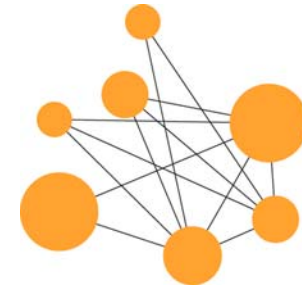
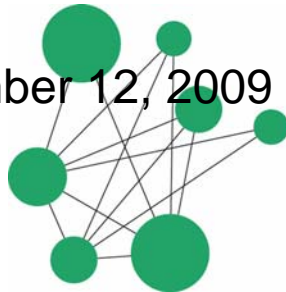


Experiments with Biological Concept Recognition tools

Karin Verspoor, Kevin B. Cohen, Helen
Johnson, Christophe Roeder, César Mejía
Muñoz,
William Baumgartner Jr., Lawrence E.
Hunter



December 12, 2009



Karin.Verspoor@ucdenver.edu

http://compbio.ucdenver.edu/Hunter_lab/Verspc



Concept Recognition in text

turn all highlighting off

date

disease

habitat

institution

organism

person

place

protein

taxon

[Top](#) | [Abstract](#) | [Author Summary](#) | [Introduction](#) | [Methods](#) | [Results](#) | [Discussion](#) | [Supporting Information](#) | [Acknowledgements](#) | [References](#) | [Data Fusion Supplements](#)

Introduction

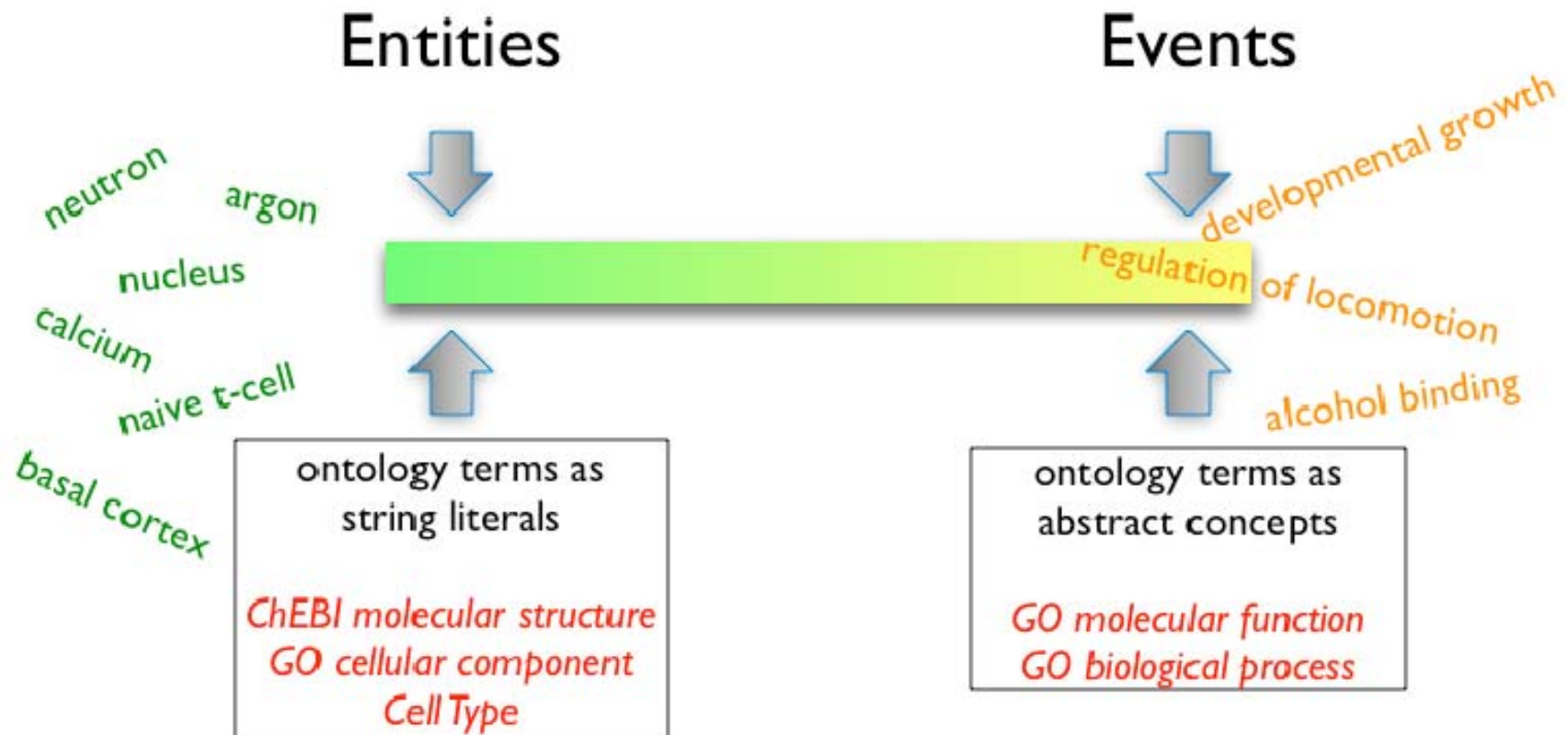
At present, one billion of the world's population resides in **slum settlements** [1]. This number is expected to double in the next 25 years [1]. The growth of large urban populations which are marginalized from basic services has created a new set of global health challenges [2],[3]. As part of the Millennium Development Goals [4], a major priority has been to address the underlying poor sanitation and environmental degradation in slum communities which, in turn, are the cause of a spectrum of neglected diseases which affect these populations [2],[3],[5].

Leptospirosis is a paradigm for an urban health problem that has emerged due to recent growth of **slums** [6],[7]. The disease, caused by the **Leptospira spirochete**, produces life-threatening manifestations, such as **Weil's disease** and severe **pulmonary hemorrhage syndrome** for which fatality is more than 10% and 50%, respectively [7]–[9]. **Leptospirosis** is transmitted during direct contact with animal reservoirs or water and soil contaminated with their urine [8],[9]. Changes in the urban environment due to expanding slum communities has produced conditions for rodent-borne transmission [6],[10]. Urban epidemics of **leptospirosis** now occur in **cities** throughout the developing world during seasonal heavy rainfall and flooding [6],[11]–[18]. There is scarce data on the burden of specific diseases that affect slum populations [2], however **leptospirosis** appears to have become a major infectious disease problem in this population. In **Brazil** alone, more than 10,000 cases of severe **leptospirosis** are reported each year due to outbreaks in **urban centers** [19], whereas roughly 3,000, 8,000 and 1,500 cases are reported annually for **meningococcal disease**, **visceral leishmaniasis** and **dengue hemorrhagic fever**, respectively, which are other infectious diseases associated with urban poverty [20]–[22]. Case fatality (10%) from **leptospirosis** [19] is comparable to that observed for **meningococcal disease**, **visceral leishmaniasis** and **dengue hemorrhagic fever** (20%, 8% and 10%, respectively) in this setting [20],[23],[24]. Furthermore, **leptospirosis** is associated with extreme weather events, as exemplified by the El Niño-associated outbreak in **Guayaquil** in 1998 [25]. **Leptospirosis** is therefore expected to become an increasingly important slum health problem as predicted global climate change [26],[27] and growth of the world's slum population [1] evolves.

Urban **leptospirosis** is a disease of poor environments since it disproportionately affects communities that lack adequate sewage systems and refuse collection services [6],[10],[11]. In this setting, outbreaks are often due to transmission of a single serovar, **L. interrogans serovar Copenhageni**, which is associated with the **Rattus norvegicus** reservoir [6], [28]–[30]. Elucidation of the specific determinants of poverty which have led to the emergence of urban **leptospirosis** is essential in guiding community-based interventions which, to date, have been uniformly unsuccessful. Herein, we report the findings of a large seroprevalence survey performed in a Brazilian slum community (*favela*). Geographical Information System (GIS) methods were used to identify sources for **Leptospira** transmission in the **slum environment**. Furthermore, we evaluated whether relative differences in socioeconomic status among slum residents contributed to the risk of **Leptospira infection**, in addition to the attributes of the environment in which they reside.



Biological Concept Recognition





Two dictionary-based tools tested against CRAFT*

- UIMA ConceptMapper

<http://incubator.apache.org/uima/sandbox.html#concept.mapper.annotator>

- stemming and case matching relaxation
- non-contiguous spans
- ignore stopwords
- order-independent lookup

- Open Biomedical Annotator

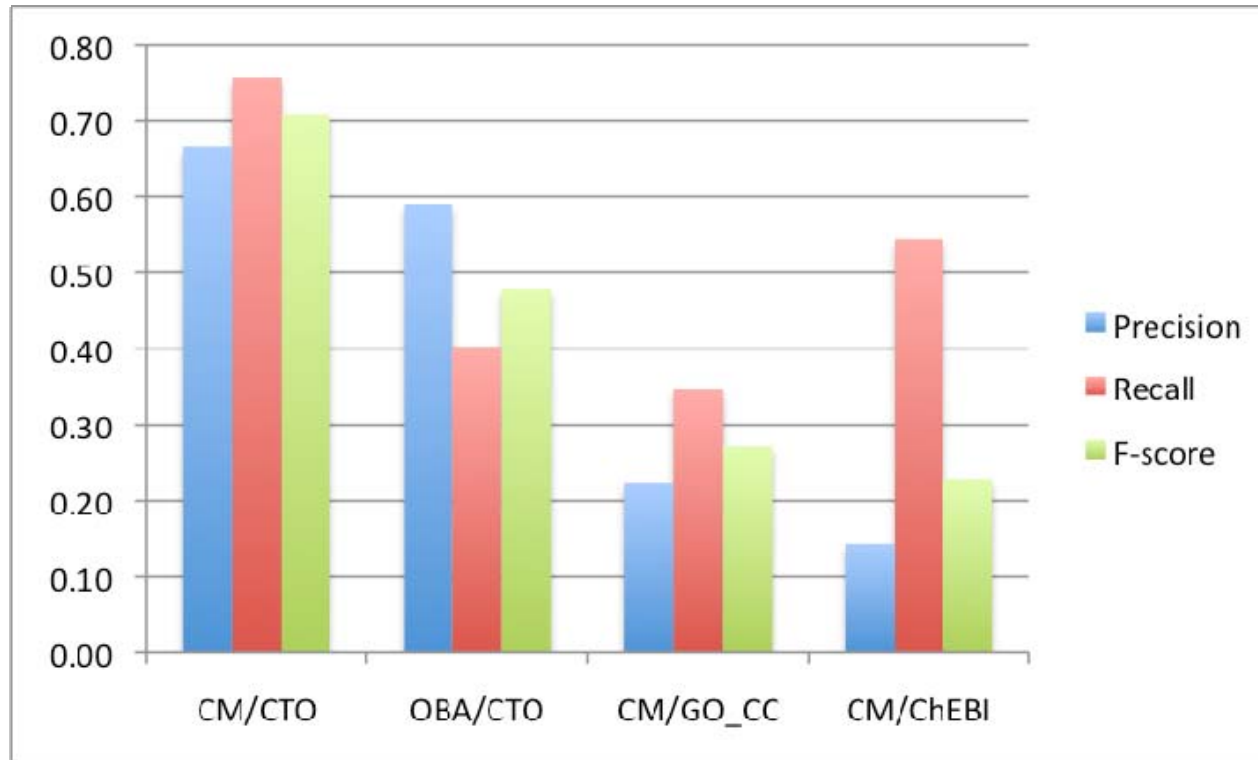
<http://bioportal.bioontology.org/annotator>

- ignore stopwords
- partial word matches

**/listen to Mike Bede's talk for details)*



Best run results



- CM/CTO: stemming + FindAllMatches: false
- OBA/CTO: using default stop words
- CM/GO_CC: stemming + caseMatch: insensitive
- CM/ChEBI: caseMatch: sensitive



Conclusions

- The kinds of terms in the ontology matter
- The strategies used in the dictionary matching tools matter
- We are working on strategies that go beyond dictionary matching ...



Acknowledgements

- Larry Hunter (Lab director)
 - Mike Bada (Ontologist)
 - Bill Baumgartner (Research engineer)
 - Kevin Bretonnel Cohen (Senior Researcher)
 - Lynne Fox (Librarian)
 - Helen Johnson (Linguist)
 - César Mejía Muñoz (undergraduate research asst)
 - Chris Roeder (Software engineer)
 - Hannah Tinney (Analyst)
- NIH 5R01LM009254
 - NIH 2R01LM008111-04A1
 - NIH 5G08LM009639-02
 - NIH 5R01GM083649-02
 - NIH 5T15LM009451-02