DEFENCE R&D DÉFENSE

DEFENCE R&D DÉFENSE

CAPTURING AND MODELING DOMAIN KNOWLEDGE USING NATURAL LANGUAGE PROCESSING TECHNIQUES

Alain Auger, Ph. D.

IKM Section / DRDC Valcartier

June 2005

(Paper 296)

Defence Research and
Development Canada

Recherche et développement
pour la défense Canada

Canada

Problem Space

- Command and control (C2) and decision-making domains are seriously threatened facing information overload and uncertainty issues
- Military have to create new ways of processing sensor and intelligence information
- **Without new means to elicit knowledge from multiple information and intelligence sources, decision-makers will have to deal with very limited knowledge and increasing levels of uncertainty in operations**
- How can we better capture and represent knowledge objects contained in sources?

Knowledge Representation Enablers

- Metadata
- Taxonomies
- Ontologies

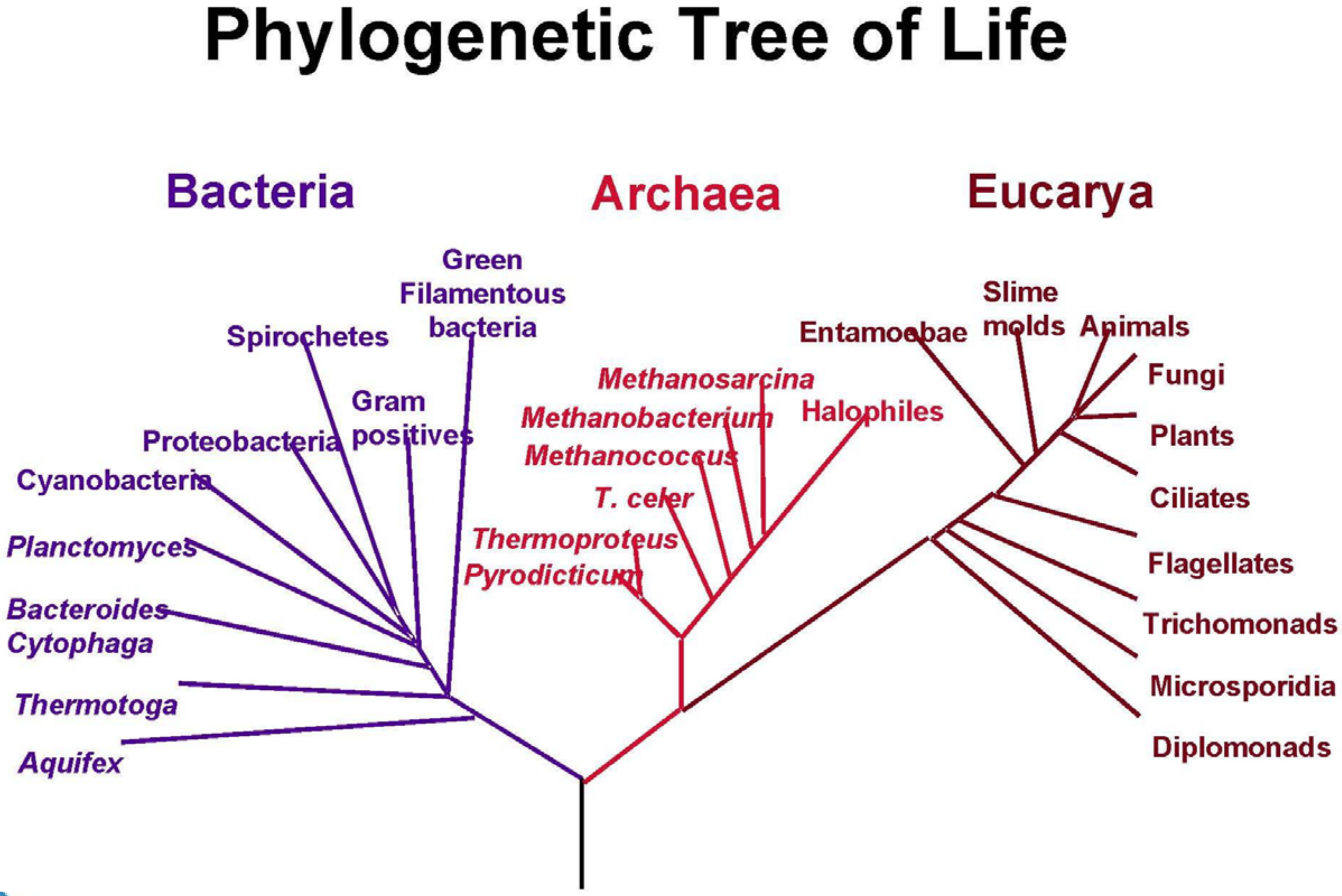
Some Metadata Sets

- **Metadata** (Greek: *meta-* + *data* "information") means « data about data ».
- Dublin Core
  - The **Dublin Core Metadata Element Set** consists of 16 optional metadata elements, any of which may be repeated or omitted. (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Audience)
- Resource Description Framework (RDF)
  - The purpose of RDF is to provide an encoding and interpretation mechanism so that resources can be described in a way that particular software can understand it, or, better put, so that software can more easily access data organized within structured parameters.
- Extended Markup Language (XML)
- Etc.

Taxonomies

- Taxonomy (from Greek ταξινομία (*taxinomia*) from the words *taxis* "order" and *nomos* "law") may refer to either a hierarchical classification of things, or the principles underlying the classification. Almost anything, animate objects, inanimate objects, places, and events, may be classified according to some taxonomic scheme. [Wikipedia]
- In taxonomies, concepts are classified using **homology**; that is, **shared characteristics that have been inherited from a common ancestor**.
- **Limitation**: IS-A or PARENT-CHILD relationship type only. Cannot express CAUSE-EFFECT relationships, for instance

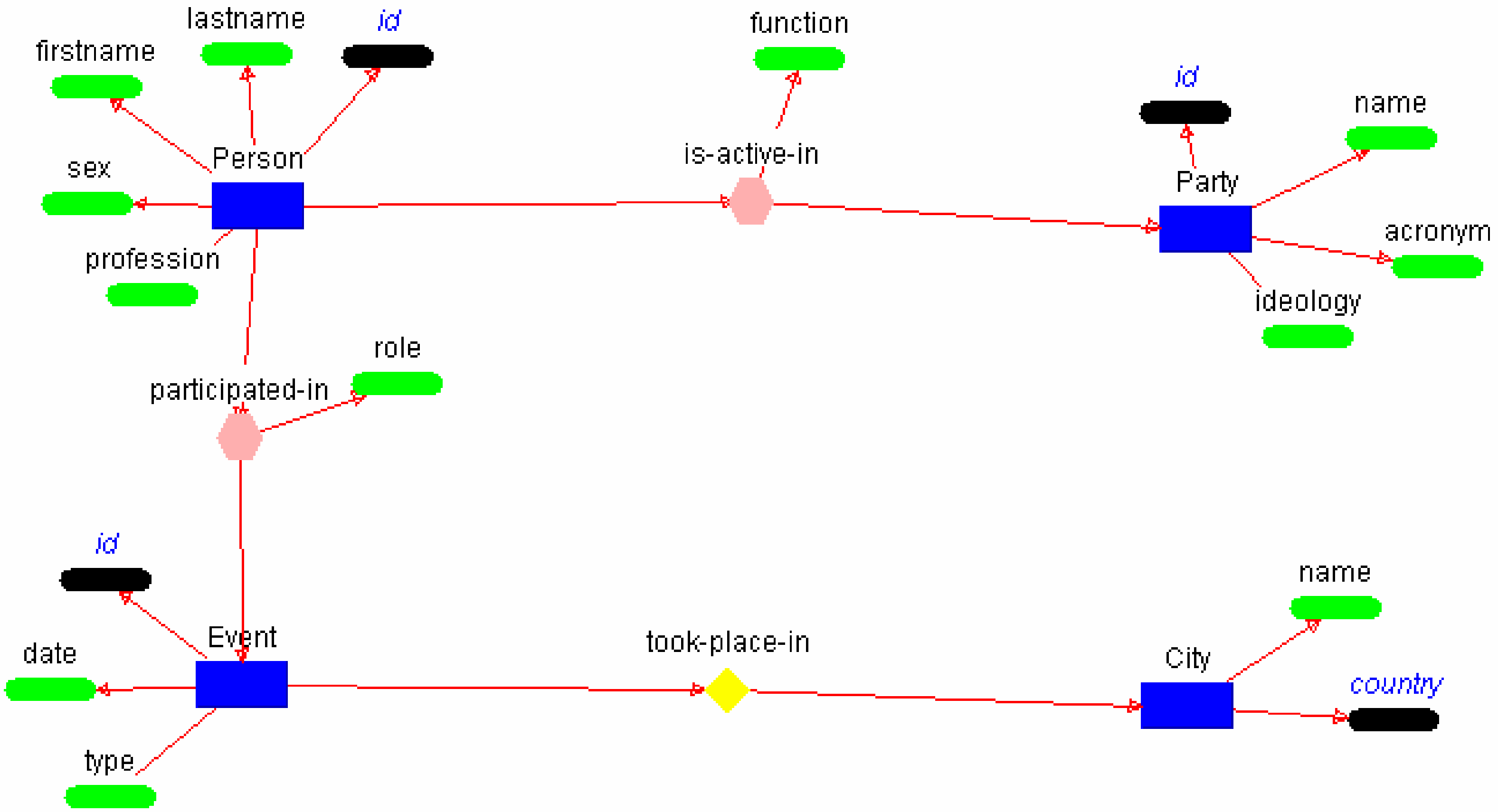
Taxonomy Sample



Ontologies

- An **ontology** is a formal, explicit specification of a shared conceptualisation [Gruber, 1993]
- An ontology is a formal explicit specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

Ontology Sample

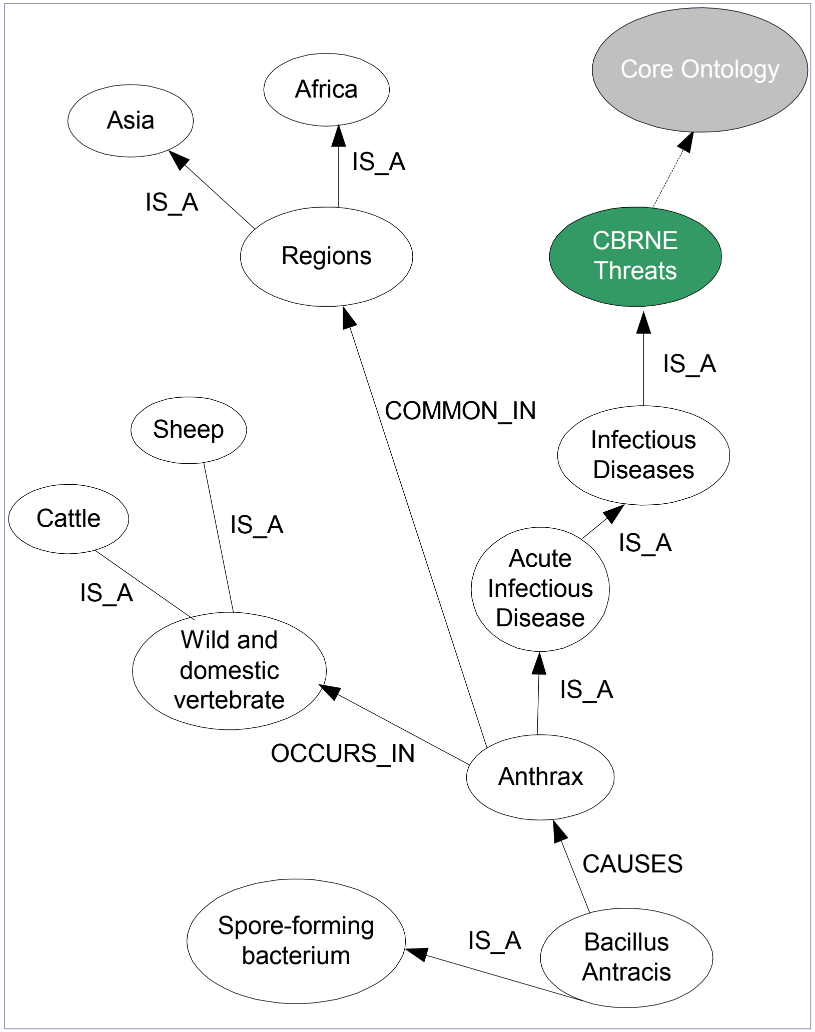


The Need for Domain Ontologies

- Domain ontologies are key elements required to enable next generation of decision support and knowledge exploitation systems with new semantic capabilities
- Ontology-engineering remains a non-trivial, time and budget consuming activity
- How can we rapidly build ontologies?

SACOT Research Project

- Aim:
  - To develop and apply natural language processing (NLP) extraction techniques to unstructured texts to capture knowledge objects they contain and represent them in the form of an ontology

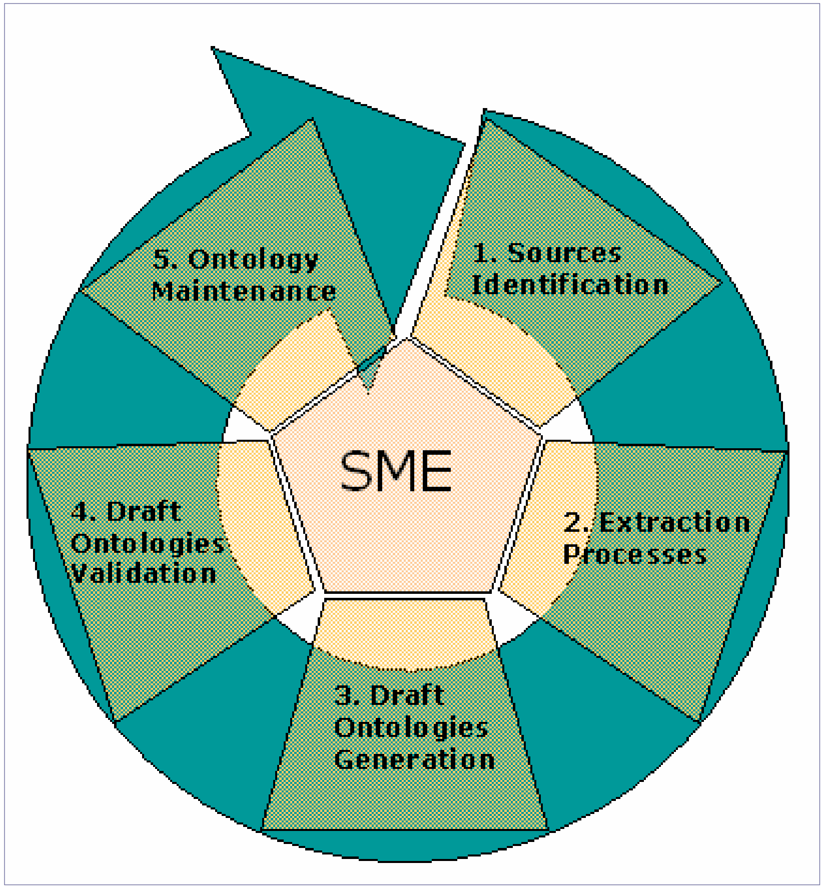


Limitations of Traditional Ontology-Engineering Approaches

- Relying on Humans
  - – Based on Subject Matter Experts
  - – Adapted to task or application ontologies
  - – Not adapted to domains ontologies (too many objects)
- Relying on Statistics
  - – e.g. computation of co-occurring words

SACOT Ontology-engineering Process

- Sources Identification
- Extraction Processes
- Draft Ontologies Generation
- Draft Ontologies Validation
- Ontology Maintenance

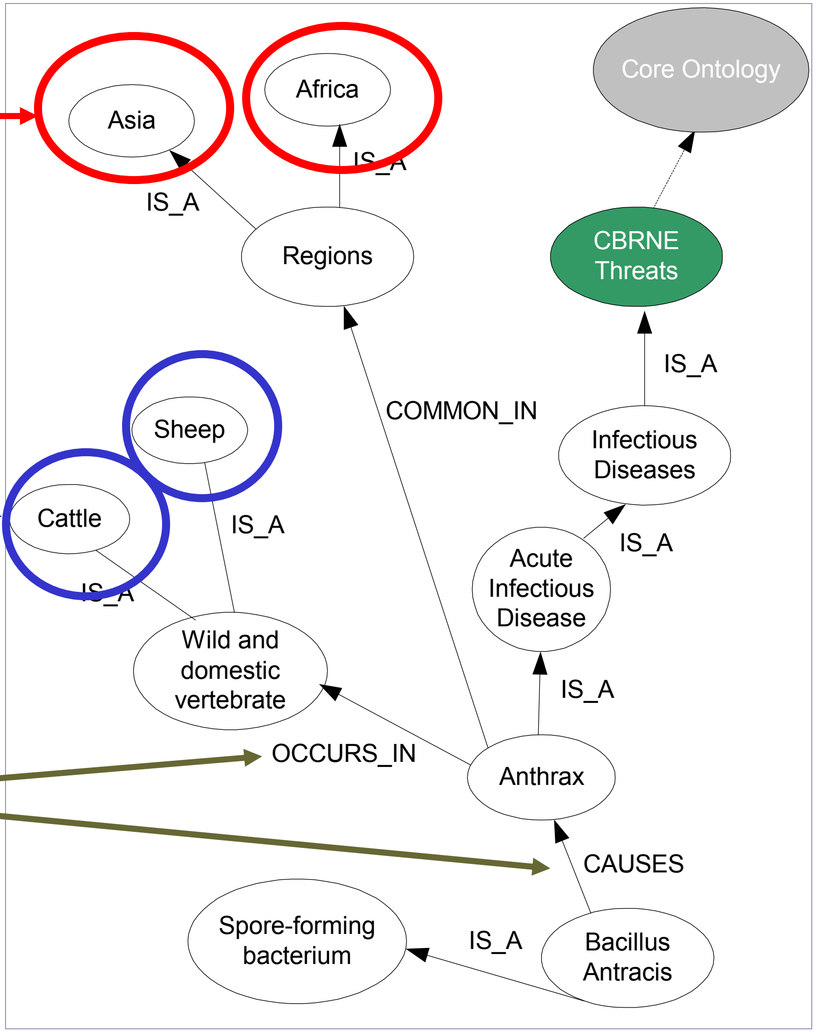

SME
1. Sources Identification
2. Extraction Processes
3. Draft Ontologies Generation
4. Draft Ontologies Validation
5. Ontology Maintenance

What are Domain Ontologies Made of?

- Named Entities expressed in texts

- Concepts expressed by terms in texts

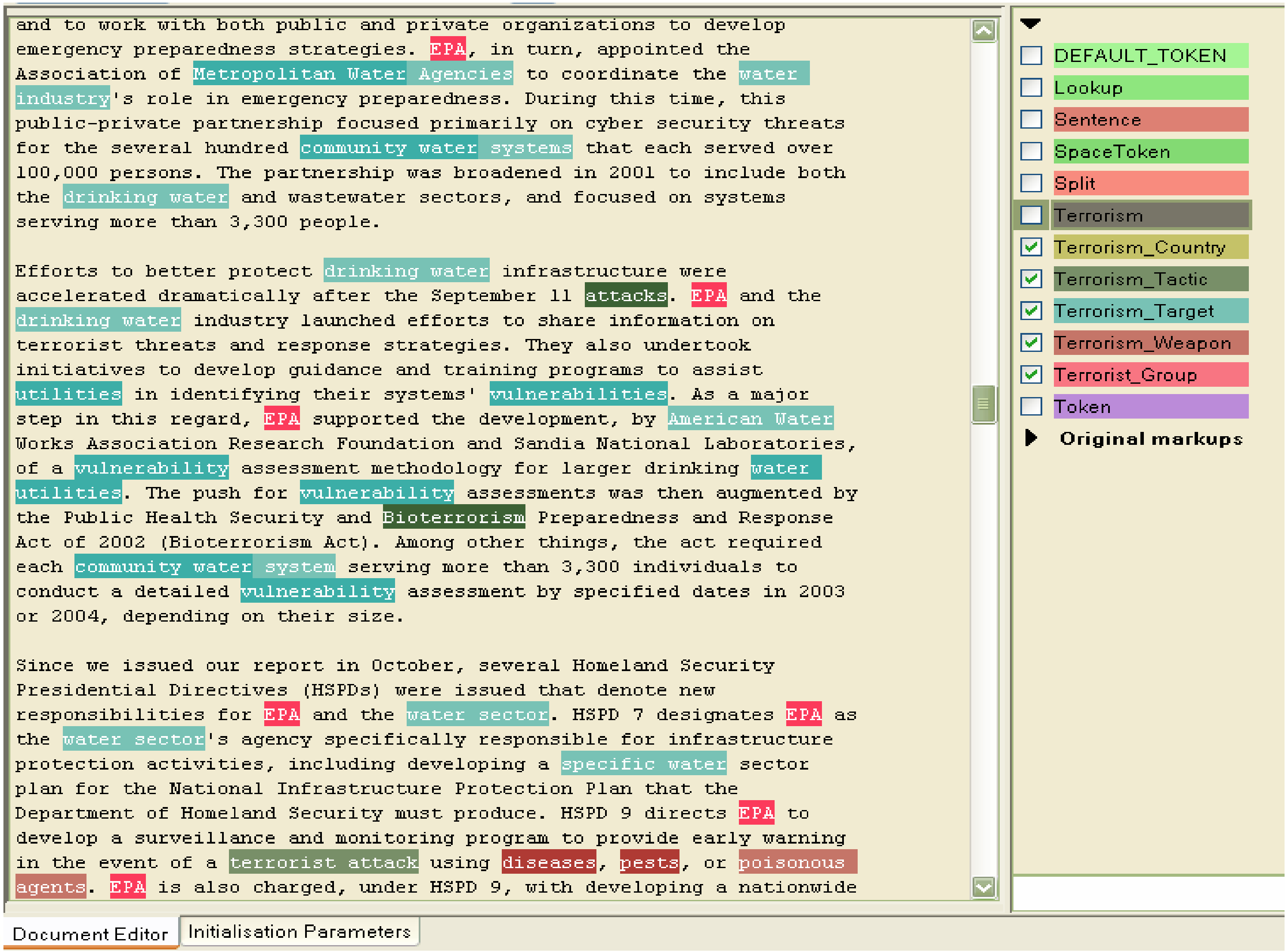
- Relations among knowledge objects



SACOT's Specifics

- Domain-specific Named Entity Extraction
- Contrastive Approach to Terminology Extraction
- Natural Language Processing (NLP) approach to semantic relations extraction

Named Entities Extraction



and to work with both public and private organizations to develop emergency preparedness strategies. EPA, in turn, appointed the Association of Metropolitan Water Agencies to coordinate the water industry's role in emergency preparedness. During this time, this public-private partnership focused primarily on cyber security threats for the several hundred community water systems that each served over 100,000 persons. The partnership was broadened in 2001 to include both the drinking water and wastewater sectors, and focused on systems serving more than 3,300 people.

Efforts to better protect drinking water infrastructure were accelerated dramatically after the September 11 attacks. EPA and the drinking water industry launched efforts to share information on terrorist threats and response strategies. They also undertook initiatives to develop guidance and training programs to assist utilities in identifying their systems' vulnerabilities. As a major step in this regard, EPA supported the development, by American Water Works Association Research Foundation and Sandia National Laboratories, of a vulnerability assessment methodology for larger drinking water utilities. The push for vulnerability assessments was then augmented by the Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (Bioterrorism Act). Among other things, the act required each community water system serving more than 3,300 individuals to conduct a detailed vulnerability assessment by specified dates in 2003 or 2004, depending on their size.

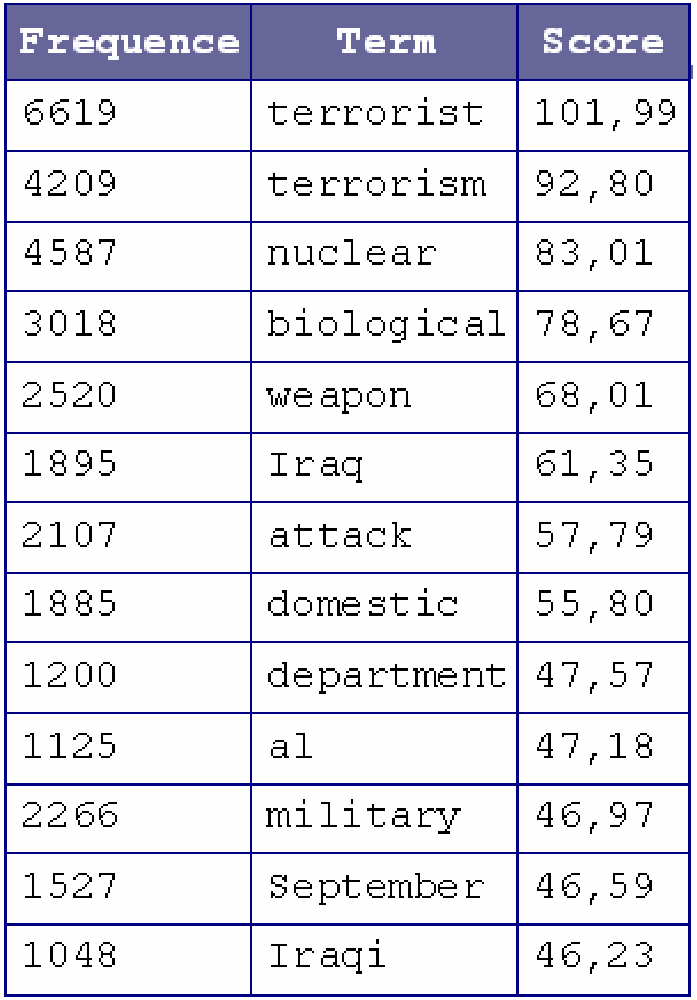
Since we issued our report in October, several Homeland Security Presidential Directives (HSPDs) were issued that denote new responsibilities for EPA and the water sector. HSPD 7 designates EPA as the water sector's agency specifically responsible for infrastructure protection activities, including developing a specific water sector plan for the National Infrastructure Protection Plan that the Department of Homeland Security must produce. HSPD 9 directs EPA to develop a surveillance and monitoring program to provide early warning in the event of a terrorist attack using diseases, pests, or poisonous agents. EPA is also charged, under HSPD 9, with developing a nationwide

▼
- ☐ DEFAULT_TOKEN
- ☐ Lookup
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Terrorism
- ☑ Terrorism_Country
- ☑ Terrorism_Tactic
- ☑ Terrorism_Target
- ☑ Terrorism_Weapon
- ☑ Terrorist_Group
- ☐ Token

▶ Original markups

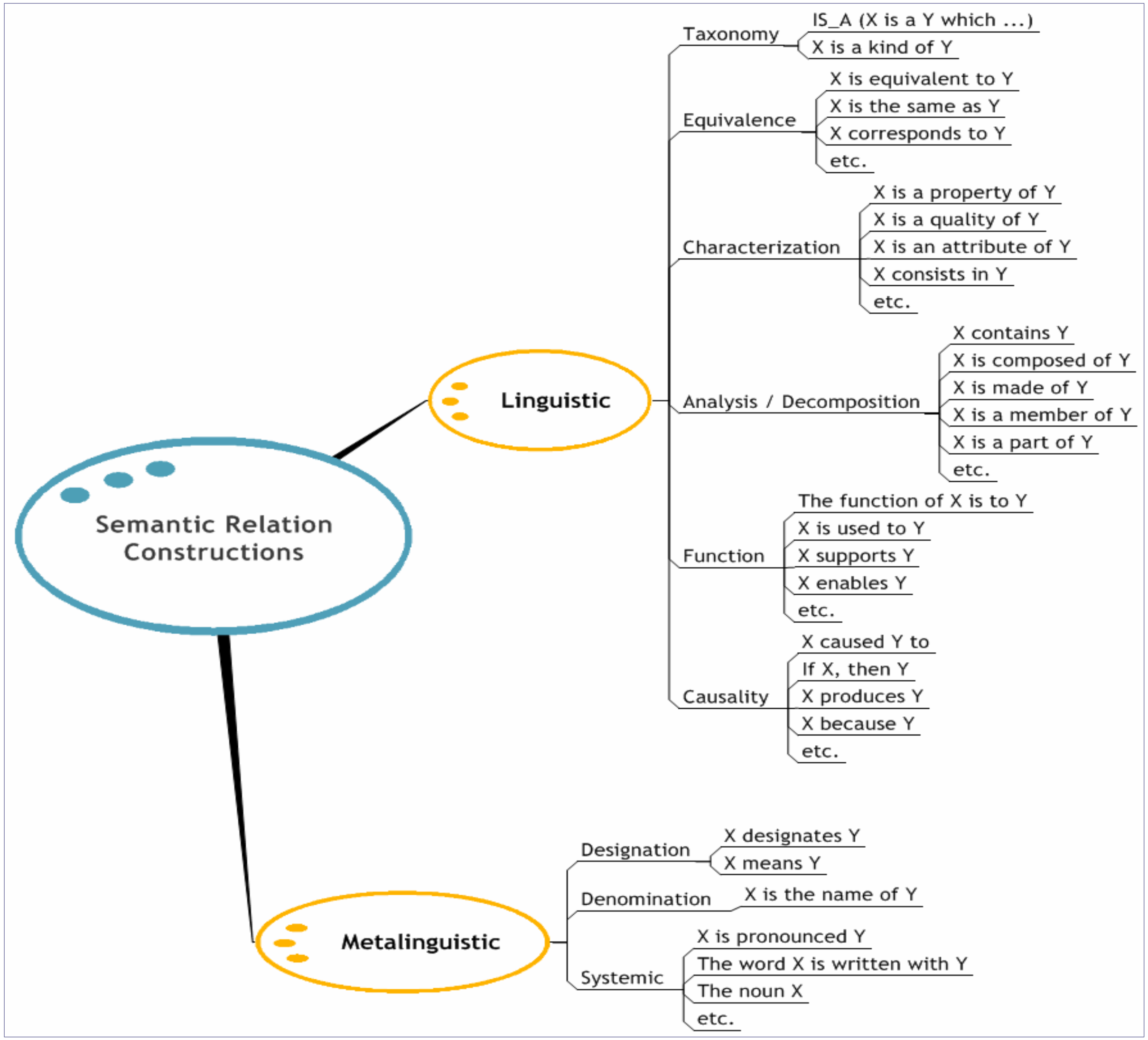
Terminology Extraction

- SACOT's Specifics:
  - Use of a contrastive approach to compute and automate candidate terms validation process



Frequence	Term	Score
6619	terrorist	101,99
4209	terrorism	92,80
4587	nuclear	83,01
3018	biological	78,67
2520	weapon	68,01
1895	Iraq	61,35
2107	attack	57,79
1885	domestic	55,80
1200	department	47,57
1125	al	47,18
2266	military	46,97
1527	September	46,59
1048	Iraqi	46,23

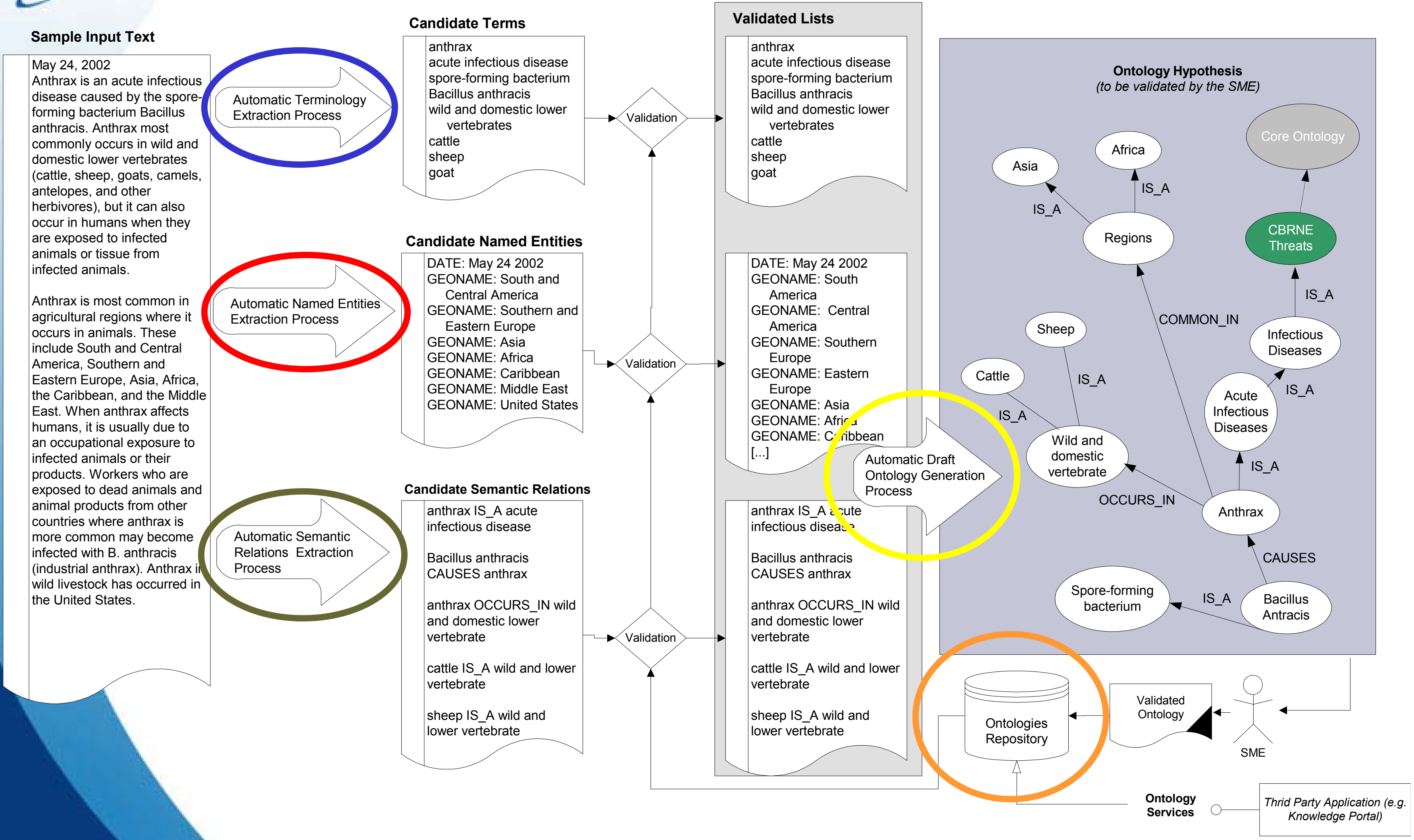
Semantic Relations Extraction



SACOT's Specifics

- Targeting specific semantic relations markers that are present in texts as explicit « indicators » to capture relations among concepts
  - e.g. *X is used to Y*, *X is located in Y*
- Not based on co-occurrence statistics
- Entirely based on semantic relation patterns
  - e.g. *is used to*, *is located in*

Putting it All Together



Sample Input Text

May 24, 2002
Anthrax is an acute infectious disease caused by the spore-forming bacterium Bacillus anthracis. Anthrax most commonly occurs in wild and domestic lower vertebrates (cattle, sheep, goats, camels, antelopes, and other herbivores), but it can also occur in humans when they are exposed to infected animals or tissue from infected animals.

Anthrax is most common in agricultural regions where it occurs in animals. These include South and Central America, Southern and Eastern Europe, Asia, Africa, the Caribbean, and the Middle East. When anthrax affects humans, it is usually due to an occupational exposure to infected animals or their products. Workers who are exposed to dead animals and animal products from other countries where anthrax is more common may become infected with B. anthracis (industrial anthrax). Anthrax in wild livestock has occurred in the United States.

Candidate Terms

anthrax
acute infectious disease
spore-forming bacterium
Bacillus anthracis
wild and domestic lower vertebrates
cattle
sheep
goat

Candidate Named Entities

DATE: May 24 2002
GEONAME: South and Central America
GEONAME: Southern and Eastern Europe
GEONAME: Asia
GEONAME: Africa
GEONAME: Caribbean
GEONAME: Middle East
GEONAME: United States

Candidate Semantic Relations

anthrax IS_A acute infectious disease

Bacillus anthracis CAUSES anthrax

anthrax OCCURS_IN wild and domestic lower vertebrate

cattle IS_A wild and lower vertebrate

sheep IS_A wild and lower vertebrate

Validated Lists

anthrax
acute infectious disease
spore-forming bacterium
Bacillus anthracis
wild and domestic lower vertebrates
cattle
sheep
goat

DATE: May 24 2002
GEONAME: South America
GEONAME: Central America
GEONAME: Southern Europe
GEONAME: Eastern Europe
GEONAME: Asia
GEONAME: Africa
GEONAME: Caribbean
[...]

anthrax IS_A acute infectious disease

Bacillus anthracis CAUSES anthrax

anthrax OCCURS_IN wild and domestic lower vertebrate

cattle IS_A wild and lower vertebrate

sheep IS_A wild and lower vertebrate

Ontology Hypothesis
(to be validated by the SME)

Automatic Terminology Extraction Process
Automatic Named Entities Extraction Process
Automatic Semantic Relations Extraction Process
Validation
Validation
Validation
Automatic Draft Ontology Generation Process
Core Ontology
Asia
Africa
IS_A
IS_A
Regions
CBRNE Threats
IS_A
COMMON_IN
Sheep
Infectious Diseases
Cattle
IS_A
IS_A
IS_A
Acute Infectious Diseases
Wild and domestic vertebrate
IS_A
OCCURS_IN
Anthrax
CAUSES
Spore-forming bacterium
IS_A
Bacillus Antracis
Ontologies Repository
Validated Ontology
SME
Ontology Services
Thrird Party Application (e.g. Knowledge Portal)

Conclusion

- Preliminary results show that the SACOT ontology-engineering framework might significantly reduces time usually required to capture the knowledge objects of a domain in traditional, fully human-based, ontology building processes.

Project Status

- Initiated in 2004, SACOT is a research project in its early stage.
- All extraction modules are still under development
- All existing modules are standalone at the moment. They are not integrated in the SACOT framework.

Way Ahead

- Measure performance of all three extraction modules
- Integrate all extraction modules in the SACOT framework
- Investigate machine learning techniques in support to SME validation of draft ontologies generated by the SACOT framework

DEFENCE R&D DÉFENSE