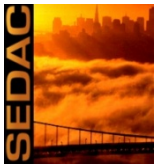
Center for International Earth Science Information Network
EARTH INSTITUTE | COLUMBIA UNIVERSITY



Data Quality Metrics for Socioeconomic Data

Robert R. Downs

rdowns@ciesin.columbia.edu

NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

ESIP 2018 Winter Meeting
Bethesda, Maryland

Session: Information Quality – Progress on Many Fronts
Thursday, 11 January 2018 11:00 a.m. - 12:30 p.m.

Socioeconomic Data Quality Metrics

- Data Quality Assessment Criteria*
  - Scientific quality
  - Quality of data products
  - Quality of data services
  - Quality of data stewardship
- Data Lifecycle Quality Assurance Processes
  - Study conceptualization
  - Study design
  - Data collection
  - Data processing
  - Data Dissemination and Product Development Planning
  - Data Archival Submission
  - Data product or service development
  - Data dissemination
  - Data use

*Consistent with Ramapriyan, Peng, Moroni, and Shie (2017)

Quality Assurance for Creating Socioeconomic Data

- Study Conceptualization
  - Data needs, data management plan, data quality control issues, peer-review plan, available expertise and resources for data quality control
- Study Design
  - Document instrument selection and/or development; data collection protocol; data quality control and processing procedures (data cleaning rules for error detection and handling missing values and outliers)
- Data Collection
  - Adherence to data collection protocol. Document data collection (including describing assumptions, concerns, deviations, and other data quality issues)
- Data Processing
  - Verification of values, ranges, formats, types. Document data processing, data cleaning, error detection, data quality issues, and modifications

Quality Assurance for Socioeconomic Data Products and Services

- Data Dissemination and Product Development Plan
  - Meets community need, methodological review
- Data Archival Submission
  - Complete and correct, documentation, provenance, no confidentiality issues, rights to modify and disseminate as open data
- Data Product or Service Development
  - Quality reviews from relevant disciplines, consistency with development plan
- Data Publication
  - Understandability, verification of documentation on applicability and limitations
- Data Use
  - Reported in peer-reviewed journals of relevant discipline(s)

Common Terminology

- Accuracy: verification of value representations
- Completeness: all files packaged; no missing values
- Confidentiality: absence of Personally Identifiable Information
- Conformity: standards compliant instruments and measures
- Consistency: outliers, not necessarily errors
- Integrity: file and version control
- Validity: inspections for errors; measurement values within acceptable thresholds; rendering of formats

SEDAC Collection Development Focuses on Human Interactions in the Environment

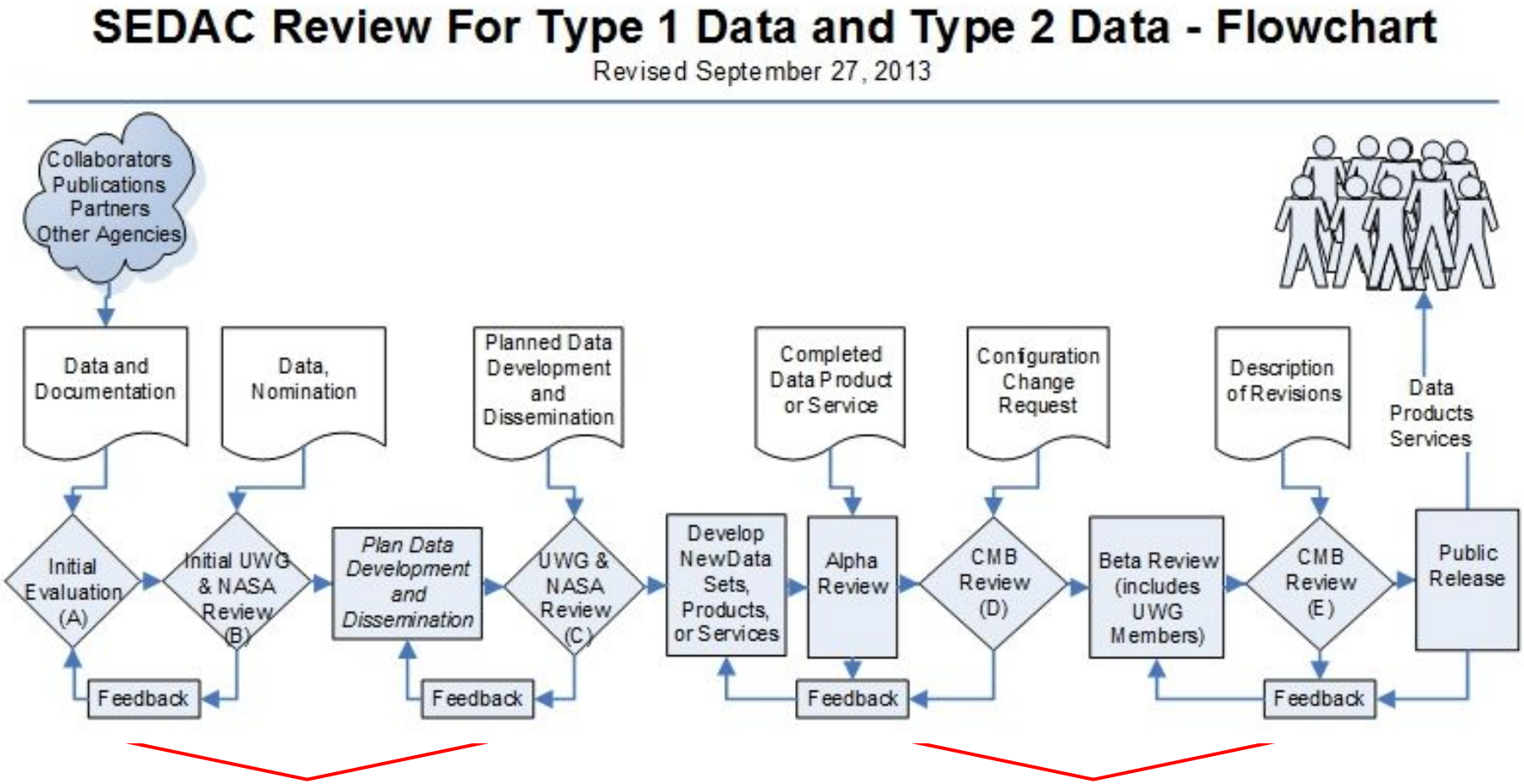
Current Themes

- Agriculture
- Climate
- Conservation
- Governance
- Hazards
- Health
- Infrastructure
- Land Use
- Marine and Coastal
- Population
- Poverty
- Remote Sensing
- Sustainability
- Urban
- Water

Selected Data Collections

- Climate Effects on Food Supply
- Compendium of Environmental Sustainability Indicators
- Energy Infrastructure
- Global Agricultural Lands
- Global Fertilizer and Manure
- Global Roads
- Global Rural-Urban Mapping Project (GRUMP)
- Gridded Population of the World (GPW), v4
- Historical Anthropogenic Sulfur Dioxide Emissions
- India Data Collection
- Indicators of Coastal Water Quality
- Intergovernmental Panel on Climate Change (IPCC)
- Land Use and Land Cover (LULC)
- Millennium Ecosystem Assessment (MA)
- Population Dynamics
- Population Exposure to Natural Disasters
- Satellite-Derived Environmental Indicators
- Spatial Economic Data
- U.S. Census Grids
- Urban Spatial Data

Comprehensive Scientific Data Product Review



Pilot Study of Quality Metrics of SEDAC Data

- Review of data quality information
  - Quality information for 21 recent (2016- 2017) data product releases
  - Limited to data quality information published with the data
- Identified and categorized sources of data quality information
  - 8 categories of sources of data quality information published with data
- Identified terminology used to describe data quality
  - 67 unique terms used for socioeconomic data quality
- Categorized the identified data quality terminology
  - 8 categories of terms used for socioeconomic data quality

Data Quality Sources for Recent SEDAC Releases

- Data files
  - Published files containing data
- Documentation
  - Documentation document published with data
- Methodological Documentation
  - Documentation published with data that addresses specific issues
- Metadata
  - Published with data and referencing documentation and article
- Article
  - Sometimes published with data or linked, depending on rights acquired
- Article Supplemental Information
  - Sometimes published with data or linked, depending on rights acquired
- Report
  - Published with data
- FAQ
  - Based on questions answered and published with data

67 Unique Socioeconomic Data Quality Terms

accuracy	disaggregation	problems
adjustments	errors	projections
aggregation	estimation	proportions
alternative sources	evaluation	quality assurance
applicable use	exceptions	quality checking
appropriate	exclusions	quality control
appropriate use	filters	quality issues
assumptions	gaps	quality problems
backcast	implications for use	rationale
bias	improvements	recommended use
caveats	inappropriate use	references on methods
coarsening	incompleteness	small errors
comission errors	inconsistencies	sources quality
comparison	inflation	substitutions
confounding factors	known issues	suitability for use
constraints on use	limitations	thresholds
corrections	log of changes by version	unavailable data
currency	matching	uncertainty
data challenges	missing values	undercounts
data quality indicators	no data	usage issues
data sources	omissions	validation
deviations	other factors	
difficulties	possible errors	

Aggregated Socioeconomic Data Quality Terms - 1

- Caveats
  - Accuracy, assumptions, appropriate, bias, caveats, comparison, confounding factors, data challenges, deviations, difficulties, evaluation, exceptions, exclusions, inconsistencies, known issues, inflation, limitations, other factors, problems, quality issues, quality problems, rationale, uncertainty, undercounts
- Correction
  - Adjustments, corrections, estimation, improvements, substitutions
- Errors
  - Comission errors, errors, possible errors, small errors
- Missing data
  - Gaps, incompleteness, missing values, no data, omissions, unavailable data

Aggregated Socioeconomic Data Quality Terms - 2

- Modification
  - Aggregation, backcast, coarsening, currency, disaggregation, filters, matching, projections, proportions, thresholds
- Use
  - Applicable use, appropriate use, constraints on use, implications for use, inappropriate use, recommended use, suitability for use, usage issues
- Quality Control
  - Data quality indicators, log of changes by version, quality assurance, quality checking, quality control, references on methods, validation
- Sources
  - Alternative sources, data sources, sources' quality