Data Science Report # 453

Commissioned by the University of Toronto

Linked Data Mash-Up Report: Food Security and Temperature Change

Written by Matt Baricevic

Section A – Intended Purpose

The scientific body known as the Intergovernmental Panel on Climate Change (IPCC) has stated in its 2007 report that the warming of the climate is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising emissions of greenhouse gases due to anthropogenic activities. (Toulmin, 2009. *Climate Change in Africa*, p. 16.)

The consensus amongst the IPCC is that climate change will increase average temperatures. The already arduous task of attaining food security will, according to projected models, become increasingly compounded through the looming presence of climate change. In particular, countries in subtropical regions will suffer agricultural losses.

It follows that the purpose of this project is to demonstrate - through metadata - the projected global temperature change and the negative physical impacts on cereal yields and dryland expansion stemming from climate change. The theoretical underpinning of the project, and the accompanying metadata modelling, may be summed up in accordance with the following variables:

Organizational Logic:

Theory	Independent Variable	Dependent Variables
Climate change reinforces global inequalities	Climate Change (in predicted temperature change)	Cereal Yields
		Drylands – Percent Total

Section B – Community and Domain Targeted by the Mash-Up

The potential domain, targeted by this mash-up is primarily agriculture producers, Non Governmental Organizations (NGO), human rights organizations, social justice groups, and the general population. The consensus amongst the scientific community is that Africa remains the most exposed and vulnerable continent in relation to climate change and specifically increases in temperature.

Moreover, the absence of academic literature on this topic served as an additional catalyst for this study. Few studies have examined the effects of climate change, through the means of metadata, on the most vulnerable populations in the least developed countries - 33 out of 48 being in Africa. (Fourth United Nations Conference on Least Developed Countries, 2011, http://www.un.org/wcm/content/site/ldc/home/Background/quick_facts#11128)

After initial brief metadata modeling attempts, I quickly arrived at the conculsion that Africa was disproportionately affected by climate change as I examined various datasets. Although my method of examining data may differ, I am not alone in this sentiment. The vice president of sustainable development for the World Bank has stated that the yield of food in Africa could decrease by nearly fifty percent in the coming twenty years years due to the effects of climate change. (Cribb, *The Coming Famine*, p. 138).

The most vulnerable populations are reliant on climate sensitive industries such as agriculture for both their livelihood and Gross Domestic Product (GDP). They remain exposed and vulnerable to the effects of temperature change. As a result, I chose to plot the significant

dependent variable of temperature change as my first map, including projected maximal temperature increases by 2060 in relation to the most vulnerable and least resilient nations.

Section C - Structured Datasets Used to Support the Mash-Up

The source of this microdata on temperature change came from the United Nations Development Programme (UNDP) and was retrieved from the University of Oxford School of Geography and the Environment (United Nations Development Program, Climate Change Country Profiles, 2012, http://www.geog.ox.ac.uk/research/climate/projects/undp-cp/). The metadata for temperature change consisted of two descriptive elements including the country name and the average temperature change projected for 2060.

The metadata for the cereal yield and percent of total drylands were provided by the Food and Agricultural Organization of the United Nations Environment Programme (UNEP): Environment for Development, and accessed through the Environmental Data Explorer. (United Nation Environment Programme via the FAO (metadata) http://geodata.grid.unep.ch/)

To contextualize these two data sets a brief excerpt from their abstracts is provided to define their purposes. The percent of total dryland area per nation is described as " areas with a potential hazard of desertification." They "are defined as arid, semi-arid and dry sub-humid zones, or areas with lengths of growing periods, or areas with lengths of growing periods of 1-179 days." (Drylands – Percent of Total Area (Metadata), Food and Agriculture Organization, 2000, http://geodata.grid.unep.ch/mod_metadata/metadata.php).

The global cereals yield per nation is defined as cereals that "also include other cereals such as mixed grains and buckwheat." The data set concerns not the harvested product but, rather, is "obtained by dividing the data stored under the production element by those recorded

under element area "harvest" and are measured in hectogrammes i.e. (100 grammes) per hectare (HG/HA" (Cereals - Yield, FAO, 2010, http://geodata.grid.unep.ch/mod_metadata/metadata. .php).

Additionally, unlike the dryland metadata, which offers no supplemental information, the cereal yields inform one that the pre-1992 data from the former USSR, in addition to the former nations of Czechoslovakia and Yugoslavia, are "based on data available after 1992"; the pre-2000 data from Belgium includes Luxembourg; the data for China includes Taiwan; the data for "Eritrea from before 1993 is shown under Ethiopia"; and that for pre-2006 Serbia includes Montenegro (Cereals – Yields (Metadata), FAO, 2010 http://geodata.grid.unep.ch/mod_metadata /metadata.php).

The metadata files for drylands and cereal yields contain the same elements and attributes with varying values. There are nine descriptive metadata elements with a total of twenty-seven different attributes, as well as one administrative element that contains three attributes.

Among the descriptive elements is citation information with the following attributes: title, publication year, data type, geographical data set type file, presentation form, geographical units, filename, an abstract, the character set, language status, and update frequency.

The topic category element contains geographical theme and data category, which helps one to identify this metadata as pertaining to descriptive information in relation to land, agricultural use and production. The keywords element contains the GEMET theme and a free form keyword description - these, in turn, distinguish between the description of dryland and desertification in the first data set and the agricultural production area, and food and soil information in the second data set regarding cereal yields.

The temporal extent element in both the datasets from the FAO specify the time covered by the data. Oddly, in the first dataset regarding drylands the values for beginning and end date are both "9999" (Drylands (Metadata), FAO, 2000, http://geodata.grid.unep.ch/mod_metadata/ /metadata.php). Whereas this element for cereal yields clearly delimits the start date as 1961 and the end date as 2009 (Cereal Yields(Metadata), FAO, 2010, http://geodata.grid.unep.ch/mod_met data/ metadata.php). Whether the dryland data set was a typo or is supposed to indicate some other field remains unknown.

The geographical extent element describes the geographic coverage of the microdata, and allows description in textual terms, such as 'global', or for more tightly defined extents the use of geographic coordinates, which was useful as this mash-up intended to look at the global supply of cereals per nation and the percent of drylands per nation. However, as time progressed, I became particularly interested in the continent of Africa because the information for this continent indicated that it was vulnerable to food shortages. These first four elements make up the core of the information required to understand the data sets; the last four elements pertain largely to describing the people and the organization responsible for the collection of the data.

The last four descriptive elements are the following: point of contact, designating the role, person, organisation, and email address needed for contact information or inquiries in relation to the data. There is also a resource constraint element, with the attribute of use constraint as public. Additionally, the contact information element specifies the online resource from which the data was retrieved - in this case it was the UNEP geodata resource located at http://geodata.grid.unep.ch and the URL original source which is the FAO address at http://faostat.fao.org/default.aspx. The final descriptive element is the coordinate reference information for the data.

Unlike the descriptive metadata, the administrative metadata is very limited in nature, containing a mere two elements and six attributes. The first is simply general information about the metadata, its language, character set, and the date. The second author element specifies the attributes of the individual responsible for creating the metadata record, their email, and the ISO.

All the data selected for this mash-up is derived from linked open data sources because they were most readily available. However, in some instances this meant that the data selected did not necessarily conform to the software used to create the visualisations. For example, I had originally hoped to overlay data on temperature change from the World Bank with that of drylands and cereal yields from the FAO. But I soon arrived at the conclusion that the data from the World Bank was far too detailed and resources of the university were too limited. In light of these considerations, the data was nearly impossible to plot using "Indiemapper" software as it blurred the underlying map, making it impossible to view continents (Indiemapper, http://indiemapper.com/).

Finding compatible file formats and software that would allow for mapping of the desired datasets was ultimately what both limited and defined my choice of data sets. I had originally tried Tableau software, however, this software proved more effective at making graphs and charts than maps (Tableau, http://www.tableausoftware.com/). I later tried Google Fusion, which would have been a possible choice had I only wanted to plot specific points on a map (Google Fusion, http://www.google.com/fusiontables/Home/). However, this required one to define the longitude and latitude of the variables, resulting in a significant amount of data modification on my behalf - especially considering that I was looking at global trends not changes in a specific location.

Finally, I found that while the World Bank's data was my initial choice for temperature information, it was unusable. Thus, I selected the data from the UNDP relating to Climate Change for Countries through the portal of the University of Oxford School of Geography and Environment to visualize the projected data. It must be noted that although both the FAO (UNEP) and the UNDP are part of the United Nations, they operate as distinct statistical divisions and produce and publish data independently of each other. Thus, the selection process for the datasets was constantly evolving: as my choice for visualisation software changed I discovered what types of data were effectively mappable within the context of my data visualization knowledge.

Section D - Importance and Innovation of the Mash-Up

In the last few years a trend that has emerged in the wake of Web 2.0 and the Open Data initiative is a growing interest in how this newly available data can be used. One response to this is data visualization, which takes microdata and creates visual representations that tell a story or help to parse large and diverse sets of data. This mash-up is characteristic of these types of efforts, and offers a glimpse into the way metadata is involved in the process.

Bringing together datasets in a single map requires metadata that describes the content and context of creation of the microdata, and metadata that supports interoperability. Microdata is provided for download in several different formats. For mapping purposes dBase (.dbf) and shapefile (.shp) formats are used to load data into mapping software. These files are binary in that they are created and used in pairs, with the dBase file containing microdata and the shapefile data that allows the plotting of values onto a map.

In the case of the map created in Indiemapper, data was download in XLS file format first in order to eliminate superfluous elements, such as null fields. Each Excel file contained two worksheets: one holding the raw data, and a second one with descriptive metadata. The metadata, identical in scope to what was described above, tells the user the scope of the data, its source and the conditions of its use, however it fails to inform on the more technical aspects of how the data was gathered and manipulated to produce the final product. Exported as CSV files, the data was then uploaded to Geocommons to generate sets of DBF and shapefiles. This step involved geocoding, which required the use of another dataset found on the Geocommons site to write in the shapefile geospatial data that corresponds to the microdata I wished to map. To enable a level of understanding between files to accomplish this, metadata in the form of ISO country codes were matched from the CSV file to the Geocommons dataset. Once the specific country codes had been read, associated vector information residing in the dataset was used to create the shapefile.

After successfully obtaining the suitable files for input into the visualization web service, files are imported one at a time as separate layers of information. Descriptive metadata held in the dbf file in the form of content headers that describe data held in each row are used by the Indiemapper to allow the user to select and manipulate data elements within the design environment. Content metadata are also by default used as legend titles.

Use of the Geocommons.org involves uploading files that then are stored in the user's personal library. Metadata is used extensively to document data sets and control access rights. However, very little of this is created automatically apart from assigning the name of the data set creator and a creation date. The task of describing content rests with the user, who is asked to document the item in several ways. By default, the item is given a name that mirrors its file

name, but it can be changed by the user. Other data fields are optional, and include such elements as descriptive tags that fit the collaborative tagging model, a data abstract, access permissions settings, source of data (including URL), data metadata URL, geocoding data, and contact information for the creator. A further set of fields, autofilled with the content headers found in the data file, encourages the creation of descriptions for each data attribute.

All this metadata simplifies search and access of data sets on the Geocommons. As a site that encourages the interchange and repurposing of data, metadata is a key part of facilitating inventory, identification of resources, and efficient retrieval. Despite, or perhaps because of, the tremendous flexibility allowed in the creation of metadata its quality is not uniform. As a result, the search terms needed to find items can be unpredictable, and the exact quality of data is often a mystery. Short of introducing a validation service to ensure quality control or a rating system that would allow the cream to rise, the current arrangement, while attempting to stimulate metadata use without being a source of irritation for users, detracts from data reliability. The critical aim here should be similar to what is the case with statistics distribution sites, namely, metadata to facilitate the identification, comparison, and evaluation of content. Without the means to ensure this, the usefulness of collections of environmental scientific data is diminished as it becomes more difficult to concatenate data and create models.

Bibliography

Cribb, J. (2011). *The Coming Famine*. U of Cal. Press, 138.

Google Fusion (2012). Google retrieved from http://www.google.com/fusiontables/Home/

Indiemapper Software (2012). Indiemapper, retrieved from http://indiemapper.com/

Tableau Software (2012). Tableau retrieved from http://www.tableausoftware.com

Toulmin, C. (2009). *Climate Change in Africa*. New York, NY: Zed Books, 16.

United Nations (2011). "Fourth United Nations Conference on Least Developed Countries." *United Nations*, retrieved from http://www.un.org/wcm/content/site/ldc/home/Backgroun dround/quick_facts# 11128

United Nations Development Program (2012). *Climate Change Country Profiles*. University of Oxford School of Geography and the Environment, retrieved from http://www.geog.ox.ac .uk /research/climate/projects/undp-cp/

United Nation Environment Programme: Environment for Development (2010). *Cereals – Yields* (Data). Environmental Data Explorer, Data retrieved from http://geodata.grid.unep.ch/ download/download.php

United Nation Environment Programme: Environment for Development (2010). *Cereals – Yields* (Metadata). Environmental Data Explorer, http://geodata.grid.unep.ch/mod_metadata/ metadata.php. Originally published by the Food and Agriculture Organization http://faostat.fao.org/default.aspx

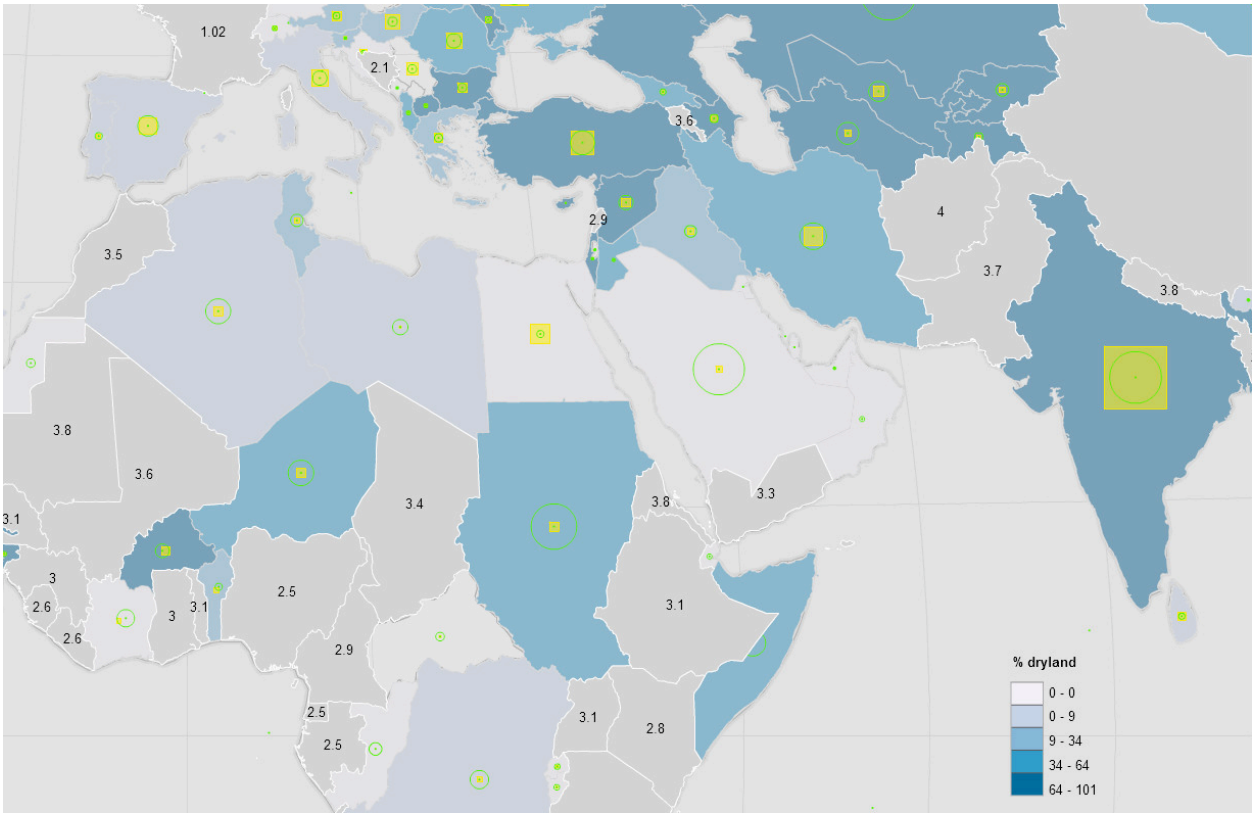
United Nation Environment Programme: Environment for Development (2000). *Drylands – Percent of Total Area* (Data). Environmental Data Explorer: United Nations retrieved from http://geodata.grid.unep.ch/mod_download/download.php. Originally published by the Food and Agriculture Organization http://www.fao.org/ag/agl/agll/terrastat/

United Nation Environment Programme: Environment for Development (2000). *Drylands – Percent of Total Area* (Metadata). Environmental Data Explorer: United Nations, retrieved from http://geodata.grid.unep.ch/mod_metadata/metadata.php

Appendices

Appendix A - Image from visualisation of predicted temperature increases by 2060

Appendix B – Image from visualisation of cereals yield and percent drylands by nation



Appendix C –Metadata for cereal yields by nation

Identification Information

Citation

Title:	Cereals - Yield
Publication year:	2010-11-01
Data type:	publication
GEO Data set type:	National
Spatial presentation form:	mapDigital
GEO Units:	Hectograms per Hectare
Filename:	cereals_yield

Abstract: Cereals also includes other cereals such as mixed grains and buckwheat. The data reported under this element represent the harvested production per unit of harvested area for crop products. In most of the cases yield data are not recorded but obtained by dividing the data stored under production element by those recorded under element: area harvested. Data are recorded in hectogramme (100 grammes) per hectare (HG/HA).

Suppl. information: **GEO AGGREGATIONS COMMENTS**: Pre-1992 (or 1993) relative shares of the former USSR, Czechoslovakia, and Yugoslavia calculated based on data available after 1992 (or 1993), they are exclusively used for the GEO aggregations. **COUNTRIES NOTES**: Data for Belgium prior to 2000 include those for Luxembourg. Data for China include those for Taiwan. Data for Eritrea before 1993 are shown under Ethiopia. Data for Serbia prior 2006 include those for Montenegro

Character set: utf8

Language: English

Status: Complete

Update frequency: annually

Topic Category

GEO theme: Land

GEO data category: Agricultural Production

Theme Keywords

GEMET theme: Agriculture

Free keywords: Agriculture, production, area, food, soil, output

Temporal Extent

Covered time: : 1961
: 2009

Geographic Extent

Coverage: World

West longitude:	East longitude:	South latitude:	North latitude:

Point of Contact

Role:

Person: ----

Organisation: Food and Agriculture Organization of the United Nations (FAO) - FAOStat
Email: FAOSTAT-Queries@fao.org

Resource Constraints

Use constraints: Public

Distribution Information

Contact Information

Online resource: http://geodata.grid.unep.ch/options.php?selectedID=839
URL original source: http://faostat.fao.org/default.aspx

Reference System Information

Coordinate Reference Information

Ellipsoid: WGS 84

Metadata Information

Language: English

Character set: utf8

Date: 2010-11-02

Metadata author

Person: Diawoye Konte
Email: geo@unepgrid.ch
Metadata version: ISO 19115:2003/19139 1.0

:

:

: weighted average
: 2221
: The data is only aggregated if at least 75% of the observations are available (i.e. % of population or % of area or % of countries) on an annual basis. The value "-9999" corresponds to "NoData".

Appendix D – Metadata for percent drylands by nation

Identification Information

Citation

Title:	Population in Drylands - Percent of Total Population
Publication year:	2000
Data type:	publication
GEO Data set type:	National
Spatial presentation form:	mapDigital
GEO Units:	Percent
Filename:	dry_pop_percent

Abstract: The concept of drylands continues to be debated. In this data set, drylands are taken as areas with a potential hazard of desertification. The hyperarid zone is not subject to desertification and is therefore excluded. Hence drylands are defined as the arid, semi-arid and dry subhumid zones, or areas with lengths of growing periods of 1-179 days. Global population database from: Tobler, W., U. Deichmann, J. Gottsegen and K. Maloy (1995), The global demography project, Technical Report 95-6, NCGIA, Santa Barbara. The population surface derives from the latest available national census data, which however varied by country from 1979 to 1994. Standard population growth models were used to derive estimates for 1994.

Suppl. information: ----

Character set: utf8

Language: English

Status: Complete

Update frequency: unknown

Topic Category

GEO theme: Land

GEO data category: Agricultural Production

Theme Keywords

GEMET theme: Soil

Free keywords: Aridity, Desertification, LGP

Temporal Extent

Covered time: : 9999
: 9999

Geographic Extent

Coverage: World

West longitude:	East longitude:	South latitude:	North latitude:

Point of Contact

Role:

Person: Freddy O Nachtergaele
Organisation: Food and Agriculture Organization of the United Nations (FAO) - TerraStat
Email: Freddy.Nachtergaele@fao.org

Resource Constraints

Use constraints: Public

Distribution Information

Contact Information

Online resource: http://geodata.grid.unep.ch/options.php?selectedID=1507
URL original source: http://www.fao.org/ag/agl/agll/terrastat/

Reference System Information

Coordinate Reference Information

Ellipsoid: WGS 84

Metadata Information

Language: English

Character set: utf8

Date: 2003-07-09

Metadata author

Person: Andrea De Bono
Email: geo@unepgrid.ch
Metadata version: ISO 19115:2003/19139 1.0