fields, actually make them less appealing for use by a single organisation. Firstly, it is **difficult for an extremely generalizable platform to be tailored to specific needs of an organisation**. Secondly, the **technical infrastructures that support large scale open data contribution are expensive to maintain, difficult to develop** and require dedicated staff with IT/DevOps expertise.

Modern solutions

While these platforms make sense for really large FAIR data initiatives, they may not make sense for us due to cost and expertise. Importantly many of these platforms were built over a decade ago (e.g. CKAN was released in 2006) with infrastructures that are in many senses out of date. This document will give a high-level overview of how we leverage emergent technologies (open source UI frameworks, cloud infrastructure and serverless computing) to build a highly customizable and resource-efficient data portal.

Open source user interface frameworks

In the past few years, large tech companies have fortunately open sourced their internal frameworks for making web content (Facebook - React.js, Vercel - Next.js). With these tools it is now able to

make incredibly customizable, scalable and complex websites all in JavaScript. These JavaScript have some really big selling points:

- They don't need virtual machines or a server to run (and so are essentially free to deploy anywhere).
- There is an incredible amount of open-source education resources for these open-source frameworks (so they are much easier to learn and master).

Even the creators of [CKAN are moving towards this direction](#).

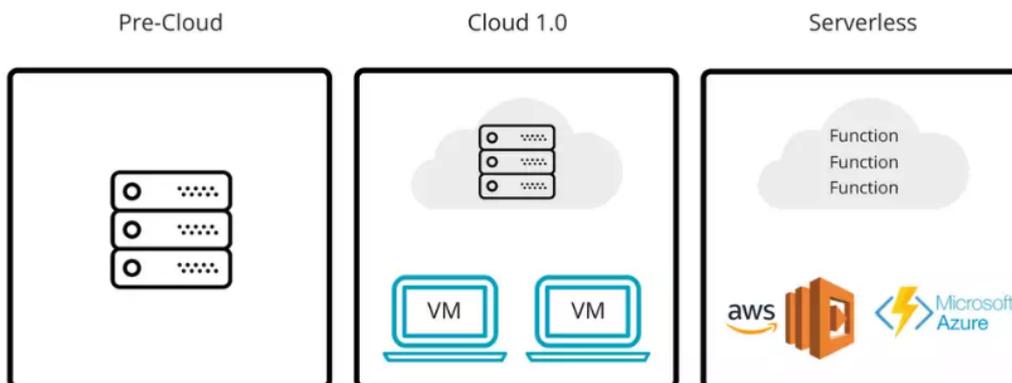
Even so, there's a high learning curve before you can build a proper application. That's because you need to learn about Python, templating, data loading and so on. If you'd like to integrate content or rich visualisations, things are even more complex.

So, we need something simple but customizable.

Think about how apps are created as a front-end developer. You create some files, write some code, load some data, and then simply deploy it. We don't have to worry about Docker, Kubernetes, data storage, Postgres etc. We are using React.js and Next.js to build the front-end of our data portal.

Serverless Computing

Evolution of Cloud



- **Pre-cloud:** buy servers, administer to servers, and run code on them.
- **Cloud 1.0** rent and administer to servers and run code on them.
- **Serverless** run code on servers that are administered by a cloud provider.

CKAN utilises Cloud 1.0: you are just renting servers which you have to administer. While you can easily get access to many servers you still face traditional IT problems: - you have to rent servers for each application; need to learn server technologies (linux, dockers, nginx) to use them; need to worry

about server scalability. What server should I rent? How much RAM do I need? How much computing power do I need?

Serverless architectures move away from renting and administering and directly to renting computing power as needed. Rather than the provider giving you a server, you give them code and they worry about where it should run. **This is incredibly cheap and lets us focus on building our application rather than server administration.**

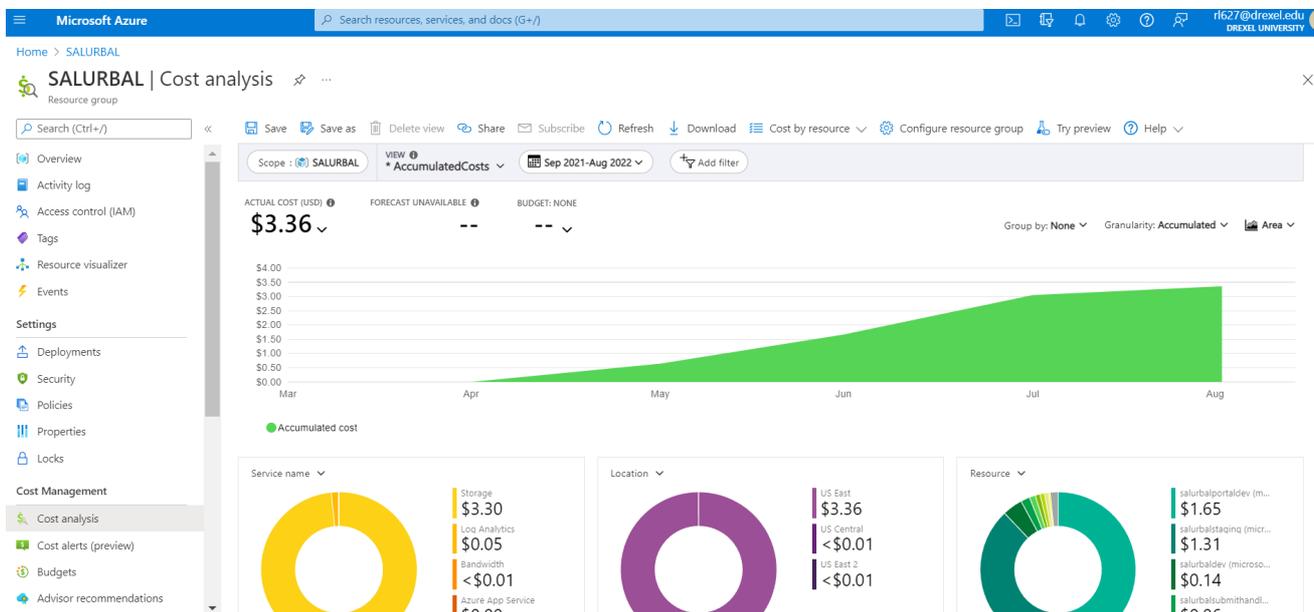
Another metaphor for serverless: Think of serverless computing like renting a furnished apartment. Instead of worrying about buying and maintaining furniture (the way you would with servers), you simply enjoy living there. The landlord (the cloud provider in our analogy) takes care of all the repairs and maintenance, so you don't have to worry about any of that.

In the serverless model, you only focus on what you want to do: living your life (or running your code, in tech speak). You also only pay for the time you are actually using the apartment, similar to how you only pay for the computing power you use with serverless. This gives you more time and resources to focus on what truly matters, whether that's living your life or building your application.

Direct Value

Highly customizable application.

Basically free.



8.3. Accessible

A1: (Meta)data are retrievable by their identifier using a standardised communications protocol. The protocol is open, free, and universally implementable. `

A1.1

A1.1 The protocol is open, free, and universally implementable.

We use HTTPS to facilitate data transfer. There are two main ways for users to download data: 1) each variable's web page contains download links for data and associated metadata 2) users can create a custom data extract on the data portal and have the data/metadata operationalised into an analytical bundle that is then emailed to them.

A1.2

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.

We deploy the data portal via the Azure Static Web Application Service. This service has built-in authentication via OAuth integrating with GitHub or Outlook to handle user authentication. We then write authorisation logic as needed in our front-end Next.js application to control information flow. Although this mechanism was developed, we do use it to primarily track user behaviour.

SALURBAL data is stored on a private encrypted server hosted by Drexel University. Data items that have contractual restrictions remain on this private encrypted server. Data items that are open to the public are then stored within Azure Blob Storage with anonymous access (anyone can have access to them).

A2

A2 Metadata are accessible, even when the data are no longer available.

We plan to submit metadata to ICPSR. Metadata will persist via their stewardship independent of our own within-project data portal.

8.4. Interoperable

Data needs to work with applications or workflows for analysis, storage and processing.

I1

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Data and metadata are stored in Azure Blob Storage in both CSV form as well as JSON format. JSON is highly interoperable as a REST API for web applications; for example, our data portal pulls from this to provide analytics/visualisations. CSV is a common data sharing format in the epi field; any

download links or downloaded data from our data portal comes in CSV format as it is more human friendly.

I2.

I2. (Meta)data use vocabularies that follow FAIR principles.
TBD

I3.

I3. (Meta)data include qualified references to other (meta)data.
TBD

8.5. Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1

R1.1. (Meta)data are released with a clear and accessible data usage licence.
Our funders mandate that all research papers fall under a CC BY 4.0 licence. Most likely publicly available data will be under the same licence, pending confirmation by Kari and Fer.

R1.2

R1.2. (Meta)data are associated with detailed provenance.
TBD

R1.3

R1.3. (Meta)data meet domain-relevant community standards.
This field is really community specific. What are the metadata requirements for SALURBAL? From a machine-actionability perspective, the minimal metadata requirements are detailed in [F2](#). These fields fall into the following categories:

- **Identifiers**: columns responsible for linkage of data and metadata
- **Categorization**: columns responsible for categorising variables into domain or subdomain

- **Details:** research related variables details, this will be useful for users who want to reuse our data/codebooks.
- **Internal:** internal project related metadata



9. Appendix B: Summary of guidelines and recommendations for FAIR practices in health survey data harmonisation

The guidelines and recommendations for FAIR practices in health survey data harmonisation that are discussed in this document will be further refined and presented in an academic article. Additional detail is presented in this Appendix.

Authors: Kari Moore, Mariana Lazo, Ana Ortigoza, D. Alex Quistberg, Brisa Sanchez, Binod Acharya, Tania Alfaro, Maria Fernanda Kroker-Lobos, Mariana Carvalho De Menezes, Olga Lucia Sarmiento, Amanda de Souza, Carolina Perez Ferrer, Luz Mery Cardenas, Akram Hernandez Vasquez, Waleska Caiaffa, Ana V. Diez Roux, And SALURBAL group

Key Features of the data used

- This data resource is part of the SALURBAL (Salud Urbana en America Latina/Urban Health in Latin America) project. SALURBAL aims to evaluate the impact of physical and social features of urban environments on health, health equity, and environmental sustainability in order to inform urban policies worldwide.
- The SALURBAL health survey data resource includes harmonised health survey data (non-communicable disease risk factors, adult and children) that can be geolinked to built, natural, and social environment data for at up to three levels: cities, subcity units (e.g. Municipios of counties that compose cities), and neighbourhoods (similar to US census tracts)
- The data includes 27 surveys from 11 countries. Survey years ranged from 2000 to 2018. The range of surveys per country is 1-5. The total number of adults (aged 18-100) and children (aged 0-17) included in the surveys are 246,738 and 69,126, respectively. In 6 countries there are multiple repeat surveys over time.
- Data available: Variables include individual-level demographic and socio-economic characteristics, alcohol/tobacco use, anthropometry, diet/physical activity, diabetes, hypertension, mental health, self-reported health, for adults' data are available for most of the survey respondents. For children, the dataset includes household-level demographic and socio-economic characteristics, and anthropometry.
- Researchers interested in the SALURBAL project can contact: SALURBAL@drexel.edu.

Data Resource

The SALURBAL health survey data includes 27 surveys from 11 countries (Table 1). Survey years ranged from 2000 to 2018, the range of surveys per country is 1-5. The total number of adults (aged 18-100) and children (aged 0-17) included in the surveys are 246,738 and 69,126, respectively. In 5 countries there are multiple repeat surveys over time. Data available include individual-level demographic and socio-economic characteristics, alcohol/tobacco use, anthropometry, diet/physical activity, diabetes, hypertension, mental health, self-reported health, for adults. For children, the dataset includes household-level demographic and socio-economic characteristics, and anthropometry.

Table 1: Summary of health surveys included in SALURBAL data resource. By country name (alphabetical order) and survey year.

Country	Survey name	Age	Child data	Survey year	SALURBAL Sample size
Argentina	Encuesta Nacional de Factores de Riesgo, ENFR (National Risk Factors Survey)	≥18 years	No	2005	25,753
				2009	16,218
				2013	21,451
Brasil	Pesquisa Nacional de Saúde, PNS (National Health Survey)	≥18 years	No	2013	29,353 in L2s 40,703 in L1ADs
				2019	33,515 in L2s 46,767 in L1ADs
Chile	Encuesta Nacional de Salud, ENS (National Health Survey)	≥15 years	No	2003	2,032
				2010	3,140
				2017	3,805
	Encuesta Longitudinal de Primera Infancia (ELPI) (Longitudinal Survey of Early Childhood)	1-12 years	Yes	2017-2018	6,723

Colombia	Encuesta Nacional de Salud, ENS (National Health Survey)	18-69 years	No	2007	43,182
	Encuesta Nacional de la Situación Nutricional en Colombia, ENSIN (National Nutritional Situation in Colombia)	0-69 years	Yes	2005	42,336 adults (18-69) 23,794 children (<18)
				2010	55,863 adults (18-69) 30,278 children (<18)
				2015	36,593 adults (18-69) 17,104 children (<18)
Costa Rica	Encuesta Multinacional de Diabetes mellitus y Factores de Riesgo, CAMDI (Multinational Survey of Diabetes Mellitus & Risk Factors, Central American Diabetes Initiative)	≥20 years	No	2005	1,427
Guatemala	CAMDI (See Costa Rica)	≥20 years	No	2002-2003	1,397
	Demographic and Health Survey (DHS)	Females 18-49 years Children <5 years	Yes	2014-2015	2,730 adult females (18-49) 983 children (<5)
Nicaragua	CAMDI (See Costa Rica)	≥20 years	No	2003	1,397
Mexico	Encuesta Nacional de Salud, ENSA (National Health Survey)	Adults ≥18 Children <5 years	Yes	2000	29,733 adults (≥18) 23,966 children (<5)
	Encuesta Nacional de Salud y Nutrición, ENSANUT (National Survey for Health and Nutrition)	All ages	Yes	2006	31,532 adults (≥18) 19,431 children (<18)
				2012	26,335 adults (≥18) 25,014 children (<18)
				2016	14,618 adults (≥18) 3,274 children (<18)
2018	27,118 adults (≥18) 7,538 children (<18)				

Panama	Encuesta Nacional de Salud y Calidad de Vida ENSCAVI (National Survey of Health and Quality of Life)	≥18 years	No	2007	11,394
Peru	Encuesta Nacional de Demografía y Salud, ENDES (National Survey of Demographics and Health)	Adults ≥15 years Children <5 years	Yes	2016	11,929 adults (≥18) 8,547 children (<5)
El Salvador	CAMDI (see Costa Rica)	≥20 years	No	2004	1,872
	Encuesta Nacional de Salud Familiar (National Family Health Survey)	Females 18-49 years Children <5 years	Yes	2008	4,297 adult females (18-49) 1,290 children (<5)
	Encuesta Nacional de Enfermedades Crónicas no transmisibles en Población Adulta de El Salvador ENECA (National Survey of Non-communicable Chronic Diseases in the Adult Population of El Salvador)	≥20 years	No	2014-2015	1,546

10 Good practices and recommendations in handling and harmonising Health Survey Data

1. Identify sources of data representative of large populations (usually national surveys or surveillance systems), that assure the data involves great representation across countries/ cities and over time.
2. Use of harmonisation approaches that are rigorous but flexible to accommodate differences across surveys.
3. Identify and collate survey questions and responses by selected domains under study (i.e., health risk factors)
4. Compare wording questions across surveys with special attention to skip patterns in the questions asked on the survey to understand whether questions across surveys share the same respondent universe (and therefore are comparable).
5. Compare and contrast quality of questions and questionnaires with surveys conducted by other Institutions such as the Centers for Disease Control and Prevention or the World Health

Organization for standard variable definitions as well as harmonisation approaches proposed by other projects.

6. Propose harmonised variable definitions and response categories with attention to differences in wording across countries.
7. Apply harmonisation processes that can balance specificity of the information retrieved versus the amount of countries/ units of analysis that can include the selected data. Some proposed strategies to consider in this process are:
 - i. Creation of multiple versions due to country differences that did not allow a single harmonised variable (e.g., diabetes, hypertension).
 - ii. Unit conversion (e.g., height, weight).
 - iii. Collapsing categories (e.g., education, self-rated health).
8. When possible, it is encouraged the use of standardised questionnaires and scales for the harmonisation of questions that assess health behaviours through different set of questions (such as physical activity, or dietary habits)
9. Create harmonisation protocols that can document the decision taken during the harmonisation process and that can assure its replication in the future, with special attention to:
 - i. The systematisation and standardisation of variable names across different years and versions of surveys
 - ii. The maintenance of consistency in the coding of answers and creation of answer categories across different years and versions
 - iii. The incorporation in the documents of the original questionnaires used for the harmonisation.
10. Create log documents that can track the different iterations made during the harmonisation process with special focus in the description of survey limitations or errors found, and the changes made in the harmonisation and coding as response to these challenges.