Semantic Similarity Computation and Concept Mapping in Earth and Environmental Science

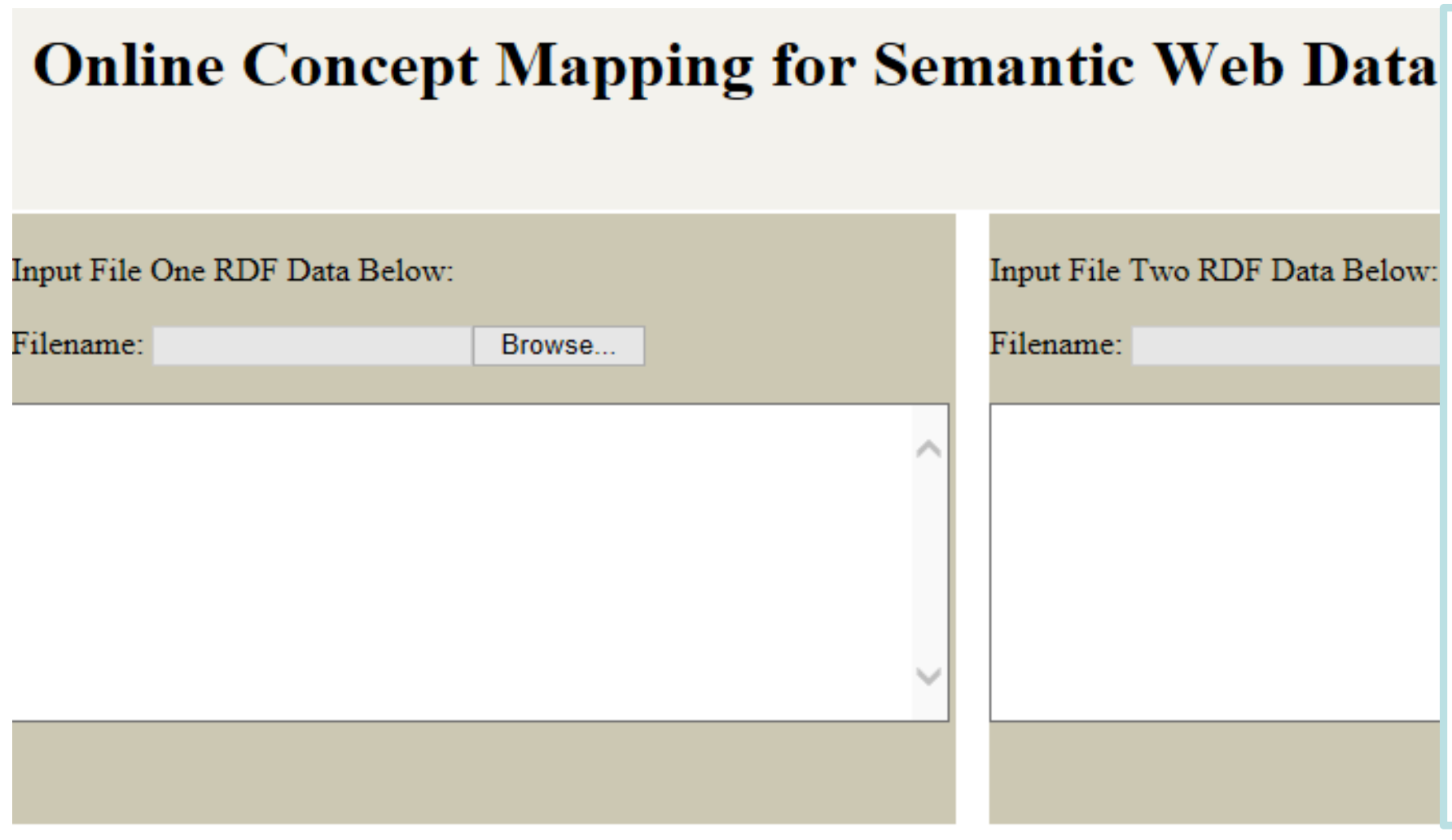
Jin Guang Zheng (zhengj3@rpi.edu), Xiaogang Ma (max7@rpi.edu), Stephan Zednik (zednis@rpi.edu), Peter Fox (foxp@rpi.edu)

Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, United States

1 ABSTRACT

Ontologies have been widely adopted and used by Earth and Environmental Science community to capture and represent knowledge in the domain. One of the major problems that prevent us to combine and reuse these ontologies to conduct real-world applications is the semantic heterogeneity issue, for example, a same term from two different ontologies may refer to two different concepts; or two terms from two different ontologies may have the same meaning. In this work, we addressed the problem by (1) developing a semantic similarity computation model to compute similarity among the concepts in Earth and Environmental Science; (2) based on the computation model, we implemented a concept mapping tool that creates alignment for concepts that are semantically the same or similar; (3) we demonstrated the effectiveness of the tool using GCMD and CLEAN vocabularies and other earth science related ontologies.

3 CONCEPT MAPPING BETWEEN GCMD AND CLEAN



Online Concept Mapping for Semantic Web Data

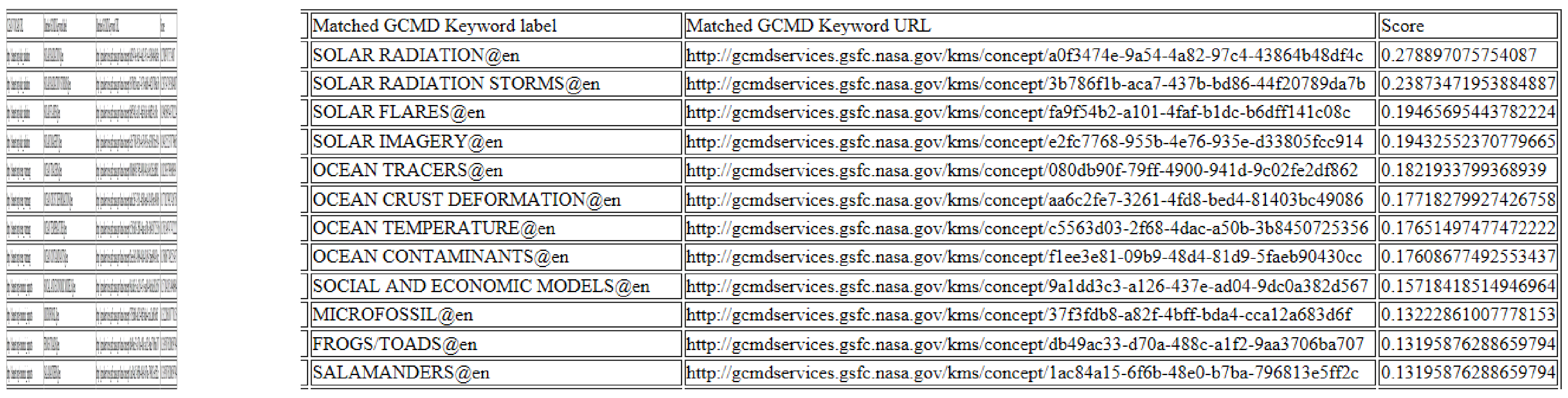
Input File One RDF Data Below:

Filename: Browse...

Input File Two RDF Data Below:

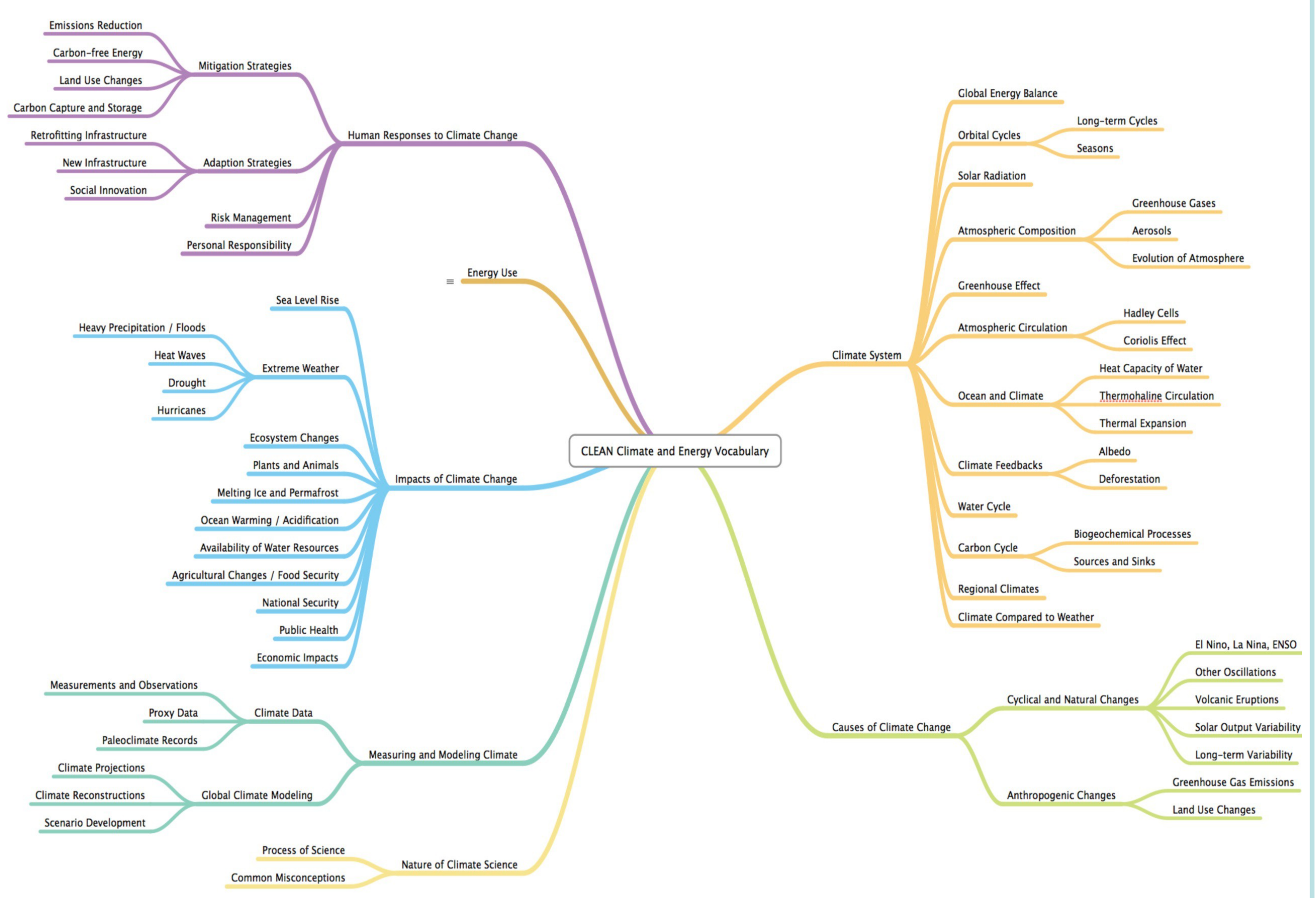
Filename:

- We provide an online interface for concept mapping service, where user can submit their RDF data.
- For each concept from ontology A, we return four most similar concepts from ontology B, where user can interact with the system to perform final selection.
- For each suggestion, we provide a similarity score computed by the system as a guide for the user.

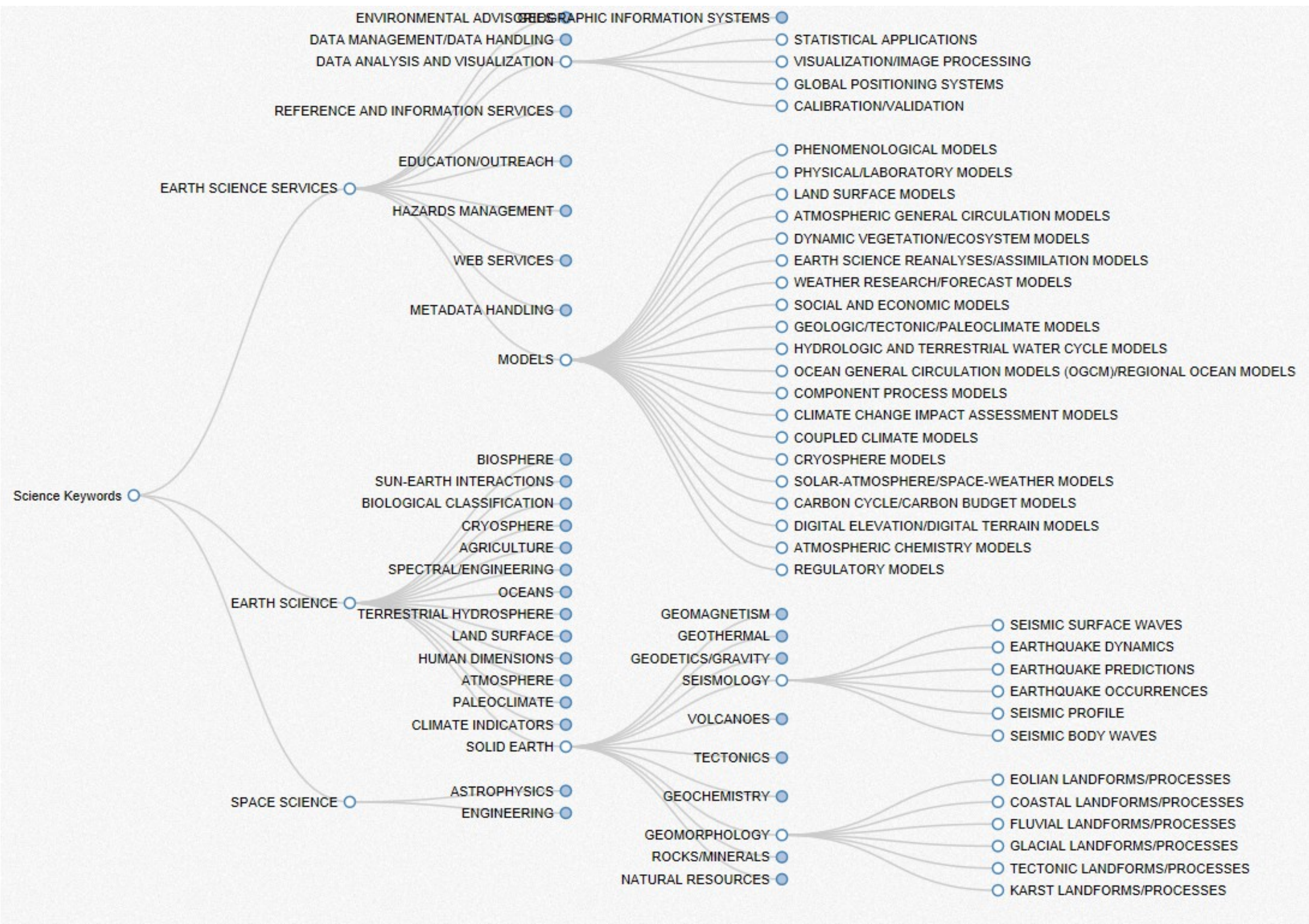


Matched GCMD Keyword label	Matched GCMD Keyword URL	Score
SOLAR RADIATION@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/a0f3474e-9a54-4a82-97c4-43864b48df4c	0.278897075754087
SOLAR RADIATION STORMS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/3b786f1b-aca7-437b-bd86-44f20789da7b	0.23873471953884887
SOLAR FLARES@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/fa9f54b2-a101-4faf-b1dc-b6dff141c08c	0.19465695443782224
SOLAR IMAGERY@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/e2fc7768-955b-4e76-935e-d33805fcc914	0.19432552370779665
OCEAN TRACERS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/080db90f-79ff-4900-941d-9c02fe2df862	0.1821933799368939
OCEAN CRUST DEFORMATION@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/aa6c2fe7-3261-4fd8-bed4-81403bc49086	0.17718279927426758
OCEAN TEMPERATURE@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/c5563d03-2f68-4dac-a50b-3b8450725356	0.17651497477472222
OCEAN CONTAMINANTS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/f1ee3e81-09b9-48d4-81d9-5faeb90430cc	0.17608677492553437
SOCIAL AND ECONOMIC MODELS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/9a1dd3c3-a126-437e-ad04-9dc0a382d567	0.15718418514946964
MICROFOSSIL@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/37f3fdb8-a82f-4bff-bda4-cca12a683d6f	0.13222861007778153
FROGS/TOADS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/db49ac33-d70a-488c-a1f2-9aa3706ba707	0.13195876288659794
SALAMANDERS@en	http://gcmdservices.gsfc.nasa.gov/kms/concept/1ac84a15-6f6b-48e0-b7ba-796813e5ff2c	0.13195876288659794

MINDMAP OF CLEAN VOCABULARIES



MINDMAP OF SUBSET OF GCMD VOCABULARIES


Science Keywords
EARTH SCIENCE SERVICES
ENVIRONMENTAL ADVISORIES
GEOGRAPHIC INFORMATION SYSTEMS
DATA MANAGEMENT/DATA HANDLING
DATA ANALYSIS AND VISUALIZATION
STATISTICAL APPLICATIONS
VISUALIZATION/IMAGE PROCESSING
GLOBAL POSITIONING SYSTEMS
CALIBRATION/VALIDATION
REFERENCE AND INFORMATION SERVICES
EDUCATION/OUTREACH
HAZARDS MANAGEMENT
WEB SERVICES
METADATA HANDLING
MODELS
PHENOMENOLOGICAL MODELS
PHYSICAL/LABORATORY MODELS
LAND SURFACE MODELS
ATMOSPHERIC GENERAL CIRCULATION MODELS
DYNAMIC VEGETATION/ECOSYSTEM MODELS
EARTH SCIENCE REANALYSES/ASSIMILATION MODELS
WEATHER RESEARCH/FORECAST MODELS
SOCIAL AND ECONOMIC MODELS
GEOLOGIC/TECTONIC/PALEOCLIMATE MODELS
HYDROLOGIC AND TERRESTRIAL WATER CYCLE MODELS
OCEAN GENERAL CIRCULATION MODELS (OGCM)/REGIONAL OCEAN MODELS
COMPONENT PROCESS MODELS
CLIMATE CHANGE IMPACT ASSESSMENT MODELS
COUPLED CLIMATE MODELS
CRYOSPHERE MODELS
SOLAR-ATMOSPHERE/SPACE-WEATHER MODELS
CARBON CYCLE/CARBON BUDGET MODELS
DIGITAL ELEVATION/DIGITAL TERRAIN MODELS
ATMOSPHERIC CHEMISTRY MODELS
REGULATORY MODELS
EARTH SCIENCE
BIOSPHERE
SUN-EARTH INTERACTIONS
BIOLOGICAL CLASSIFICATION
CRYOSPHERE
AGRICULTURE
SPECTRAL/ENGINEERING
OCEANS
TERRESTRIAL HYDROSPHERE
LAND SURFACE
HUMAN DIMENSIONS
ATMOSPHERE
PALEOCLIMATE
CLIMATE INDICATORS
SOLID EARTH
GEOMAGNETISM
GEOTHERMAL
GEODETICS/GRAVITY
SEISMOLOGY
SEISMIC SURFACE WAVES
EARTHQUAKE DYNAMICS
EARTHQUAKE PREDICTIONS
EARTHQUAKE OCCURRENCES
SEISMIC PROFILE
SEISMIC BODY WAVES
VOLCANOES
TECTONICS
GEOCHEMISTRY
GEOMORPHOLOGY
EOLIAN LANDFORMS/PROCESSES
COASTAL LANDFORMS/PROCESSES
FLUVIAL LANDFORMS/PROCESSES
GLACIAL LANDFORMS/PROCESSES
TECTONIC LANDFORMS/PROCESSES
KARST LANDFORMS/PROCESSES
ROCKS/MINERALS
NATURAL RESOURCES
SPACE SCIENCE
ASTROPHYSICS
ENGINEERING

- Both CLEAN and GCMD provides rich set of terms to describe Earth and Environment related concepts and knowledge, and are widely used by the scientists
- Some of the terms in both GCMD and CLEAN are describing same concepts
- Create a concept mapping between GCMD and CLEAN will enable more interesting works such as data integration
- Using this mapping tool, we performed concept mapping between GCMD and CLEAN, subset of result is shown in the table.

Get the poster at:

DEFINITIONS

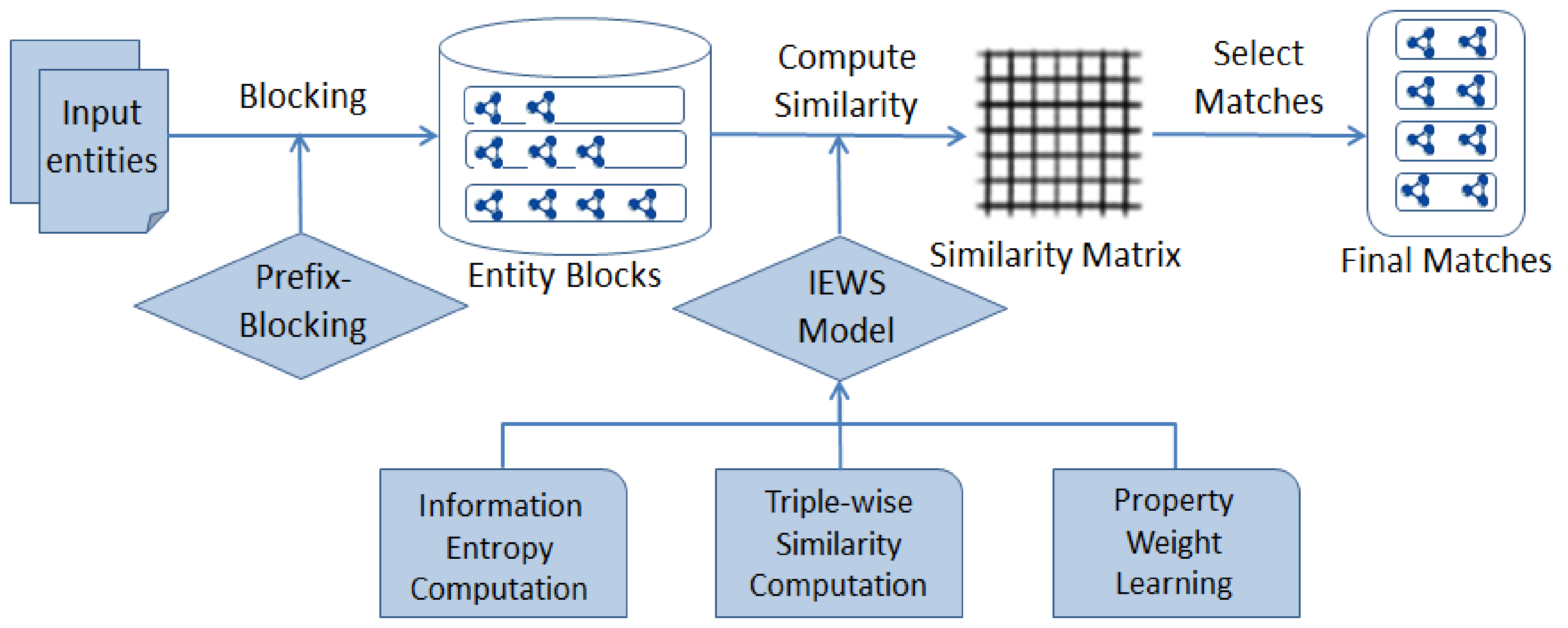
Semantic Similarity: Semantic similarity measures how alike two things are at semantic and concept level.

Information Entropy: Information entropy measures uncertainty of a given information. [1]

Global Change Master Directory (**GCMD**): The GCMD holds more than 28,000 Earth science data set and service descriptions, which cover subject areas within the Earth and environmental sciences. [2]

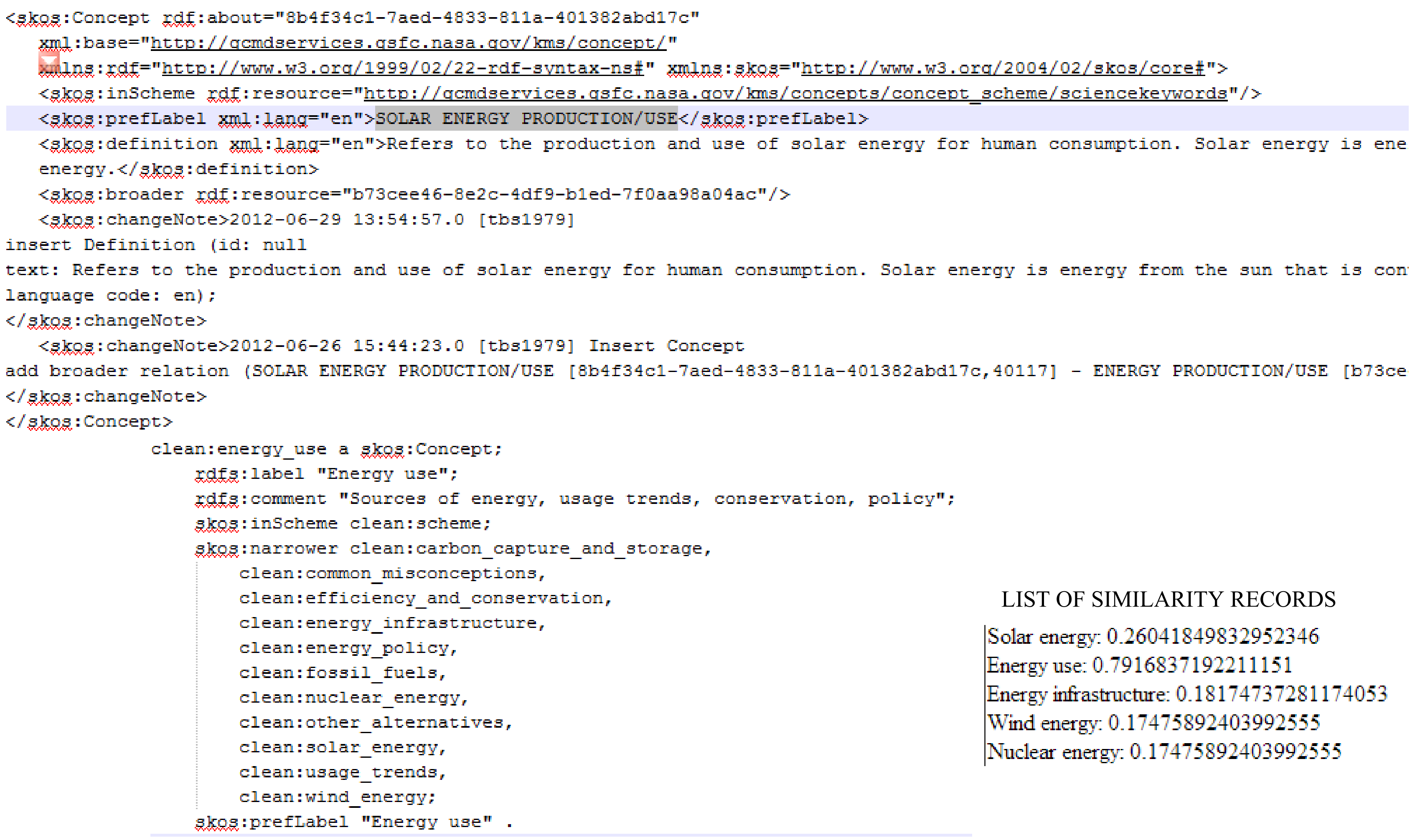
Climate Literacy and Energy Awareness Network pathway (**CLEAN**): Digital resources for teaching about climate science, climate change and energy awareness – resources are reviewed by educators and scientists, and annotated and aligned with standards and benchmarks. [3]

2 SEMANTIC SIMILARITY COMPUTATION MODEL



- Semantic similarity score is computed using Information Entropy and Weighted Similarity (IEWS) Model
- IEWS Model consists three components: Information Entropy Computation Component, Property Weight Component, Triple-wise Similarity Computation Component
- Information Entropy Computation Component and Property Weight Component computes importance and amount of information are given by the description of entities. Triple-wise Similarity Computation component uses these information to compute triple-wise similarity score. Combining these triple-wise similarity score, we get a final similarity score for each pair of entities
- $\mathrm{Sim}^{F}(e_1,e_2)=\mathrm{H}(\mathrm{P})\dfrac{\sum Sim^{wpv}}{\sum Sim^{wpv}+\propto(\sum WP_1-\sum Sim^{wpv})+\propto(\sum WP_2-\sum Sim^{wpv})}$

4 TRACING SIMILARITY COMPUTATION



```
<skos:Concept rdf:about="8b4f34c1-7aed-4833-811a-401382abd17c"
   xml:base="http://gcmdservices.gsfc.nasa.gov/kms/concept/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:skos="http://www.w3.org/2004/02/skos/core#">
   <skos:inScheme rdf:resource="http://gcmdservices.gsfc.nasa.gov/kms/concepts/concept_scheme/sciencekeywords"/>
   <skos:prefLabel xml:lang="en">SOLAR ENERGY PRODUCTION/USE</skos:prefLabel>
   <skos:definition xml:lang="en">Refers to the production and use of solar energy for human consumption. Solar energy is ene
   energy.</skos:definition>
   <skos:broader rdf:resource="b73cee46-8e2c-4df9-b1ed-7f0aa98a04ac"/>
   <skos:changeNote>2012-06-29 13:54:57.0 [tbs1979]
insert Definition (id: null
text: Refers to the production and use of solar energy for human consumption. Solar energy is energy from the sun that is con
language code: en);
</skos:changeNote>
   <skos:changeNote>2012-06-26 15:44:23.0 [tbs1979] Insert Concept
add broader relation (SOLAR ENERGY PRODUCTION/USE [8b4f34c1-7aed-4833-811a-401382abd17c,40117] - ENERGY PRODUCTION/USE [b73ce
</skos:changeNote>
</skos:Concept>
            clean:energy_use a skos:Concept;
                rdfs:label "Energy use";
                rdfs:comment "Sources of energy, usage trends, conservation, policy";
                skos:inScheme clean:scheme;
                skos:narrower clean:carbon_capture_and_storage,
                    clean:common_misconceptions,
                    clean:efficiency_and_conservation,
                    clean:energy_infrastructure,
                    clean:energy_policy,
                    clean:fossil_fuels,
                    clean:nuclear_energy,
                    clean:other_alternatives,
                    clean:solar_energy,
                    clean:usage_trends,
                    clean:wind_energy;
                skos:prefLabel "Energy use" .
```

LIST OF SIMILARITY RECORDS

Solar energy:	0.26041849832952346
Energy use:	0.7916837192211151
Energy infrastructure:	0.18174737281174053
Wind energy:	0.17475892403992555
Nuclear energy:	0.17475892403992555

- We provide a list of triple-wise similarity records between two entities.
- Each record tells how similar two entities are according to one of these entities' features or property descriptions.
- Features or property descriptions can be dereferenced for further human analysis (why/how these features are similar/different).
- Scientists can study the similarity records to decide whether or not to accept the mapping.

5 CONCLUSIONS AND FUTURE WORK

- In this work, we presented Information Entropy and Weighted Similarity Model, which computes semantic similarity among entities from different ontologies
- We developed an online concept mapping tool and performs concept mapping on GCMD and CLEAN. The result demonstrates we can find and match similar concepts between Earth science related ontologies.
- We implemented an explanation feature, where scientists will be able to see why such a mapping is created, and how two concepts are similar.
- We will also like to perform concept mapping using more Earth and Environmental related ontologies to demonstrate the applicability of this concept mapping tool in Earth and Environmental related studies.

REFERENCES:

[1] http://en.wikipedia.org/wiki/Entropy_(information_theory)
[2] http://gcmdservices.gsfc.nasa.gov/index.html
[3] http://cleanet.org/index.html

SPONSORS:

ESIP 'Funding Friday' Award 2013

UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH