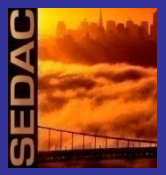
Designing A Data Commons for Sustainability Science: Lessons Learned from a World Data Center

Alex de Sherbinin
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

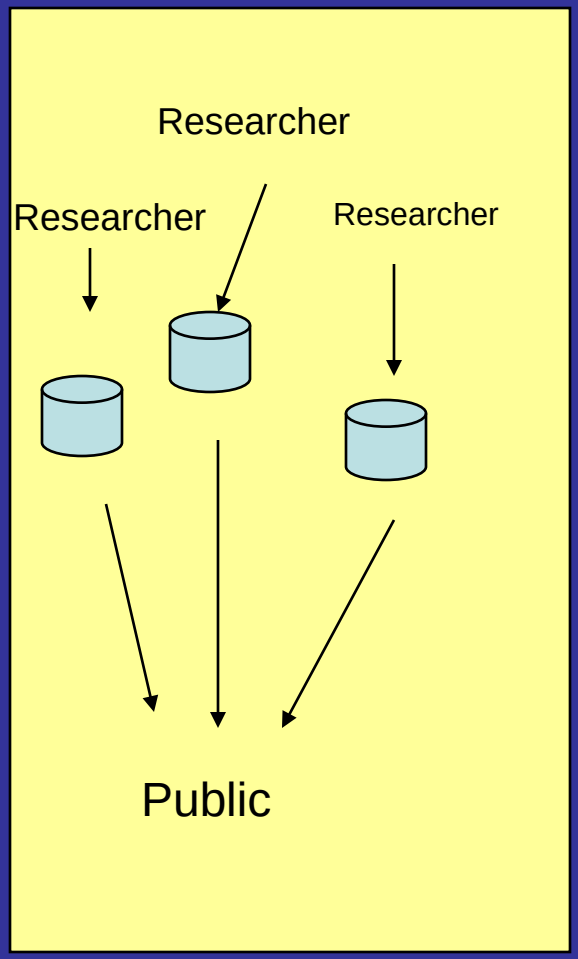
Presentation to the International Workshop on Open Access to Scientific Literature and other Digital Scientific Information Resources in Central America and the Caribbean: Focus on Education and Health for Sustainable Development
3-4 September 2008
Havana, Cuba

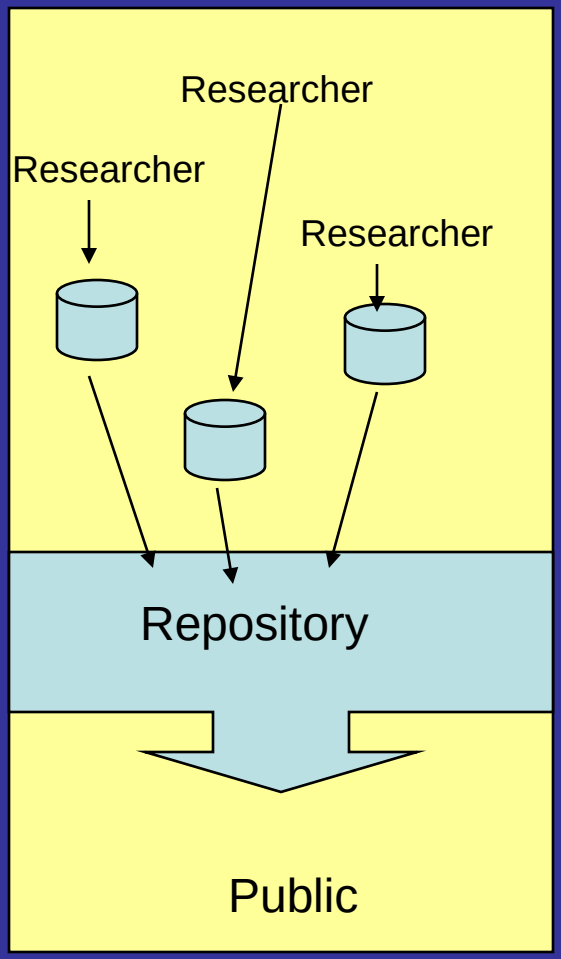


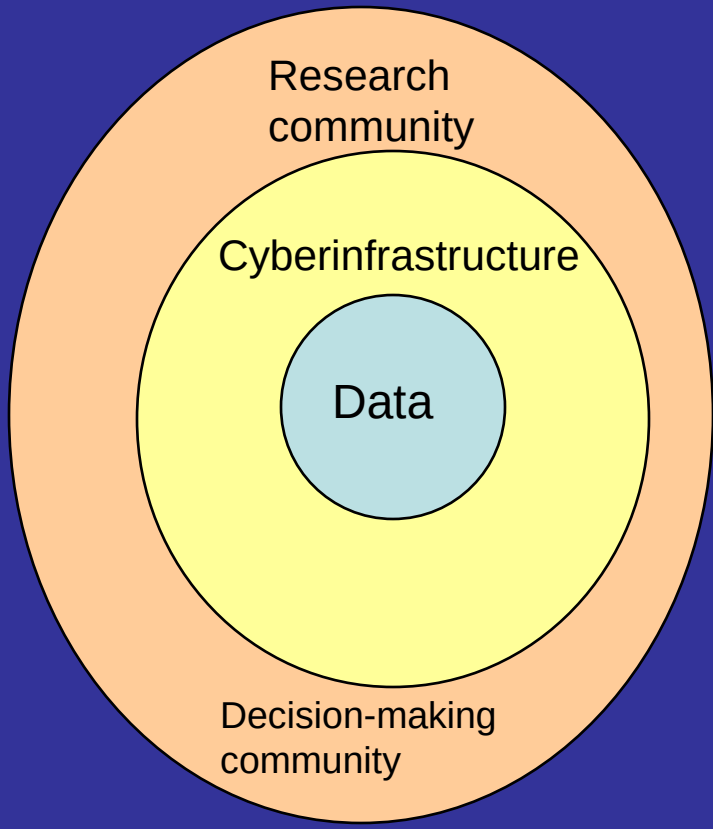
Overview

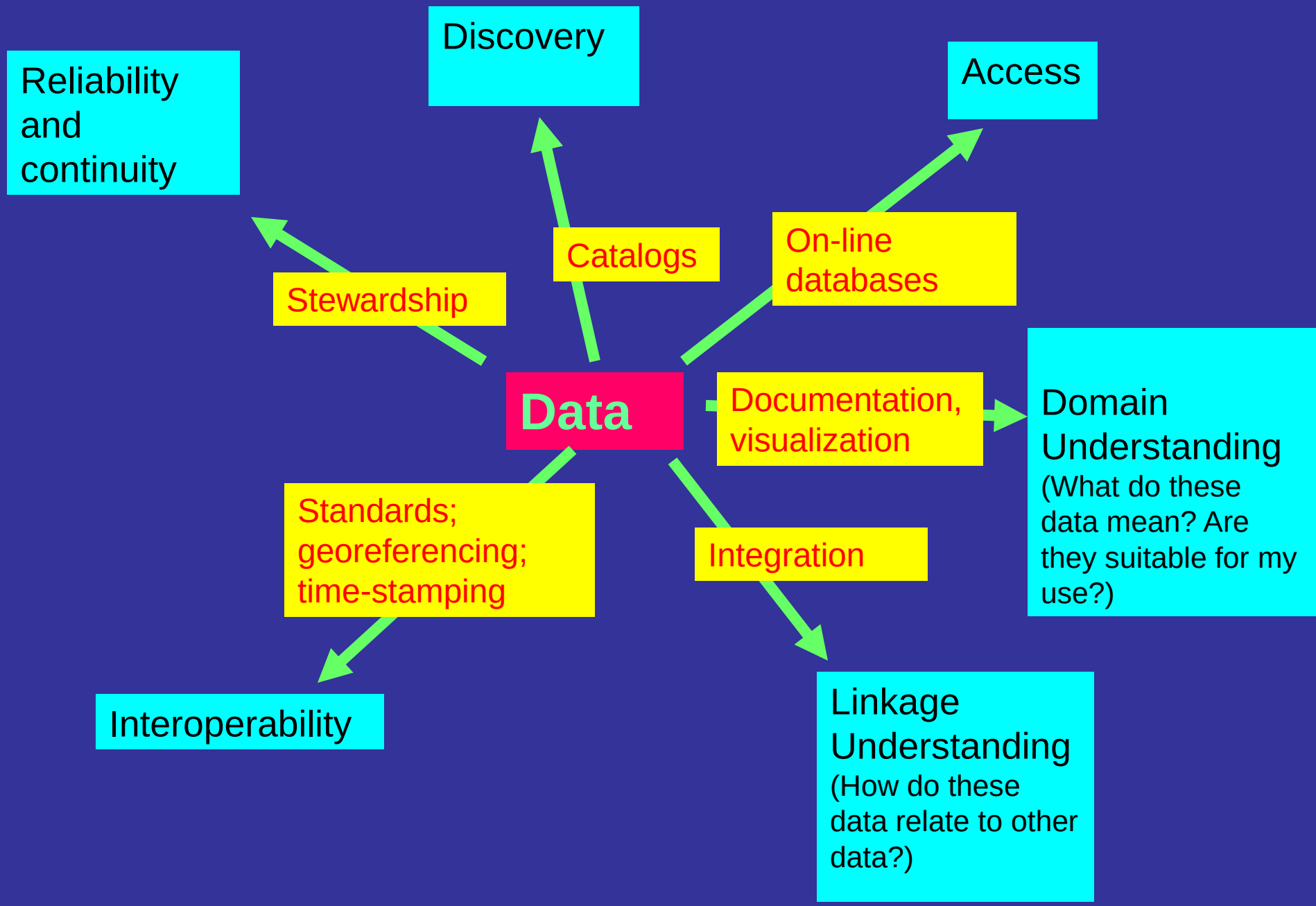
- Conceptual approach: the modern data center
- The World Data Center for Human Interactions in the Environment
- The case for open access (OA)
  - International principles supporting OA
  - Issues around OA
  - The growing movement towards OA

Evolution of the Data Provider Model





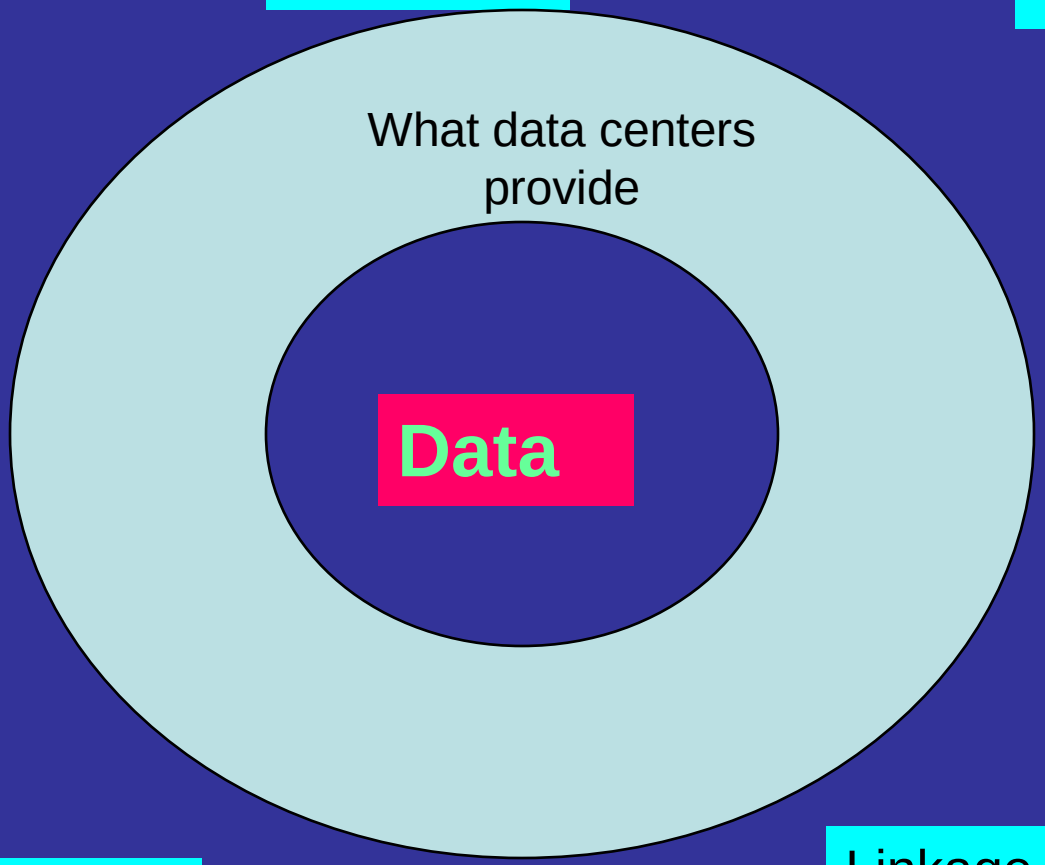



Reliability and continuity
Discovery
Access
Stewardship
Catalogs
On-line databases
Data
Documentation, visualization
Domain Understanding
(What do these data mean? Are they suitable for my use?)
Standards; georeferencing; time-stamping
Integration
Interoperability
Linkage Understanding
(How do these data relate to other data?)

Discovery

Access

Reliability and continuity



Domain Understanding
(What do these data mean? Are they suitable for my use?)

Interoperability

Linkage Understanding
(How do these data relate to other data?)

CIESIN's World Data Center

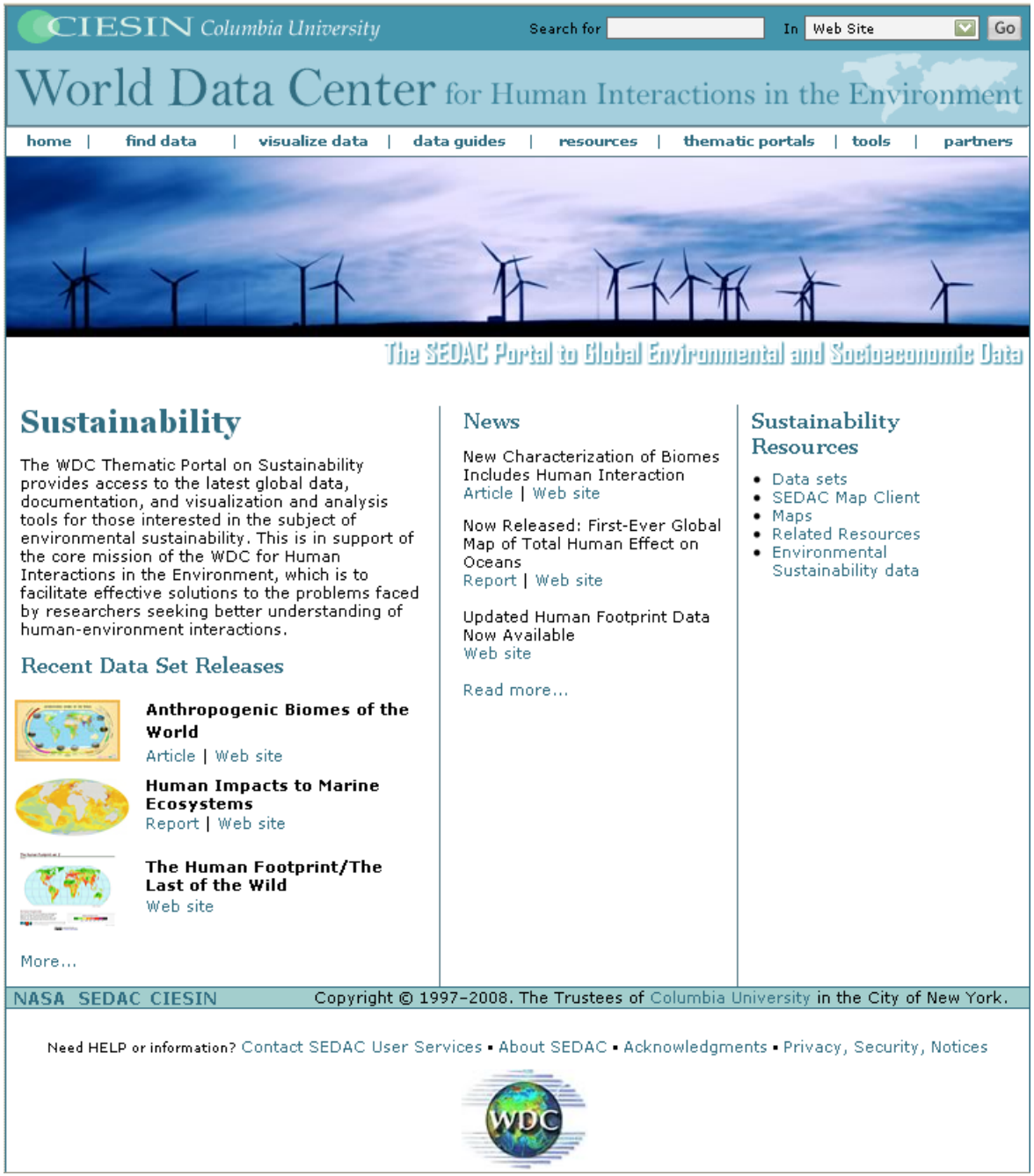
- CIESIN is a data and research center that was established in Michigan in 1989 and which joined The Earth Institute at Columbia University in 1998
- CIESIN runs the Socioeconomic Data and Applications Center (SEDAC), which is one of eight Distributed Active Archive Centers (DAACs) run by NASA's Earth Observing System Data and Information System (EOSDIS)
- CIESIN's World Data Center (WDC) for Human Interactions in the Environment is part of the International Council of Science (ICSU) WDC system
  - A "virtual" data center
  - Supported by SEDAC
  - Focus on *global* scale *geospatial* data

CIESIN's World Data Center

Climate
Conservation
Governance
Hazards
Health
Population
Poverty
Sustainability

Discovery
Domain Understanding

Sustainability Thematic Portal



Discovery

Catalog

Utilizes opensource Geonetwork by FAO

Distributed search using Z39.50 protocol

FGDC metadata

Short profiles with thumbnails

Full metadata records available

All data freely available for download

Domain Understanding

Map Gallery

CIESIN Columbia University Search for In Web Site Go

World Data Center for Human Interactions in the Environment

home | find data | visualize data | data guides | resources | thematic portals | tools | partners

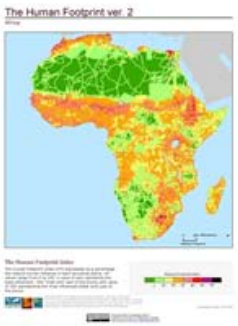
The SEDAC Portal to Global Environmental and Socioeconomic Data

View: All Maps | Climate | Conservation | Governance | Hazards | Health | Population | Poverty | Sustainability

Conservation

Human Footprint

Africa

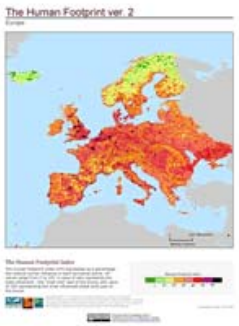


download: PDF | JPG

Asia

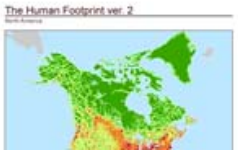
download: PDF | JPG

Europe



download: PDF | JPG

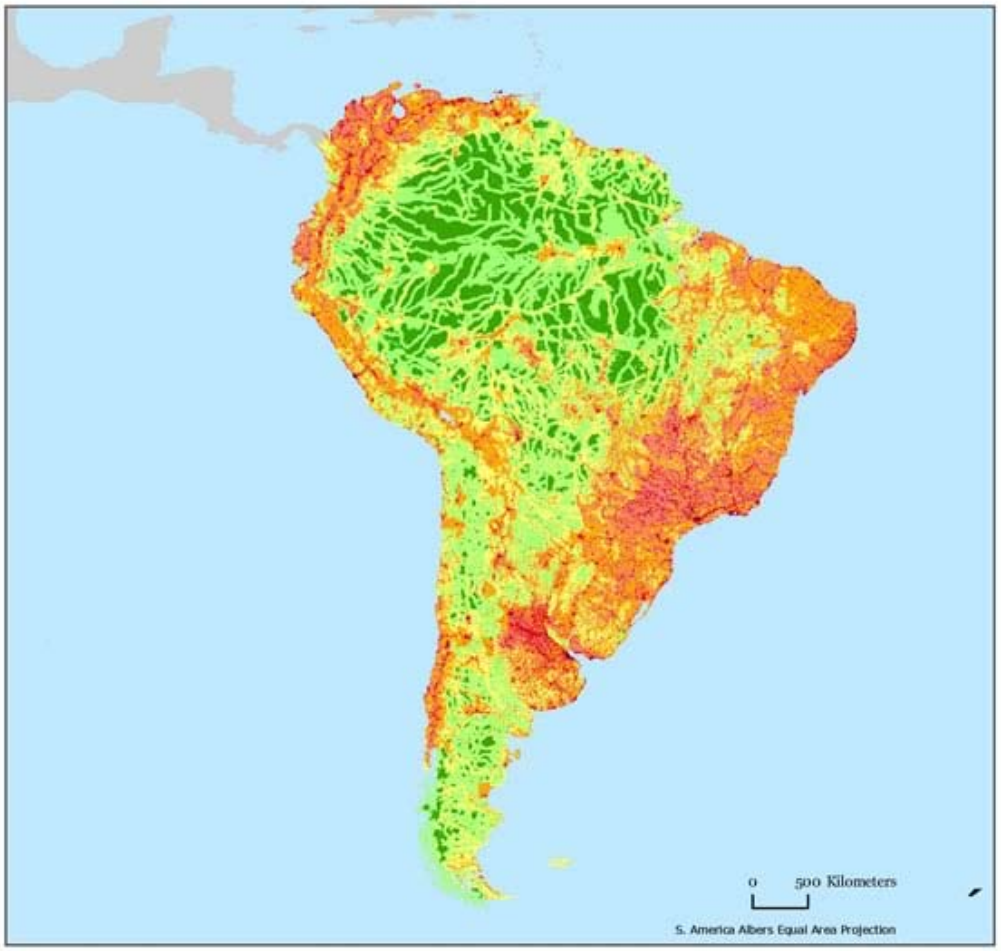
North America



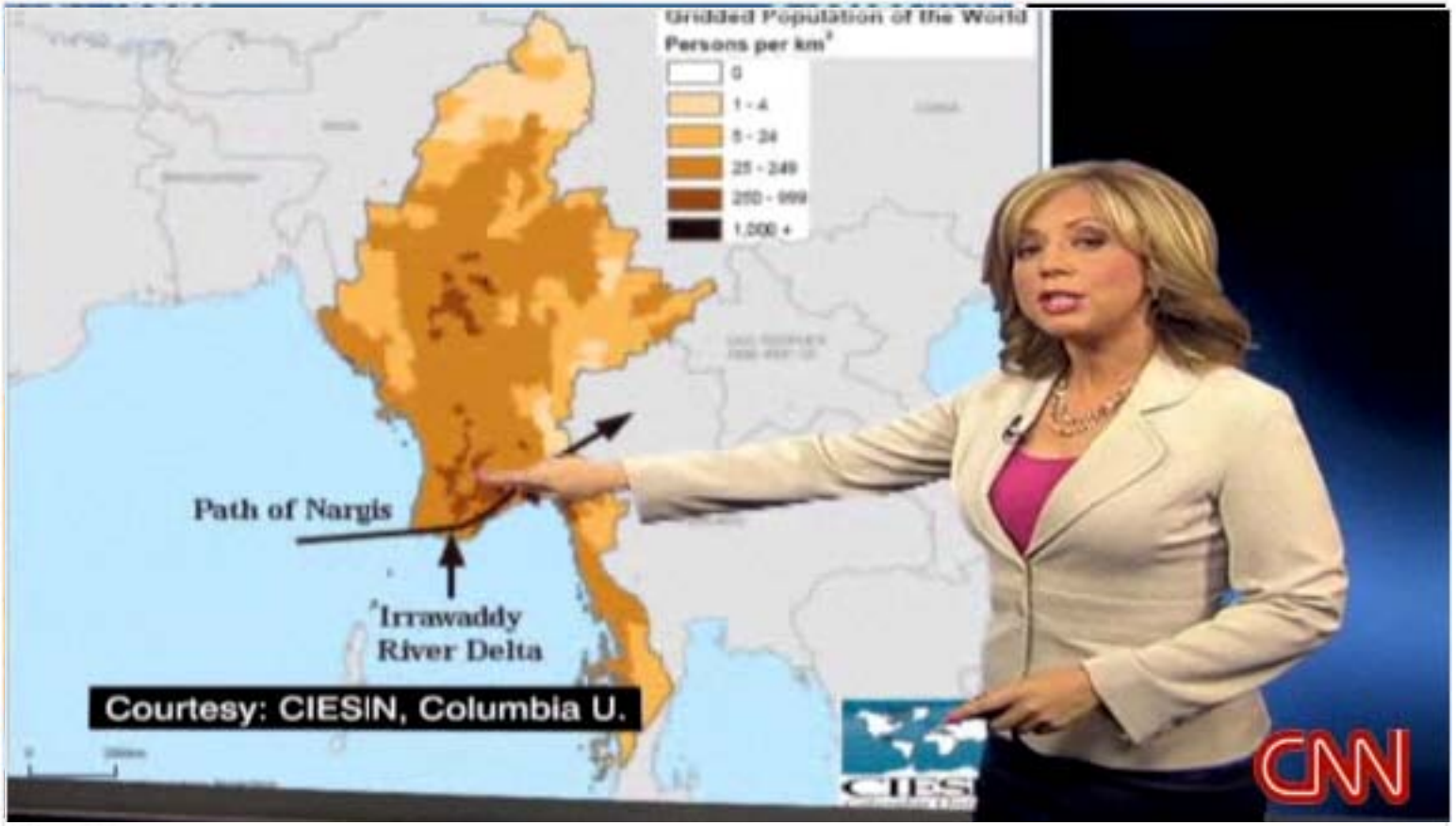
Oceania

South America

Sample Map



Maps are distributed using a Creative Commons license.



1:13 / 2:29 EMAIL FULL SCREEN

Cyclone surges explained 2:29

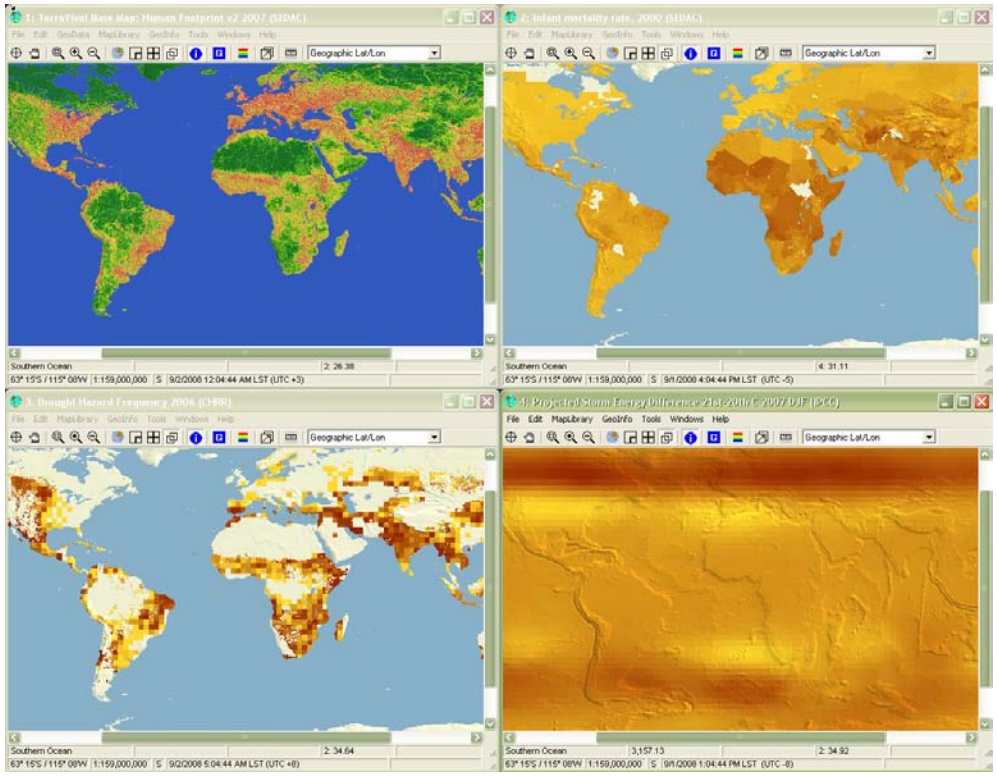
CNN's Bonnie Schneider explains what caused the surges that left areas of Myanmar submerged.

- Myanmar cyclone death toll exceeds 22,000
- Impact Your World - Special Reports from CNN.com

Source: CNN
Added On May 6, 2008

Domain Understanding
Linkage Understanding

2008 TerraViva! SEDAC



Stand-alone data exploration tool

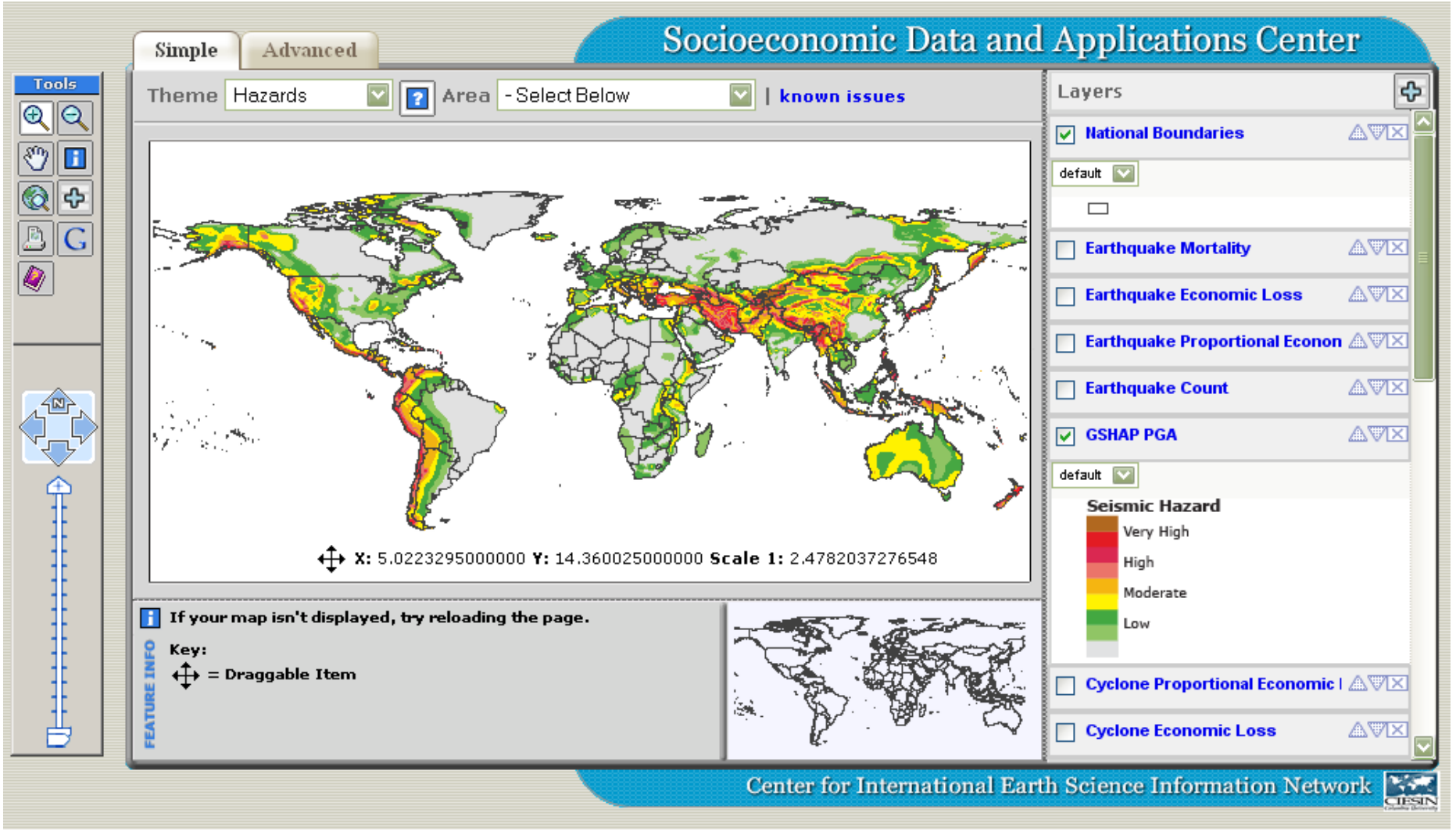
Visualization and interpretation

Carry out simple spatial analyses

http://sedac.ciesin.columbia.edu/terraVivaUserWeb/

Interoperability
Domain Understanding
Linkage Understanding

OGC-compliant Web Mapping



Helping Users Make Wise Choices

Traditional Documentation not enough

Multi-faceted approach required

- **Comparative Guides**
- **Visualizations**
- **Common Pitfalls**
- **Examples**
- **Citations**

SEDAC Citation Index

Browse references citing SEDAC data, applications, publications, and projects:

- Archive of Census Related Products (ACRP)
- Central American Vegetation / Land Cover Classification and Conservation Status Data (PROARCA)
- China Dimensions Data Collection
- Crop Climate Datasets
- Demographic Data Viewer (DDViewer)
- Environmental Treaties and Resource Indicators (ENTRI)
- Environmental Performance Index (EPI)
- Environmental Sustainability Index (ESI)
- Geographic Correspondence Engine (Geocorr)
- Georeference Population Data Sets of Mexico
- Global Rural-Urban Mapping Project (GRUMP)
- Gridded Population of the World (GPW)
- Human Footprint and Last of the Wild Datasets
- IPCC Socioeconomic Data Distribution Centre (DDC)
- Model Visualization and Analysis (MVA)
- Ozone and Human Health
- Population-Environment Research Network (PERN)
- Public Use Microdata Samples (PUMS)
- Remote Sensing and Environmental Treaties Workshop
- Thematic Guides

SEND US YOUR CITATIONS!

If you know of any publication citing data or information obtained from SEDAC, please **send us** the citation!

CITATION GUIDELINES

Guide to formatting citations.

http://sedac.ciesin.org/citations/CitationIndex.html

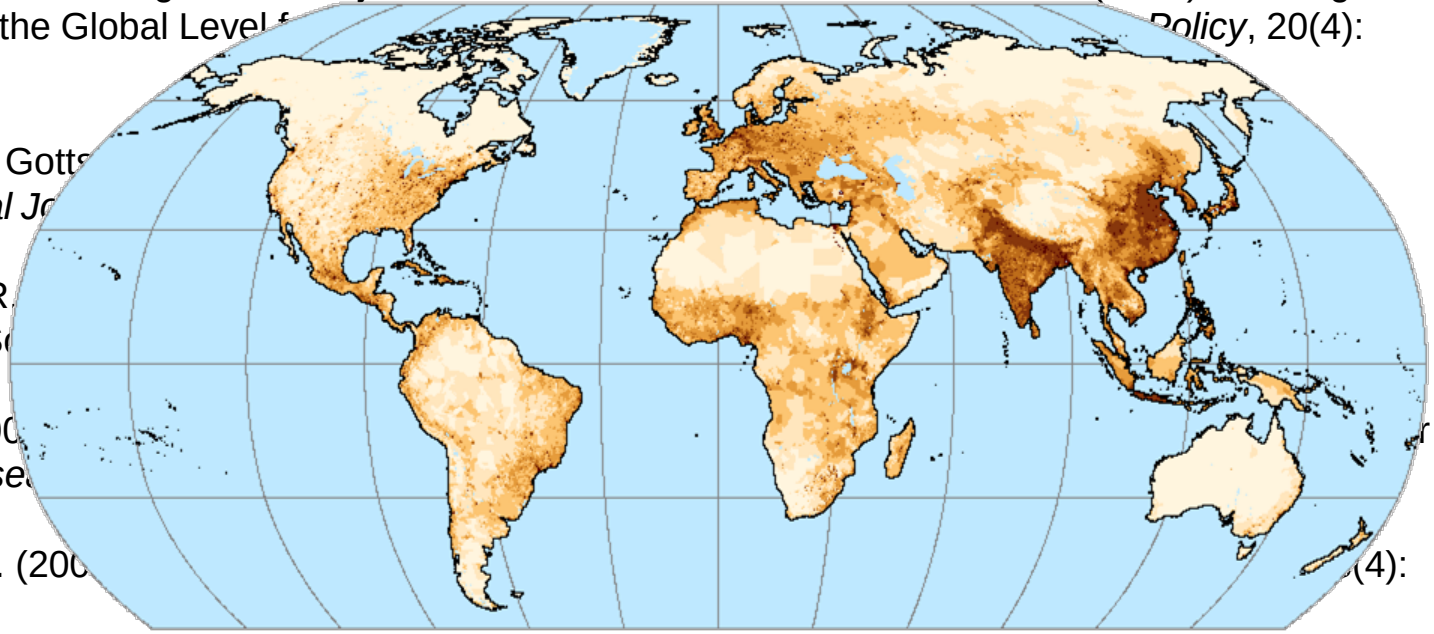
Poor Livestock Keepers at the Global Level f… *Policy*, 20(4): 311-322.

Tobler, W., Deichmann, U., Gotts… Quadrilaterals. *International Jo…*

van Lieshout, M., Kovats, R.… of the SRES Climate and S…

Vassolo, S. and Döll, P. (200… Use. *Water Resources Rese…*

Verburg, P.H., and Chen, Y. (200… (4): 369-385.



Viviroli, D. and Weingartner, R. (2004) The Hydrological Significance of Mountains: from Regional to Global Scale. *Hydrology and Earth System Science,* 8(6): 1016-1029.

Vorosmarty, C.J., Green, P., Salisbury, J., and Lammers, R.B. (2000) Global Water Resources: Vulnerability From Climate Change Acid Population Growth. *Science*, 289(5477): 284-288.

Vorosmarty, C.J., and Sahagian, D. (2000) Anthropogenic Disturbance of the Terrestrial Water Cycle. *Bioscience*, 50(9): 753-765.

White, M.A., Hoffman, F., Hargrove, W.W., and Nemani, R.R. (2005) A Global Framework for Monitoring Phenological Responses to Climate Change. *Geophysical Research Letters,* 32(L04705): 4pp.

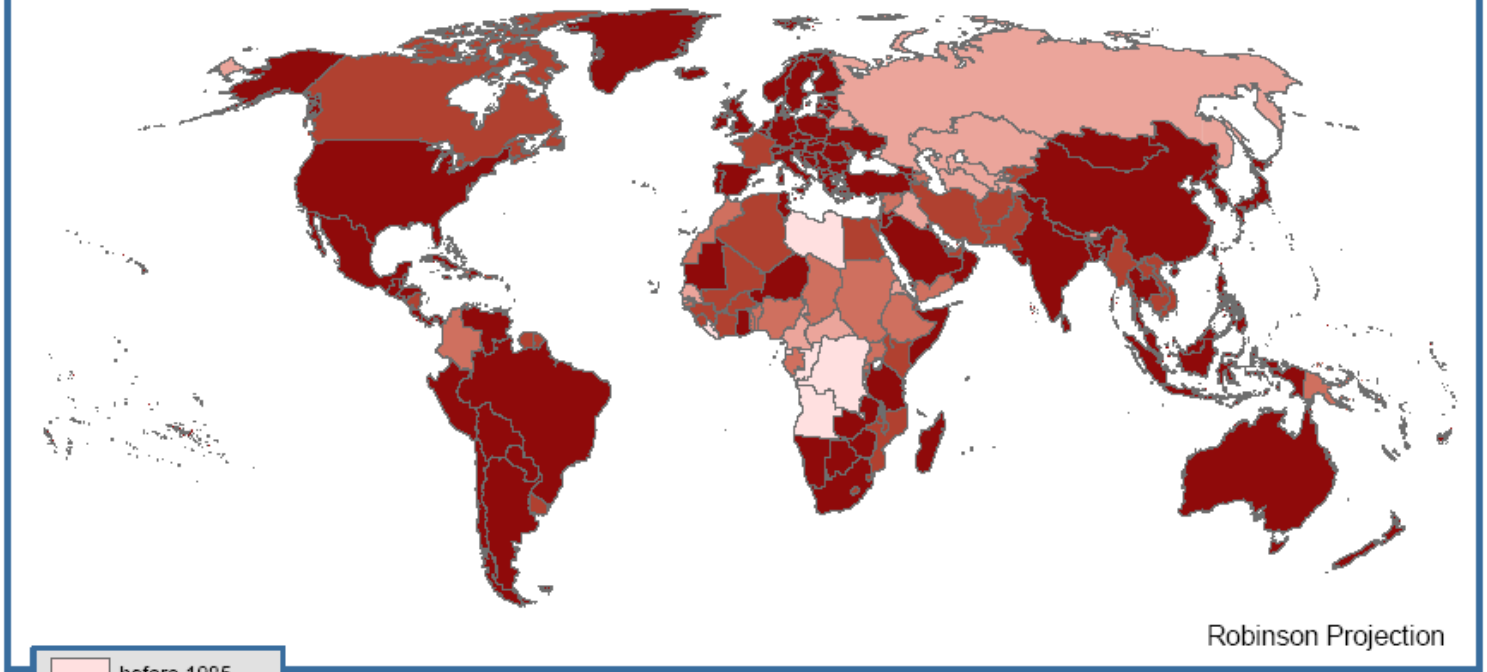
White, M.A., Nemani, R.R., Thornton, P.E., and Running, S.W. (2002) Satellite Evidence of Phenological Differences Between Urbanized and Rural Areas of the Eastern United States Deciduous Broadleaf Forest. *Ecosystems*, 5(3): 260-273.

Wilson, S.J., Steenhuisen, F., Pacnya, J.M., and Pacnya, E.G. (2006) Mapping the Spatial Distribution of Global

Information on data quality is critical to judging appropriate uses

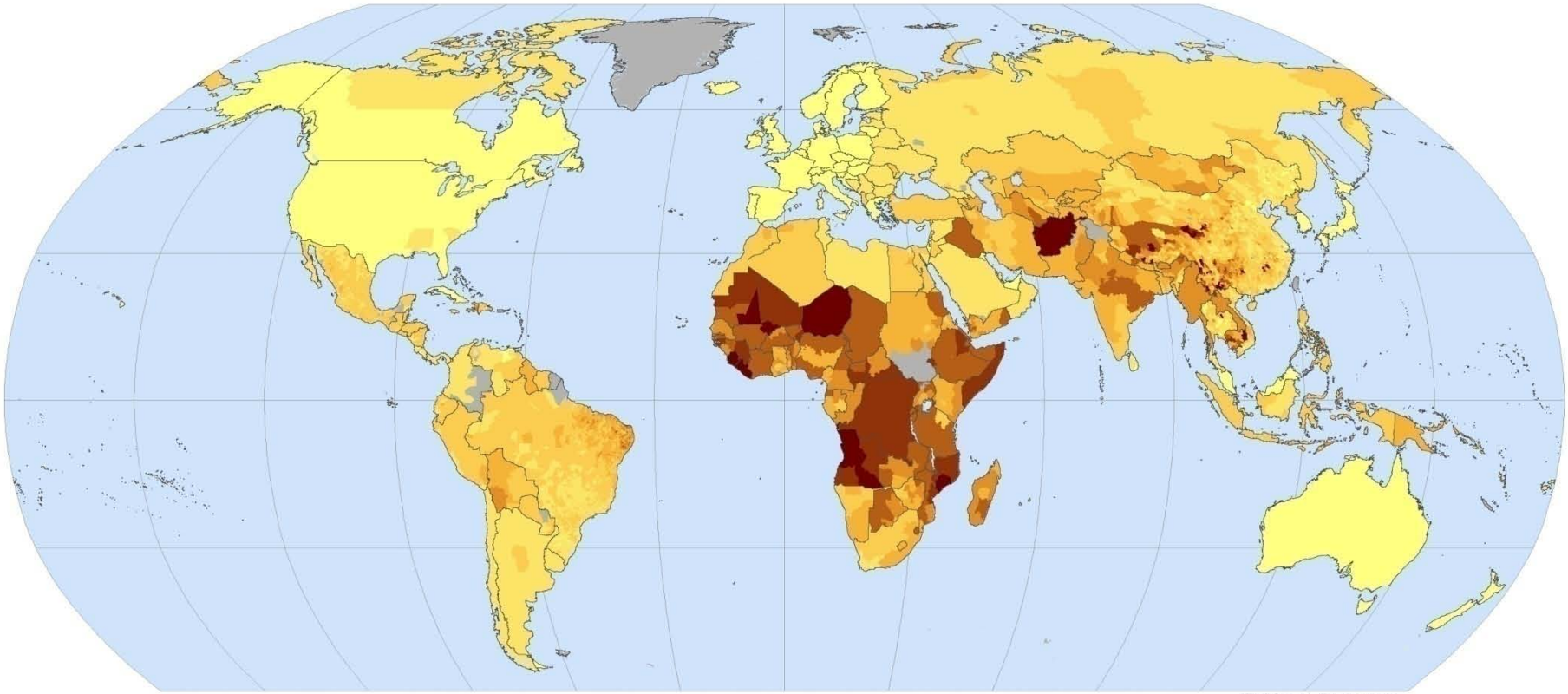
Figure 2. Number of administrative units per country [v3]

Figure 4. Most recent population data year GPW



Sample Analysis

- Data documentation helps to determine fitness of use for given analyses
- How is the global distribution of poverty associated with the distribution of natural hazards?



Robinson Projection

Measures of Poverty

Infant Mortality Rates

By Subnational Administrative Unit

Subnational mortality rates are adjusted to 2000 using national trend data.
Original data for 96% of countries are from 1995 or later. All data are from 1990 or later.

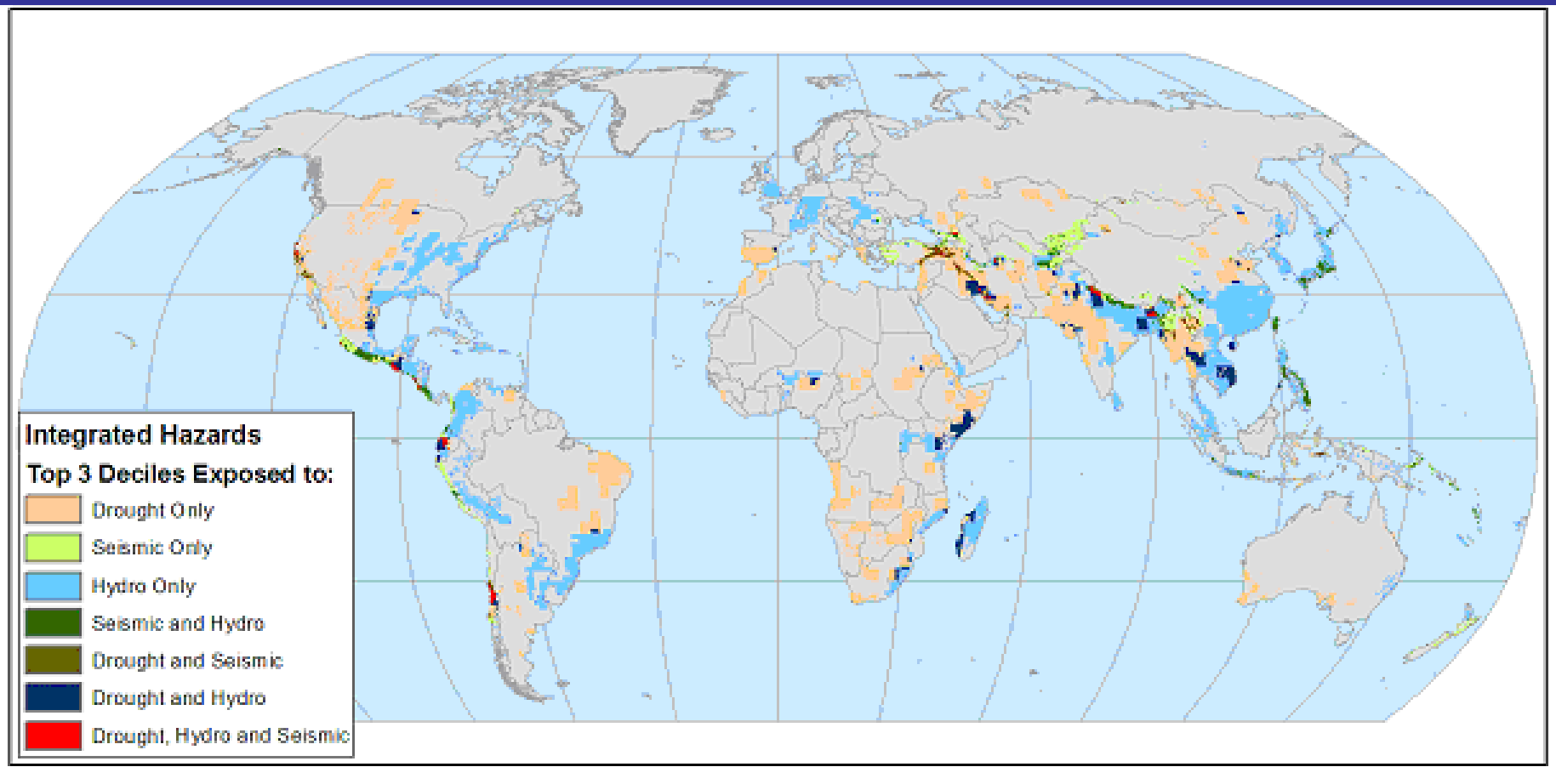
Infant mortality rate, 2000
(deaths per 1000 live births)

- < 9.1
- 9.1 - 25.0
- 25.1 - 50.0
- 50.1 - 75.0
- 75.1 - 100.0
- 100.1 - 125.0
- 125 - 150.0
- 150.0 <
- no data

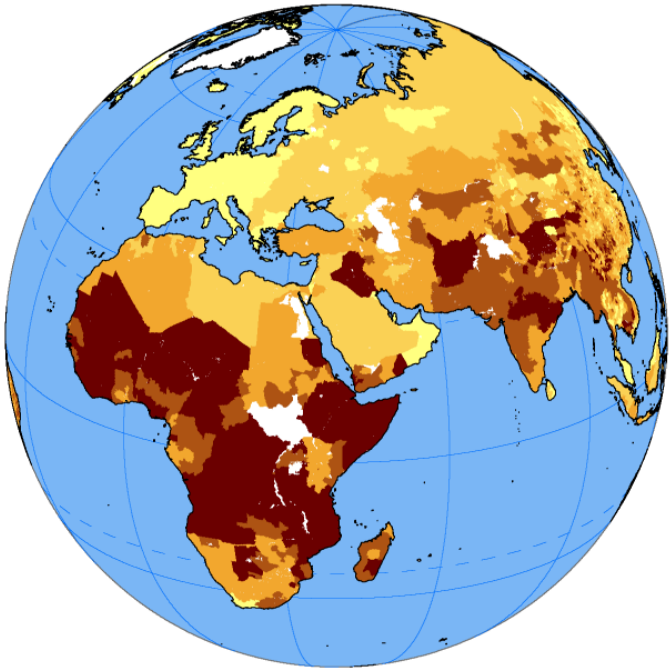
National Boundary

Subnational boundaries have been removed from countries for clarity.

Exposure to Multiple Natural Hazards



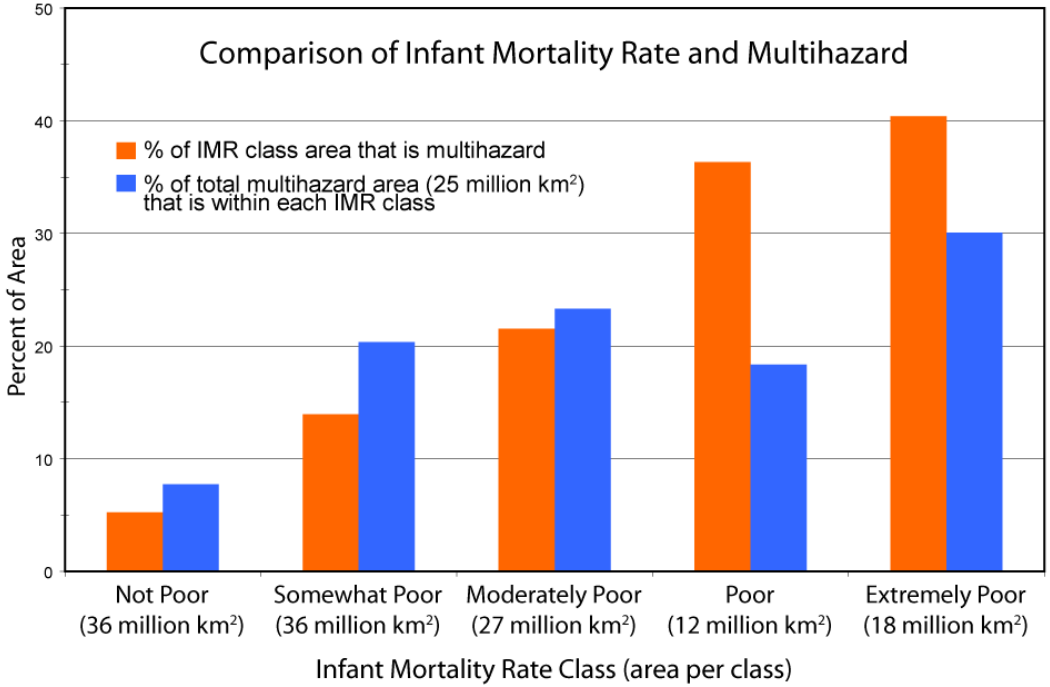
Seismic hazards include earthquakes and volcanoes; *hydrological hazards* include floods, cyclones, and landslides

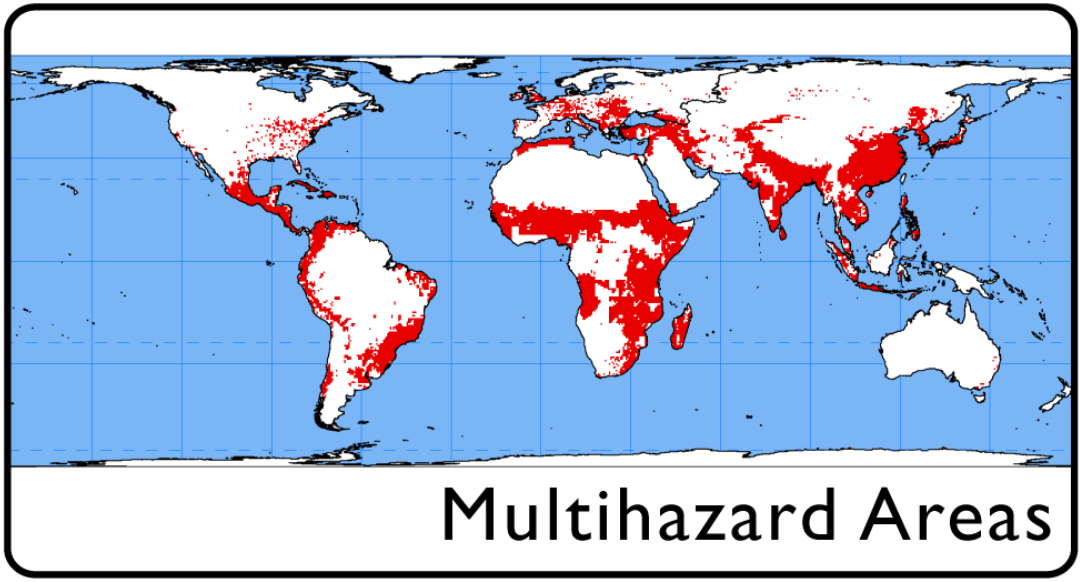


Infant Mortality Rates

(per 1,000)

Not Poor	1.9 - 15
Somewhat Poor	15.1 - 32
Moderately Poor	32.1 - 65
Poor	65.1 - 100
Extremely Poor	> 100





Multihazard Areas

Helping users make wise choices is a community-building and community-strengthening task

Need for Long Term Archiving

- Much information is now “born digital”
- Data are being lost at an alarming rate – either *not preserved at all* or *unreadable* due to changing computer systems
- Important to preserve the discovery and access capabilities – not just back up the data on offline tapes

LTA Architecture Review

- CIESIN reviewed commercial and open source systems to facilitate ingest, preservation, and access
  - Digital asset management systems
  - Electronic records management systems
  - Document management systems
  - Digital repository systems
- We decided to focus on open source approaches to avoid proprietary dependencies
  - Dspace
  - Eprints
  - Fedora
  - Greenstone
- We selected the Flexible Extensible Digital Object Repository Architecture (Fedora)
  - Developed by Cornell University and the University of Virginia
  - Modular approach to facilitate enhancement
  - Active user community of developers and implementers
- We purchased VITAL with Fedora from VTLS

Source: Downs, R. and Chen, R. (2008). Implementing a Digital Repository for the Preservation of Interdisciplinary Data. Presented to the International Association for Social Science Information Services & Technology (IASSIST) 2008 Conference, Palo Alto, CA.

The Case for Open Acess (1)

- The US set a benchmark by dedicating much federal data to the public domain
  - Government funded data are often required to be distributed freely, even though creators may retain copyright
- Europe is gradually following suit
- Some developing countries are remarkably far along

The Case for Open Access (2)

- OECD Principles and Guidelines for Access to Research Data from Public Funding.
- GEOSS data sharing principles:
    - There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation.
    - All shared data, metadata, and products will be made available with minimum time delay and at minimum cost.
    - All shared data, metadata, and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

Regional Examples

- Inter-American Institute's Data and Information Service (IAI-DIS)
- NASA's SERVIR for hazards & meteorological data
- Centro Internacional de Agricultura Tropical (CIAT)
- Comisión Económica para América Latina y el Caribe (CEPAL) – databases and statistics
- Spatial Data Infrastructure (SDI):
  - Central American Geographic Information Project (PROCIG)
  - Brazilian Institute of Geography and Statistics (IBGE)
  - China-Brazil satellite data (CBERS)

Barriers to Open Access

- Researchers may not want to release data for fear of findings being “scooped”
- Data limitations are too costly/time consuming to correct and would reflect poorly on the researcher
- The transaction costs of data documentation and metadata creation may be too high
- Distributing data freely will undermine institutional objectives:
    - Cost recovery
    - Power
    - Quid pro quo

OA Data Creates Opportunities

- Creates opportunities for new scientific discovery
- Creates opportunities for sustainable development
  - By providing data that solve real problems
  - By creating economic opportunities for small or medium sized enterprises
- Creates opportunities for international exchange and collaboration

Gracias!

CIESIN World Data Center

http://sedac.ciesin.columbia.edu/wdc/