Building and Using Geospatial Ontology in the BioCaster Surveillance System

Son Doan[1], Quoc-Hung Ngo[2], Ai Kawazoe[1], and Nigel Collier[1]

[1] National Institute of Informatics, Tokyo, Japan

[2] University of Information Technology, Vietnam National University (HCM), Vietnam

Motivation

▪ Disease and location names are important information in any reported news story.

▪ There is currently a lack of available geospatial ontology that covers all countries and major cities with the longitude/latitude information for visualization purposes.

Key idea

- Build a geospatial ontology from Wikipedia (http://www.wikipedia.org).
- Use advanced natural language processing techniques to detect events (including location names) within news stories.

Results

- A geospatial ontology with two administrative levels: 243 names in level 1 and 4,025 names in level 2 with their part-whole relationship and longitudes/latitudes.
- Geospatial ontology was integrated into BioCaster ontology and freely downloaded at http://biocaster.nii.ac.jp/index.php?page=ontology.
- Algorithms for detecting locations are implemented inside the Global Health Monitor, publicly available at http://biocaster.nii.ac.jp.
- We collected data from a 10-week period (Dec 20, 2007 – Feb 20, 2008):
  - 7,412 English news.
  - Covering 110 countries and 360 sub-countries.
  - News by continents: 58.00% Africa, 18.23% Asia, 11.37% South America, 5.30 % North America, 3.40% Middle East, 2.86% Europe and 0.34% Ocean.
- The system successfully detected ebola in Uganda (Bundibugyo, Kampala, Mbarara), yellow fever in Brazil (Goias, Sao Paulo), avian influenza in Indonesia (Jakarta, Banten), and cholera in Vietnam (Ha Noi, Ha Tay).

Challenges and Future work

- Geo-coding: To disambiguate location names, e.g., Camden can be a area in London (UK) or a town in New South Wales (Australia).
- Extend to deeper administrative levels like districts and sub-districts (wards, towns, villages).
- Evaluate and compare to other available resources, i.e., GAZ, dbpedia will be considered.

Acknowledgements

This work was supported by Grants-in-Aid from the Japan Society for the Promotion of Science (grant no. 18049071) and Trandisciplinary Integration Grant from ROIS.

Building a geospatial ontology from Wikipedia

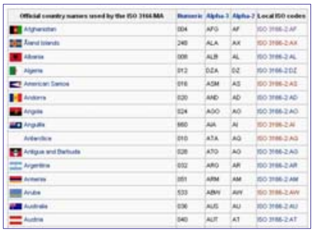
Step 1 Extract names and ISO 3166-1 code of countries and dependent territories (level 1).
For each country, extract names and ISO 3166-2 codes of country subdivisions and dependent areas (level 2).

Step 2 Re-create the part-whole relations between level 1 and level 2 of geographical ontology.

Step 3 Verify results by manually checking local administrative websites.

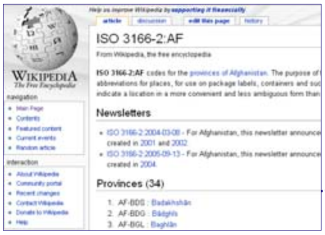
In addition we extracted geo-coordinates of all countries and sub-countries for visualization purpose.

http://en.wikipedia.org/wiki/ISO_3166-1



Official country names used by the ISO 3166/MA	Numeric	Alpha-3	Alpha-2	Local ISO codes
Afghanistan	004	AFG	AF	ISO 3166-2:AF
Åland Islands	248	ALA	AX	ISO 3166-2:AX
Albania	008	ALB	AL	ISO 3166-2:AL
Algeria	012	DZA	DZ	ISO 3166-2:DZ
American Samoa	016	ASM	AS	ISO 3166-2:AS
Andorra	020	AND	AD	ISO 3166-2:AD
Angola	024	AGO	AO	ISO 3166-2:AO
Anguilla	660	AIA	AI	ISO 3166-2:AI
Antarctica	010	ATA	AQ	ISO 3166-2:AQ
Antigua and Barbuda	028	ATG	AG	ISO 3166-2:AG
Argentina	032	ARG	AR	ISO 3166-2:AR
Armenia	051	ARM	AM	ISO 3166-2:AM
Aruba	533	ABW	AW	ISO 3166-2:AW
Australia	036	AUS	AU	ISO 3166-2:AU
Austria	040	AUT	AT	ISO 3166-2:AT

http://en.wikipedia.org/wiki/ISO_3166-2:AF



Help us improve Wikipedia by supporting it financially

article | discussion | edit this page | history

ISO 3166-2:AF

From Wikipedia, the free encyclopedia

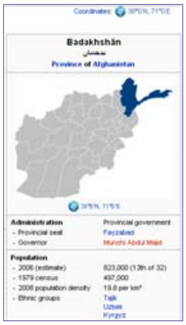
ISO 3166-2:AF codes for the provinces of Afghanistan. The purpose of t
abbreviations for places, for use on package labels, containers and suc
indicate a location in a more convenient and less ambiguous form than

Newsletters

- ISO 3166-2:2004-03-08 - For Afghanistan, this newsletter announce created in 2001 and 2002.
- ISO 3166-2:2005-09-13 - For Afghanistan, this newsletter announce created in 2004.

Provinces (34)

1. AF-BDS : Badakhshān
2. AF-BDG : Bādghīs
3. AF-BGL : Baghlān



Coordinates: 38°0′N, 71°0′E

Badakhshān

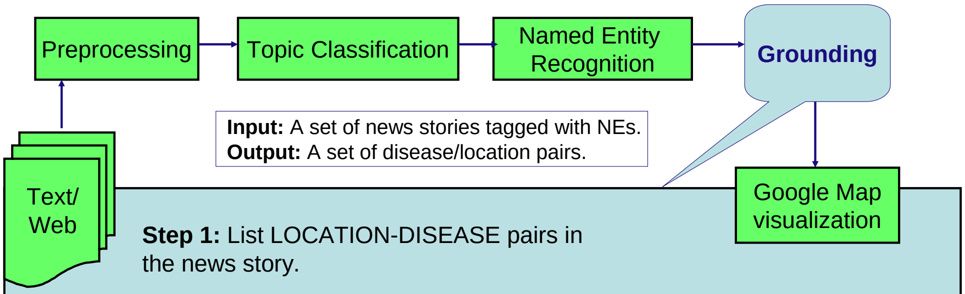
بدخشان

Province of Afghanistan

38°0′N, 71°0′E

Administration	Provincial government
- Provincial seat	Fayzabad
- Governor	Munshi Abdul Majid
Population	
- 2006 (estimate)	823,000 (13th of 32)
- 1979 census	497,000
- 2006 population density	19.8 per km²
- Ethnic groups	Tajik Uzbek Kyrgyz

Algorithm for grounding disease/location into BioCaster ontology



Input: A set of news stories tagged with NEs.
Output: A set of disease/location pairs.

Step 1: List LOCATION-DISEASE pairs in the news story.

Step 2: Calculate the frequency of LOCATION-DISEASE pairs in the news database for the last 12 hours.

Step 3: Rank LOCATION – DISEASE pairs by the frequencies calculated in Step 2. Use a threshold to choose top LOCATION - DISEASE names.

Step 4: Map disease and location names:

- If sequences within DISEASE matches (regular expression matching) to a synonym in BioCaster ontology then DISEASE was assigned to that disease name.
- If sequences within LOCATION matches (regular expression matching) to a location in geospatial ontology then LOCATION was assigned to that location name (countries and cities).

Step 5: Re-map into news stories: Match detected diseases and locations within the first half of each news story. If both disease and location are matched then they are stored; otherwise, skip.

NII – National Institute of Informatics - Japan

BioCaster ontology is freely downloaded from http://biocaster.nii.ac.jp