resilience.io

data collection strategy (abbreviated)

This is a brief description of the data collection strategy that has been developed for the **resilience.io** platform which seeks to model the spatial and temporal dynamics of a city and its supply hinterland. This strategy provides generic procedures that can be applied to the 14 sectors which fit within the 5 building blocks as described in the data specification (abbreviated) publication.

To be able to build these components, there are four types of datasets that need to be collected, cleaned and verified for use:

Geospatial data - the data to define the spatial surface area of a city-region, that forms the base map from which other datasets are attached, will be organised using three tables. The first accounts for naming the area, calculating the surface area and type of area (district, neighbourhood etc) . The second accounts for spatial points and vertices, the boundary of an area. The third covers land use type, extent and relevant attributes.

Population and household data - a set of seven tables is used to organise data describing the city population and households including people's activities, employment, education, health and demographics information. Additional socio-economic data can be added as required given differing city region contexts.

Economic flow data - is used to represent any data value related to a flow, including physical flows in terms of demand and supply, as well as financial flows and quality characteristics thereof. The economic flow datasets serve to assess demands, to inform cost and to aid in the establishment of model relationships e.g. the relation between water consumption and socio-economic characteristics of households.

Infrastructure data - data on infrastructure processing technologies and distribution networks will be captured e.g. water treatment plants and related distribution infrastructure within the city region. These two types of infrastructure data will be organised in two tables each including, name, type, operational status, capacity, location and implementation year.

With this structure in place, the data quality requirements for population of **resilience.io** must be fulfilled as follows:

1. **Data screening & collection** - a data inventory is developed with a summary of the relevant datasets, the level of detail/coverage at spatial, temporal, or categorical levels, and meta-data for the source plus the utilised data collection method.

2. **Raw data collection** - existing data is extracted, into an XLS/CSV format, and meta-data expanded to include data collection methodology; sample size, granularity, data age and collection time.

3. **Data quality checks** - once raw datasets have been collected, standardised procedures are employed to assess gaps, suspicious values, and inconsistencies. Any data value that is found to fall under one of the check categories will be noted in the raw dataset. These include checks for consistency (match to known totals or distribution); missing values (within a time series or categories); duplicate values; discontinuities (to account for measurement error and adjustment in measurement techniques) and outliers. Once this is complete, raw data sets are copied to ensure no original data is lost and cleaning adjustments as described above are applied.

4. **Data quality verification** - order of magnitude analysis and multiple dataset similitude comparison is used for final iterative quality checks and investigation. These serve as additional final quality checks for entire datasets to assess whether the datasets need further investigation.

5. **Data harmonization** - a large number of datasets need to be integrated using rule sets, derived from smaller samples, to ensure validation and robust representation of the city region. Data can then be used to populate **resilience.io.**

This approach will be used for early **resilience.io** prototype versions. As we move through the development phases we will be automating these processes managed through our data brokerage system, an interface that will find, validate, incorporate and connect disparate data sources to the model and users, for example, census data, corporate databases and sensor data, as well as relevant region specific initiatives like INSPIRE and GEOSS which will connect to **resilience.io**. Stakeholders will be able to use this system to select and connect data and access indicators to show impacts and performance e.g. related to the Sustainable Development Goals.

For more information please visit resilience.io