CHIDS Database Technical Design

July 1, 2014

Introduction

The Community Health Information Database System (CHIDS) is the primary data platform for the Community and School Together Project (CAST[1]). Rather than a single database, the CHIDS system comprises a set of databases linked across two institutions: the Oregon Research Institute (ORI[2]) and Lane Council of Governments (LCOG[3]). Together, the two institutions manage their relevant CHIDS data collection and storage activities. Summary data sets, reports, and supporting documents are exchanged between the two institutions, based on the functions of the institution within the CAST project (Figure 1).

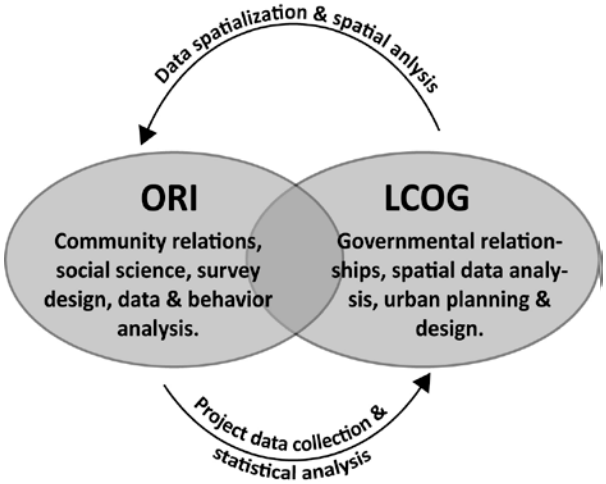

Figure 1. Functional institutional relationships on CHIDS.

CHIDS structure and administration

The databases of the CHIDS system come from a variety of sources (Figure 2), and each institution collects, maintains, and summarizes the information collected as project needs arise. Figure 2 highlights the data sharing around school student data and street-segment built environment data. School Student data is maintained by the Bethel School District, a partner in the CAST project , while street segment data is maintained by LCOG.

Research staff at ORI developed close working relationships and administrative structures by establishing a project staff school coordinator position within the Bethel School District . This mutually hired employee, paid by ORI, split their time between the District and ORI which allowed for the effective and secure access to anonymous confidential data of those students participating in the CAST project. Following each data transfer, the ORI technical staff would forward a refreshed dataset to LCOG.

---

[1] The Communities and Schools Together (https://cast.ori.org/) project is a unique partnership among the Bethel School district, Oregon Research Institute, and several community organizations. The mission of CAST is to support schools, parents, and community groups in reducing childhood obesity. CAST does this by working together to increase neighborhood health and safety for elementary school children.

[2] Oregon Research Institute (http://www.ori.org/) is an independent behavioral sciences research center dedicated to understanding human behavior and improving the quality of human life through the prevention and treatment of health, educational, and social problems.

[3] Lane Council of Governments (http://www.lcog.org/) is a one-stop destination for GIS and other services to local governments, agencies and research groups in the Lane County region and beyond.

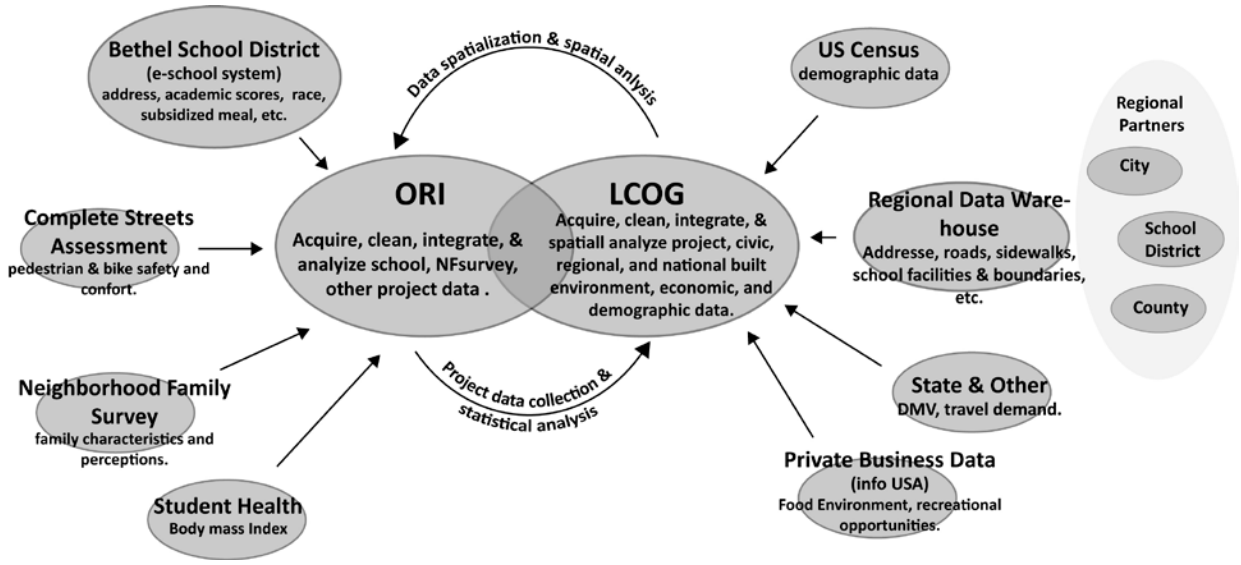

Figure 2. Data sources and flow processes.

ORI staff was responsible for maintaining year-to-year cohort tables and all collected student attributes. These tables are refreshed on an annual basis, typically in October of each school year, to include active current CAST students as well as new K students and new enrollees in grades 1-5. In addition to grade and school, variables collected include race, ethnicity, spoken language (ie, English or Spanish), DOB, gender and free and reduced lunch status, as well as academic scores from the prior school year.

Following is a description of the CAST Goals and separate sections on details about the datasets maintained at LCOG and ORI.

Project Goals

The overarching goal of CAST is to understand the relationship of complex social, environmental, nutritional, and physical activity factors to childhood obesity. A related objective is the development of effective interventions for obesity prevention in elementary school children. In support of these objectives the project uses a Community Based Participatory Research (CBPR) model to involve community participants and local organizations in addressing the some six specific sequential aims:

Aim 1: To establish a facilitated, community-based participatory research collaboration that that designs, implements, and evaluates project activities useful in understanding and responding to local multi-systemic issues relevant to the prevention of childhood obesity.

Aim 2: To develop a community-based, multicomponent health monitoring system that contains information regarding local community demographics, school health profiles, social and built environments, and food systems.

Aim 3: To use the information derived from the developed health monitoring system to examine relationships between social and built environment, existing food systems, and other obesity risk factors faced by elementary school children and their families.

Aim 4: To develop and design a community-based intervention that focuses on training and educating parents of school children with respect to information and knowledge about social and built

environment and healthy food consumption that contribute to children's healthy lifestyle.

Aim 5: To implement and evaluate the effectiveness of the parent training and education program in a community setting.

Aim 6: To summarize research findings of this CBPR project and convert them into a CAST research model that can be used for effective dissemination.

Aim 7: To communicate the results from this community-driven CBPR project among participating partners, and to disseminate them in a broader community setting.

The proposed research is significant because it is the first study to develop a comprehensive community health monitoring system for obesity prevention and intervention in a major public health setting—a school community. Such an effort is critical in tracking rates of children's obesity, analyzing environmental and social influences of obesity, and developing strategies for effective and timely community-based obesity prevention and intervention programs. The proposed approach is also innovative because it involves and engages local community participants and key stakeholders to work collaboratively on the basis of mutual interest, trust, and equality in all major phases of the project, including community needs, availability of resources, and system effectiveness. By establishing partnerships among schools, community organizations, and academic institutions, the outcomes of this project are expected to have important public health ramifications with respect to program development, evaluation, implementation, and dissemination in typical settings

The management and storage of CAST data within CHIDS involves two related and essential considerations: the nature of the data in its use within the project, and how—specifically, in what data formats—the data are stored.

LCOG Maintained Datasets

The geo-spatial data comprising the CAST project can be classified into three general groups of data: project-data, census data, and cartographic reference data.

1. Project data includes all data originating from the CAST project via ORI's efforts: school student data, family survey data, student health data, and CSAT built environment data. It also includes subsets of regionally maintained data, such as street segment centerline data, sidewalks, and addresses, often with improvements or additions, such as household data from the NFS applied to the regional address file, with the subset of addresses surveyed entering into CAST project data.
2. Census data is data collected primarily for the 200 census, with 2010 decadenal and 1-3 year American Community Survey (ACS) data to be incorporated as it becomes available. This census-based data serves as a demographic backdrop for comparing CAST project participants to their geographic area, and for comparing the area of the bethel School District to the Eugene-Springfield metropolitan area. Some census data for the 1990 census and for more historic 1960, 70, and 80 censuses have been collected as well.
3. Cartographic reference data refers to those data sets that are necessary for providing contextual reference in project-produced maps but is not an active or intimate used data set within the CAST project. These data themes are features such as roads or streams, but may also include project data stored in alternate formats for cartographic display, such as school attendance boundaries stored as lines rather than areas purely for cartographic production.

ORI Maintained Datasets.

All individual data collected from participating CAST students and their parents used a unique identifier (CASTID for students and an address derived household identifier, HHID, for parents) to identify the respondents and permit linking of records across datasets. These datasets were the source for' project data' (see above) maintained at LCOG.

- **School Individual Student Data** – CAST Student information from the district electronic records system (e.g., student height, weight, address, attendance, achievement scores, race/ethnicity, DOB, and eligibility for free/reduced lunch) was made available to the project during the fall quarter of each school year in an anonymous format (i.e., school data provided to the project used CASTID with a link file relating CASTID to student ID maintained only by the district)
- **School District Data** – aggregate data at both school and district levels were updated each fall to capture enrollment, proportion of minorities, eligibility for school lunch subsidies, and proportion of students meeting academic achievement standards.
- **BMI** – Data elements required to calculate a Body Mass Index for elementary school students (height, weight, age, gender) were collected for all CAST students as part of the district health screening process conducted during the fall of each of the 5 project years. The Epi Info (Centers for Disease Control, Atlanta, Georgia) analysis tool, NutStat, a nutrition anthropometry program that calculates BMI from age and gender normed tables, was used to calculate raw BMI, *z*-score, and percentile equivalents using CDC (2011) guidelines.
- **Family Survey** - The Family Survey included questions about family choices surrounding food and eating habits , nutrition, shopping practices, exercise, perceptions of neighborhood safety, transportation, and accessibility to recreation. The survey also included basic demographic data and a short food security questionnaire. The 4 year longitudinal survey was conducted with a random sample of CAST parents beginning in year 2 of the project, augmented by the addition of parents of new kindergarten students in years 3 and 4.
- **Farm to School Questionnaire** - The Farm to school project included a pre- and post test given to the students in grades 2 and 3 in two District schools who participated in the project sponsored Farm to School program during years 2-4 of the project. The survey was designed to probe student knowledge and behavior in four areas; 1) Preference for fruits and vegetables; 2) Consumption of fruits and vegetables; 3) Knowledge of where food comes from; and, 4) garden concepts.
- **SRTS Survey and Teacher Tally** - The Safe Routes to School (SRTS) survey consisted of two standardized forms originating from the National Center for Safer Routes to School: - an in classroom student transportation tally and a parent transportation survey. The in-class student collected data pertaining to how students arrived and departed from school each day of ta typical week. The parent survey is collected data on student modes of travel and parent perspectives on important issues impacting their children's health and safety. SRTS data were collected during the first and third years of the project from students and parents in seven CAST elementary schools. Unlike the other datasets SRTS surveys could not be linked to a CAST ID , thus restricting their use in analyses to school level predictors.
- **Healthy Moves Measures -** In conjunction with a school based physical activity program, a survey of reported physical activities as well performance on tests of physical activity was collected for CAST students who participated in the program during year 3 of the project .

Data Storage Format and Management.

CAST data at ORI are stored in a relational database format (Microsoft Access database), from which data are exported to the Statistical Package for the Social Sciences (SPSS) environment for calculation of basic descriptives and creating a codebook for each dataset . Data for more complex analyses are extracted from access tables using an ODBC connection and processed using packages from the R statistical and graphical software environment. Microsoft Excel is also used for data entry and transfer between applications and staff, typically when individual tables or simple query outputs are needed. These Excel files are always password-protected.

CAST data stored at LCOG is maintained in a combination of several Microsoft Access Jet-4 database applications, including Environmental Research Systems Institute's (ESRI) implementation of this database format as the ArcGIS Personal Geo-database, and ESRI's file-based geo-database format. Regional data at LCOG is stored in a combination of formats, with a general institutional migration to ESRI's Spatial Database Enterprise (SDE) architecture. Currently, all forms of ESRI data types are used for both maintenance and analysis.

The decision of what data sets to store in what data storage formats arises directly from the advantages and disadvantages offered by (1) the storage formats, the software that can access and make use of these formats, and (2) the intended use of the data . ESRI's ArcGIS software is capable of performing a range of spatial analysis and spatial database functions, such as geo-coding, network routing, and spatial overlay, and can access a range of data storage formats, but is generally poorly equipped to execute true RDBMS relational analysis. Use of the MS Jet-4 database engine has some limitations, such as a 2 Gb size limit, but can be accessed from both Microsoft's Access application and ESRI's ArcGIS suite of applications through Open Database Connections (ODBC). ESRI's file-based geo-database lacks ODBC connections, but does allow faster data retrieval and display in ESRI software and an unlimited file storage size.

Given these considerations and the three general data groups discussed previously, the following configuration is used for the storage, spatial and relational analysis, and cartographic display of CAST data. CAST project geo-data is stored in an ESRI personal geo-database, and paired with an MS Access database which contains links to the feature attribute tables of the geo-data for use in SQL queries. Census data is stored similarly, with features and their attributes stored in a ESRI geo-database, and downloaded tables and summary queries stored in a linked MS Access database. Cartographic reference data is usually larger in spatial extent for ease of map creation, and is not called upon for analytic uses, and so is stored in an ESRI file-based geo-database.

This combination of storage formats allows for full and direct use of GIS spatial analysis applications, robust query and summary analysis of both tabular and geo-processed or feature attribute data, and tabular and cartographic display of base data and analysis results. This architecture is described further in Table 1, below.

Database Name	**Format**	***Contents***
CAST_geo.mdb	ESRI Jet-4 Personal Geo-database	*All project-based geographic data sets and some related tabular data.*
CAST_qry.mdb	Microsoft Access Jet-4 Database	*All tabular data imported from various sources, links to geographic data sources in CAST_geo.*
Census_geo.mdb	ESRI Jet-4 Personal Geo-database	*Census reporting boundaries for Lane County for 1960-2010. Selected Census tables for demographic characteristics.*
Census_qry.mdb	Microsoft Access Jet-4 Database	*Downloaded (1990-2010) or constructed (1960-1980) US Census tabular data. Queries summarizing census data or relating census and project data sets.*
CAST_Cartographic.gdb	*ESRI File-based Geo-database*	*Cartographic reference data, such as roads and hydrologic features, for productions of maps.*

Table 1. CAST CHIDS databases.

Data Relationship Management

The use of spatial databases allows the association of entities with each other purely by their spatial relation to each other. This poses special challenges not typically encountered in tabular database management, but in the end can be handled through the association of database keys after a spatial relationship has been established.

Primary Key, CAST_ID. The most atomistic unit of analysis in the cast project is the child. Each child in the CHIDS database is identified by a unique CAST_ID value, the primary database key for the dataset. Each new cohort entering into the project database is assigned a new unique id.

Foreign key, Household ID. A household ID is established simply by running a frequency on normalized addresses and assigning a unique identity to each address found in the subject data. Several other foreign keys are used in the CHIDS, some, such as the US Census STF ID used for identifying census based geographic units, and Bethel building ID, used for identifying different schools. Database Keys used in the CHIDS are summarized in Table 2, below.

Key field name	Dataset	Database	Description
CAST_ID	CAST year cohort counts, CAST student health data, derivative data sets .	CAST_geo.mdb CAST_Qry.mdb	Unique Identifier for each student in the CAST project
HH_ID	CAST year cohort counts, CAST student health data, derivative data sets .	CAST_geo.mdb CAST_Qry.mdb	Unique identifier for CAST participant households.
Building - E/W*	School identity from the Bethel School district	CAST_geo.mdb CAST_Qry.mdb	Unique Bethel-identifier for schools.
ATTEND	School identity from regionally maintained facility and attendance boundary data.	CAST_geo.mdb CAST_Qry.mdb	Unique LCOG-identifier for schools.
STFID	US Census unique geography identifier.	Census_geo.gdb Census_qry.gdb	Unique identifier for Census-based geography.
DL_ID_NUMBER	Unique identifier for Department of Motor Vehicle data (ODOT).	CAST_geo.mdb CAST_Qry.mdb	Drivers license id number for identifying licensed vehicle drivers.

* Field names in GIS databases cannot use special characters such as "/" or "–"; filed stored in GIS-based databases will substitute "_" for spaces or special characters.

Table 2. CAST CHIDS database keys

Data Update and Transfer

As mentioned above, each October a series of tables are requested by ORI from the school district. In general, the transfer of data from the district to ORI each fall comprises three tables: A demographics file, an attendance report and a test scores output. All tables are received in Excel format. The tables are then imported (as opposed to being merged, which comes later) into the main Access database.
The merging of the Demographics file is the most complex procedure, since the worksheet contains the majority of the student data in "vectored" or stacked format. After checking that the file has been filtered for parents who declined the project, the records are scanned with an eye toward students with multiple records. The basis for an additional record or "vector" is a student changing address, which results in an addition line in the data in which the address changes, but other variables are copied into the new row.

These multiplicity of vectors, as well as the appearance of new students every year, requires a multi-step process for merging in the new year's records. In general the process first isolates existing students, verifying whether their address has remained the same or been updated. Existing students with new addresses then have their old addresses archived. Then new students are simply appended to the dataset. The final step is to query for existing/former CAST students who have returned to the current dataset after moving or becoming inactive during the past school year.
Once the merge is complete, families are assigned a Household ID, which is meant to link siblings and families members residing at the same address. This variable was created due to the lack of a family linking ID within the school district dataset. At that point, the new year's cohort is established for the current school year, and is the basis for eligible students from which our Family Survey sample (among other data collection efforts) is established.

While the format of the tables has been refined slightly ever year, each "pull" of student data has followed a similar pattern in terms of the number and structure of the Excel tables received. Race and ethnicity is an example of a variable (or variables) that eventually comprised their own table. This occurred after the school district updated their storage of race/ethnicity to be more in line with federal

census standards. Updating their system meant more data for each child, and more complexity in reporting, so a table dedicated to just race/ethnicity is now a standard procedure.

Per agreement with the school district, CAST maps will visually display household information not at the single household level, but aggregated at the level of three.
School data with student attributes is geocoded and then provided to ORI in summary or derivative form, such as imputed household relationship, geographic placement regarding district or school transfer, or imputed route to school.

To impute routes to school, built environment data was sourced from the City of Eugene Public Works department in terms of sidewalk data and street segment data. This data was processed by project staff at LCOG into several related data sets, including streets with additional attributes, intersections, and networked sidewalk data sets. The street segment data was then used in the CSAT project to contain the segment-based assessment of pedestrian and bike access. This data was then shared with Eugene public works staff for use in transportation planning.

References

Centers for Disease Control (CDC). *About BMI for children and teens. Updated June 2, 2011.* Available at: http://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html. Accessed August 1, 2012.