Knowledge Discovery in Climate Change Domain

Pinar Öztürk, Erwin Marsi

Norwegian University of Science and Technology (NTNU), Norway

Natalia Manola

University of Athens, Greece

Outline

- Introduction of Ocean-Certain (OC) EU- project
- Knowledge Discovery in OC
- Decisions underlying OC's Knowledge Discovery system
  - Type of knowledge to focus on
  - Corpus
  - Text mining subtasks
  - Technology/tool
  - External sources
- Some results (and examples) so far
- Conclusions

EU Project "Ocean Certain"

- Title: **Ocean** Food web Patrol – Climate Effects: Reducing Targeted Un**certain**ties with an Interactive Network
- Work programme topic : F7- ENV.2013.6.1-1
- EU funding: 7.1 Mill Euro
  - Our workpackage : 40 man/month
  - 3 years, with start Nov. 2014

List of participants:

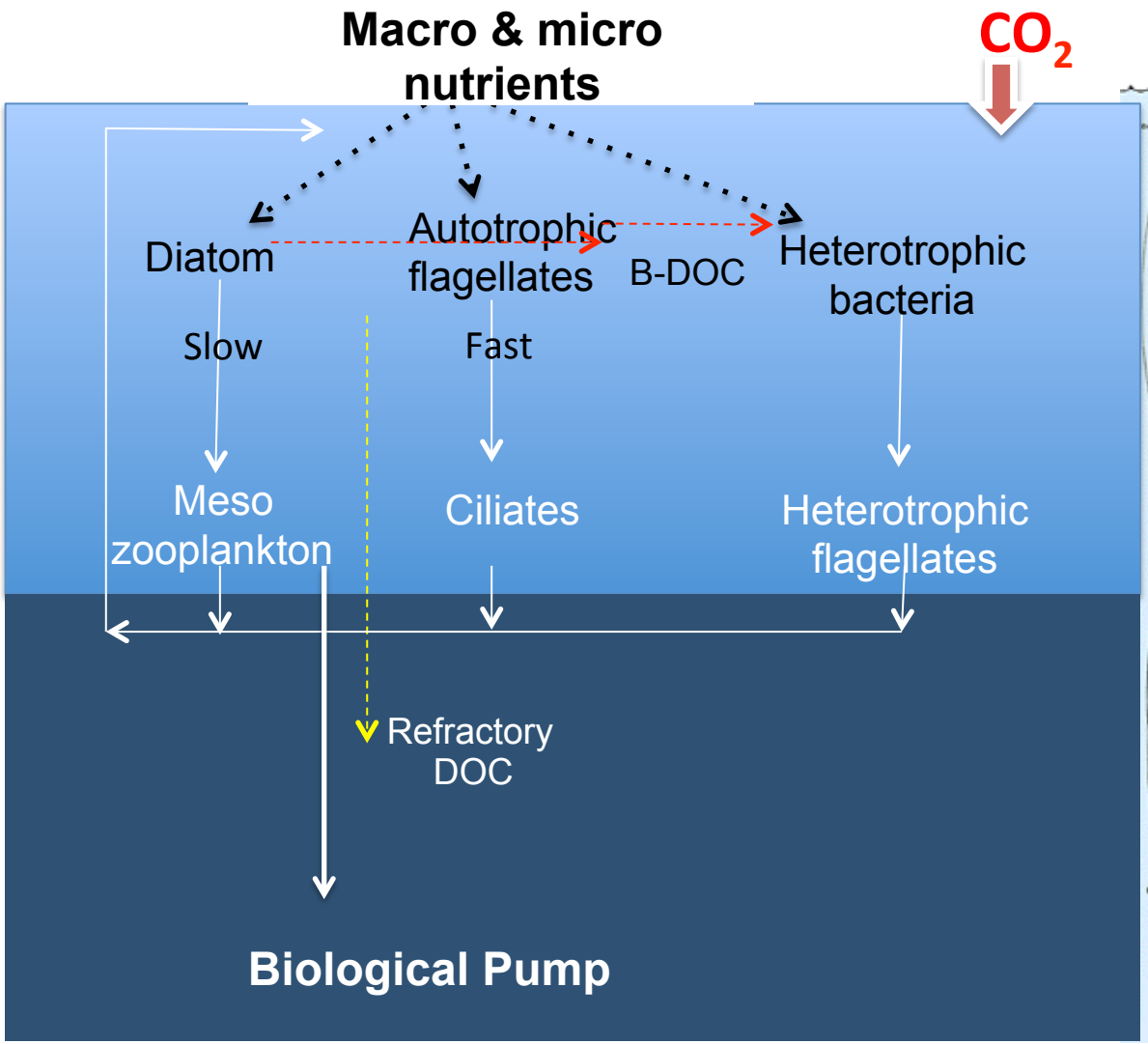
Partner no. *	Participant organisation name	Country
1 (Coordinator)	Norwegian University of Science and Technology	Norway
2	University of Bergen	Norway
3	GEOMAR Helmholtz Centre for Ocean Research Kiel	Germany
4	Vlaamse Instelling voor Technologisch Onderzoek	Belgium
5	DEU-IMST	Turkey
6	University of Gothenburg	Sweden
7	Griffith University	Australia
8	Universidad Austral de Chile	Chile
9	National Research Council and Institute of Marine Sciences	Italy
10	Centre for Environment, Fisheries & Aquaculture Science	UK
11	World Ocean Council	UK
12	Universidad de Concepción	Chile

Overarching goal of Ocean-Certain

- Identifying the interactions (impacts and feedbacks) between the climate related oceanic processes, food web and biological pump

- Determining qualitative and quantitative changes in the functionalities of the “food web” and estimating the efficiency of the “biological pump” in exporting carbon to deep sea

Climate change domain



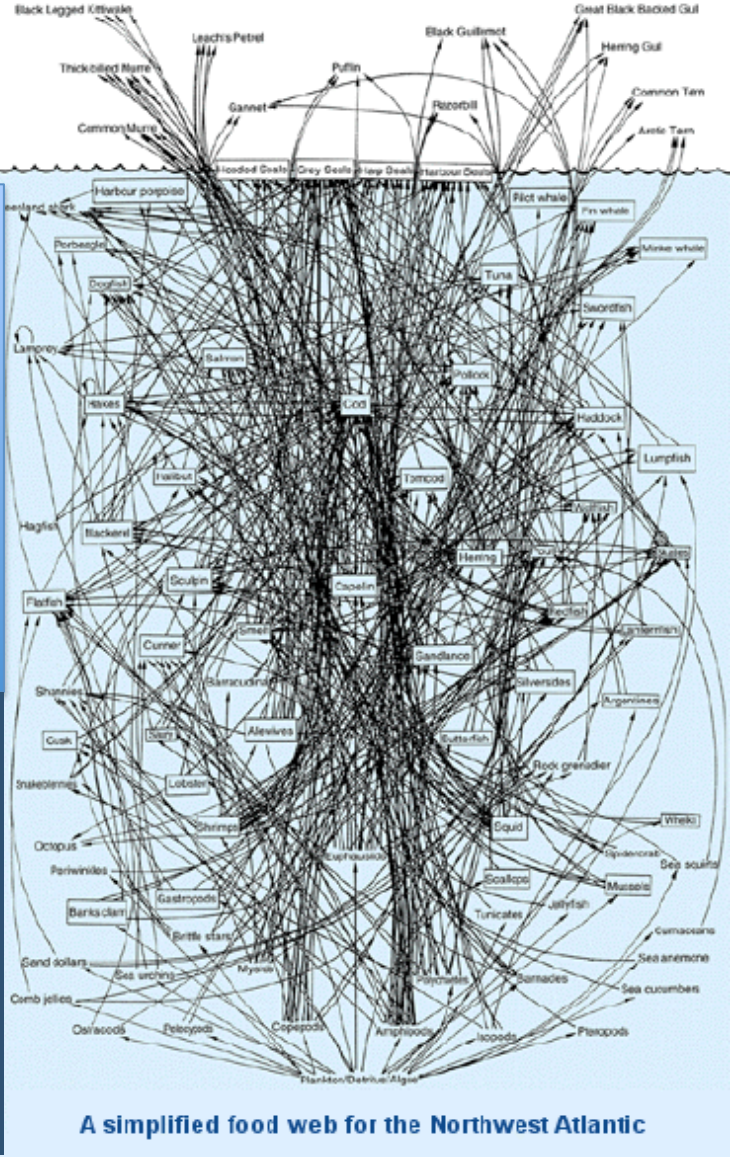


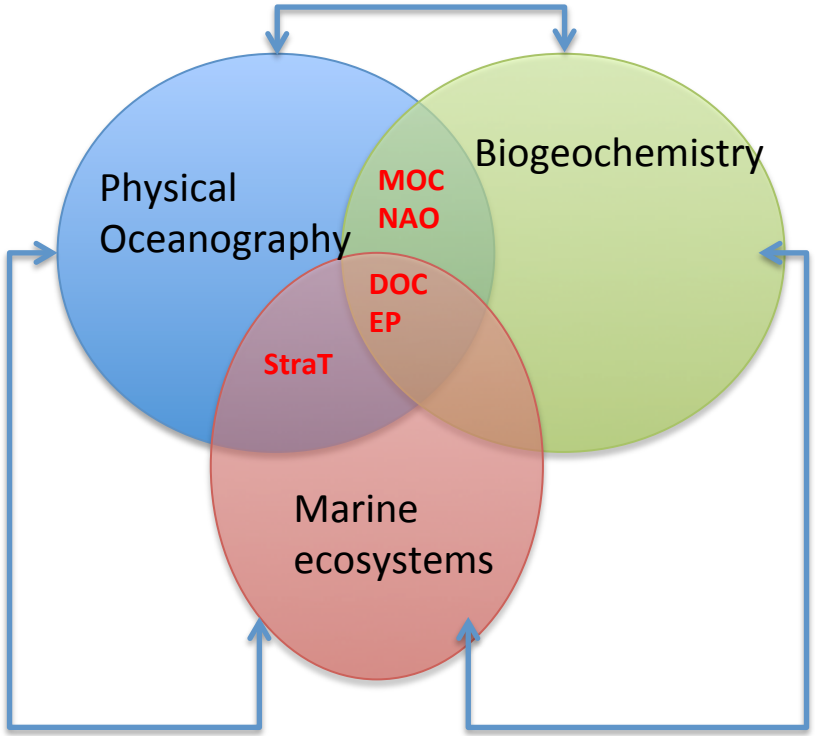
Figure from Thingstad et al. (2008)

Main obstacle of scientific discovery

- is often not lack of scientific research and reporting of these – i.e., not knowledge
- is the lack of ability of *linking* various disciplines and making sense out of the accumulated/documented knowledge across disciplines – i.e., inferring new knowledge from the existing knowledge

Why knowledge linking is challenging

- Vast amount of literature and growing – info overload
- Increased specialisation
- Isolated research communities and literatures – research silos
- Different conventions and terminology



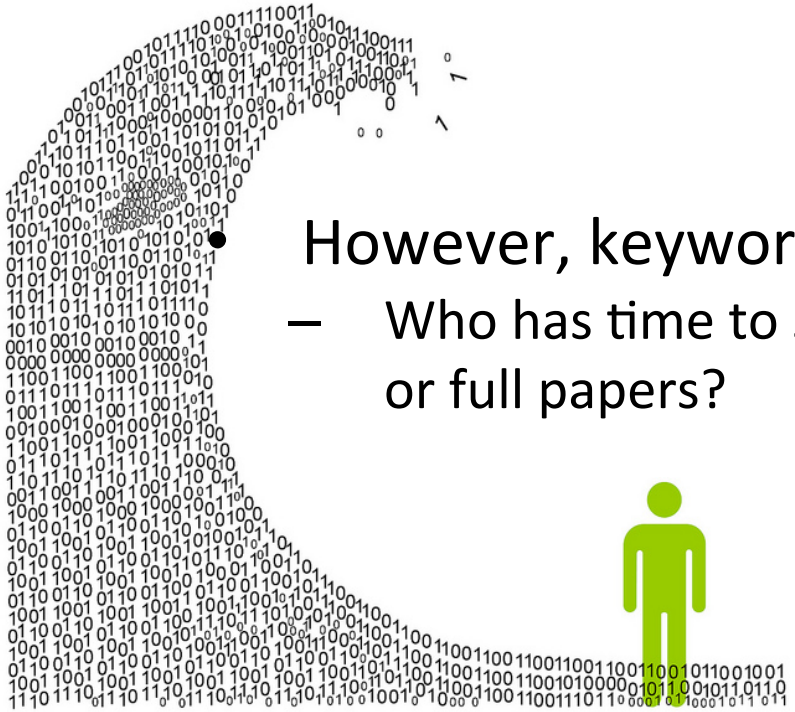
StraT: Stratification
DOC: Dissolved organic carbon
EP: Carbon Export (*synonym*: biological pump)
MOC: Meridional Overturning current
NAO: North Atlantic oscilliation

Computational support for handling scientific text

- Support the user in various ways
  - Search
  - Question-answering
  - Citation analysis
  - Trend discovery
  - Hypothesis generation – literature-based knowledge discovery

Search

- Literature **search** works reasonably well
  - ScienceDirect, Google Scholar, Medline/PubMed, ...



- However, keyword search only returns articles
  - Who has time to sift through hundreds/thousands of abstracts or full papers?

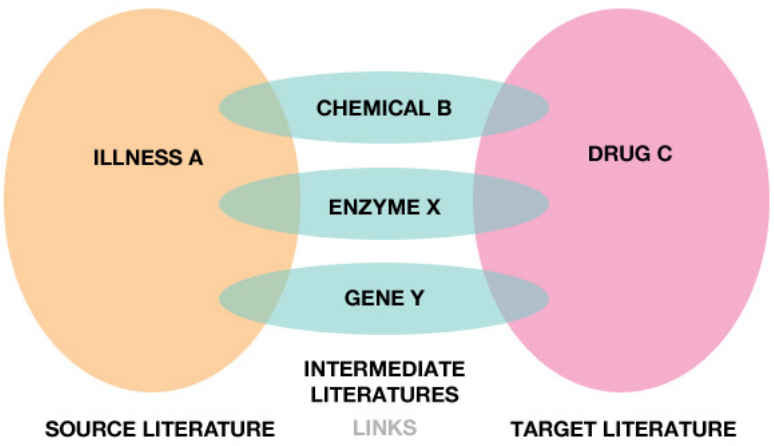
Hypothesis generation

- Two main cognitive tasks
  - Identifying important knowledge pieces
  - Inferring new knowledge from these pieces

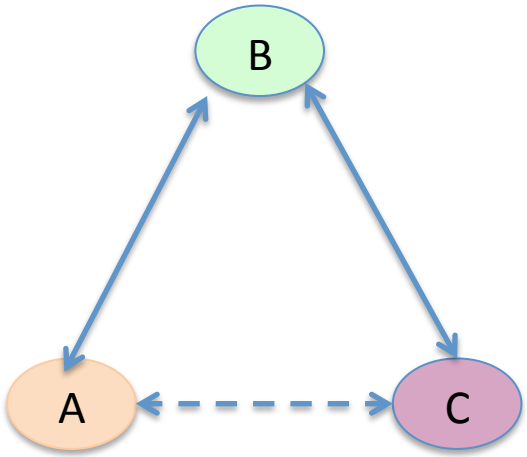
- Focus in this presentation: Identification of knowledge pieces in scientific papers

- Computational method:
  - Literature-based knowledge discovery (LBKD)

LBKD history – Swanson example



(From Wikipedia)

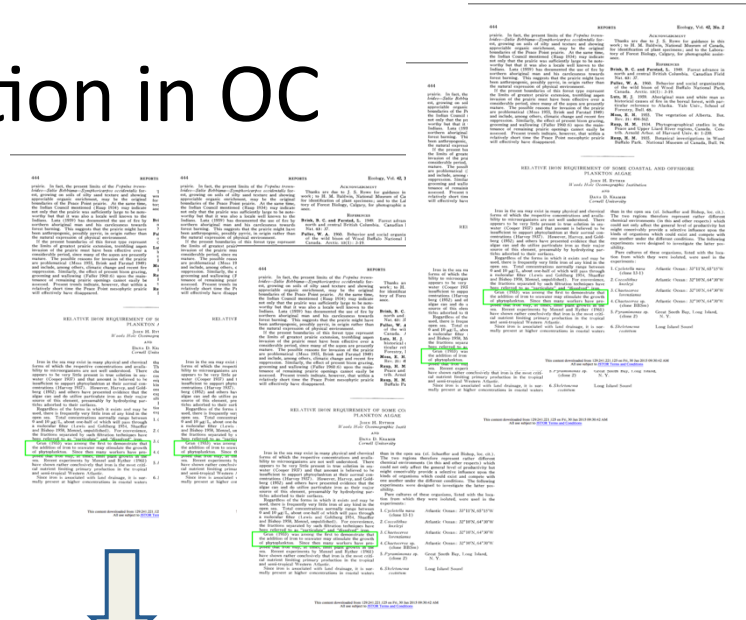


Inference:
A influences B
B influences C
Hence A influences C

1. Relation of spreading depression to the visual scotomata of classical migraine
2. Magnesium in the extracellular cerebral fluid can prevent or terminate spreading depression

3. INFER: migraine ←-→magnesium deficiency

Example: Hypothesis Generation in OC



Identify important knowledge pieces

↑iron → ↑phytoplankton

↑phytoplankton → ↑photosynthesis

↑photosynthesis → ↓CO_2

Infer new knowledge from these pieces

↑iron → ↓CO_2

Text mining for extraction of knowledge pieces

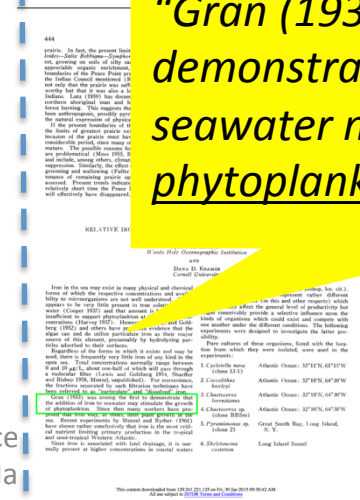
- Text mining deals with identification and extraction of phrases/sentences of interest

- Techniques : natural language processing, information retrieval, machine learning, information extraction, various statistics-based techniques

Design of a text mining system for OC -1

1. Decide what type of knowledge to attend to

Process	Change events	Variables	Relationships between events
Biological Pump	Increase/decrease /change	pH, temperature Chemical compounds Biological species	Causal/correlational



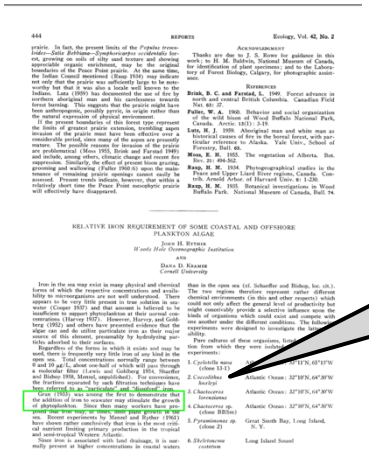
"Gran (1933) was among the first to demonstrate that the addition of iron to seawater may stimulate the growth of phytoplankton."

*.....that the addition of iron to seawater **may stimulate** the growth of phytoplankton."*

↑iron
↑phytoplankton

↑iron → **↑**phytoplankton

Change events



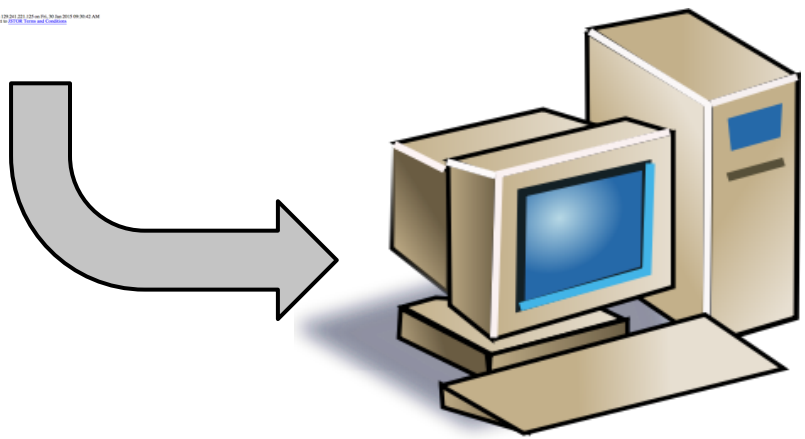
RELATIVE IRON REQUIREMENT OF SOME COASTAL AND OFFSHORE PLANKTON ALGAE

JOHN H. RYTHER
Woods Hole Oceanographic Institution

AND

DAVID D. KEMPER
Cornell University

"Gran (1933) was among the first to demonstrate that the addition of iron to seawater may stimulate the growth of phytoplankton."



- ↑iron
- ↑phytoplankton

Event expressions in natural language

- Same event may be expressed in various ways in natural language
- E.g., “increase”:
  - “Rise in atmospheric CO2 levels…”
  - “…addition of iron…”
  - “elevated value of …..”

- E.g., “decrease”:
  - “…to slow down calcification in corals..”
  - “decreasing temperature…”
  - “…reduced pH value…”

Design of a text mining system for OC- 2

2. *Design and construct Corpus*
   - Decide which disciplines, publishers, journals
     - Ensure sufficient coverage, i.e number and variety of publications
     - Currently 10 K papers from Nature
     - Problems with open access – text mining &sharing rights

3. *Determine text mining subtask(s):*
   - Event extraction
   - Causal/Correlational relationships between events

   - Recognizing Entity mentions
   - Linking to ontologies and generalization of terms

Design of a text mining system for OC - 3

4. Identify the tools to be used

5. Decide the external sources to be used

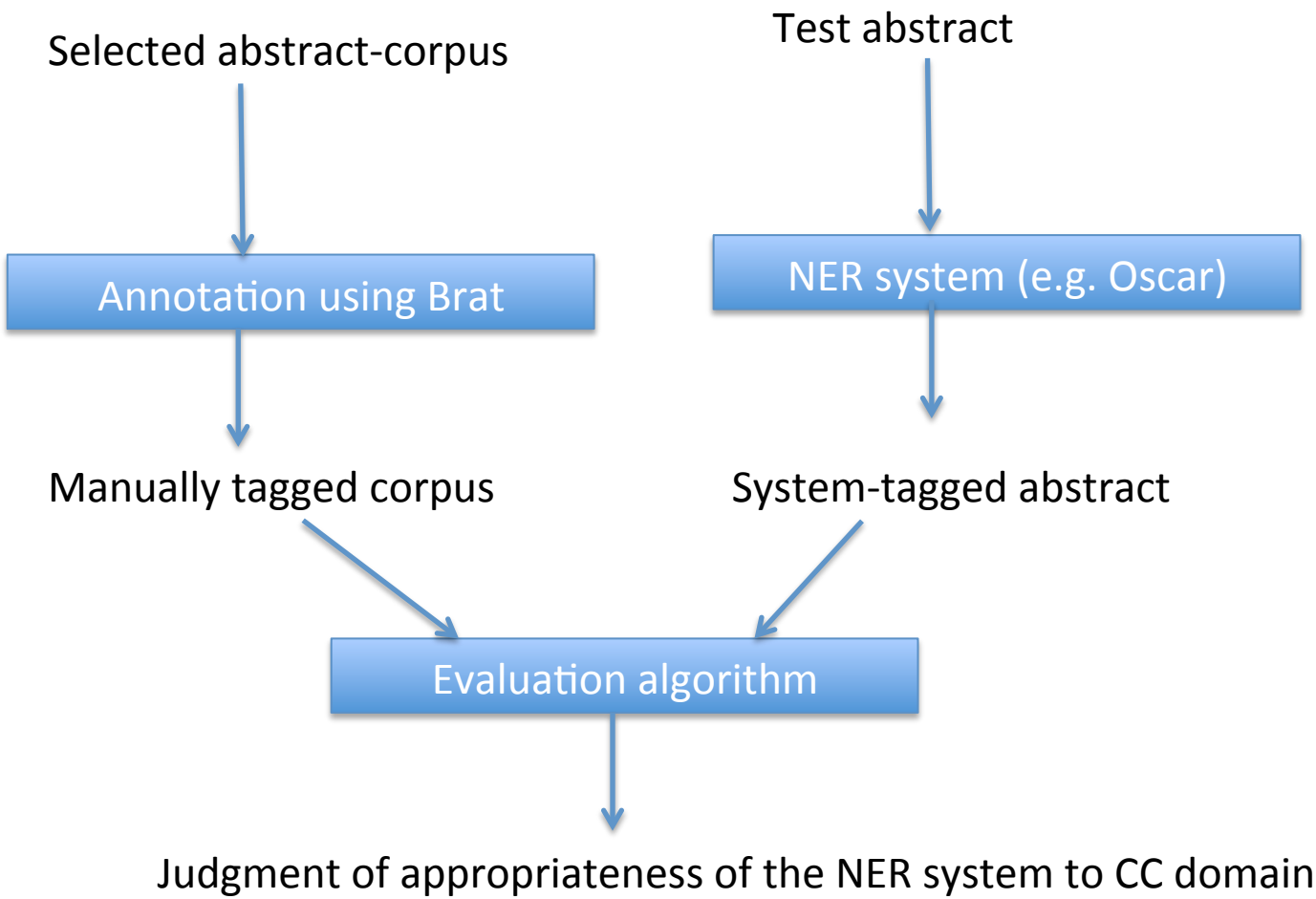
Design of a text mining system for Climate Change domain - 4

- Preliminary yet
- *Our Strategy*: try to map if/which of the existing tools, methods, and external sources developed for other domains (e.g., biomedicine, news text, digital heritage etc) are relevant
- Tools:
  - NLP tools & Annotation tools, e.g., Stanford's NLP, GATE, Brat annotation tool
- External resources
  - Controlled vocabularies, terminologies, thesauruses, ontologies, data bases
  - Examples: dbpedia, Wiki, WordNet, Chebi, Oscar, ChemSpot,, linnaeus2

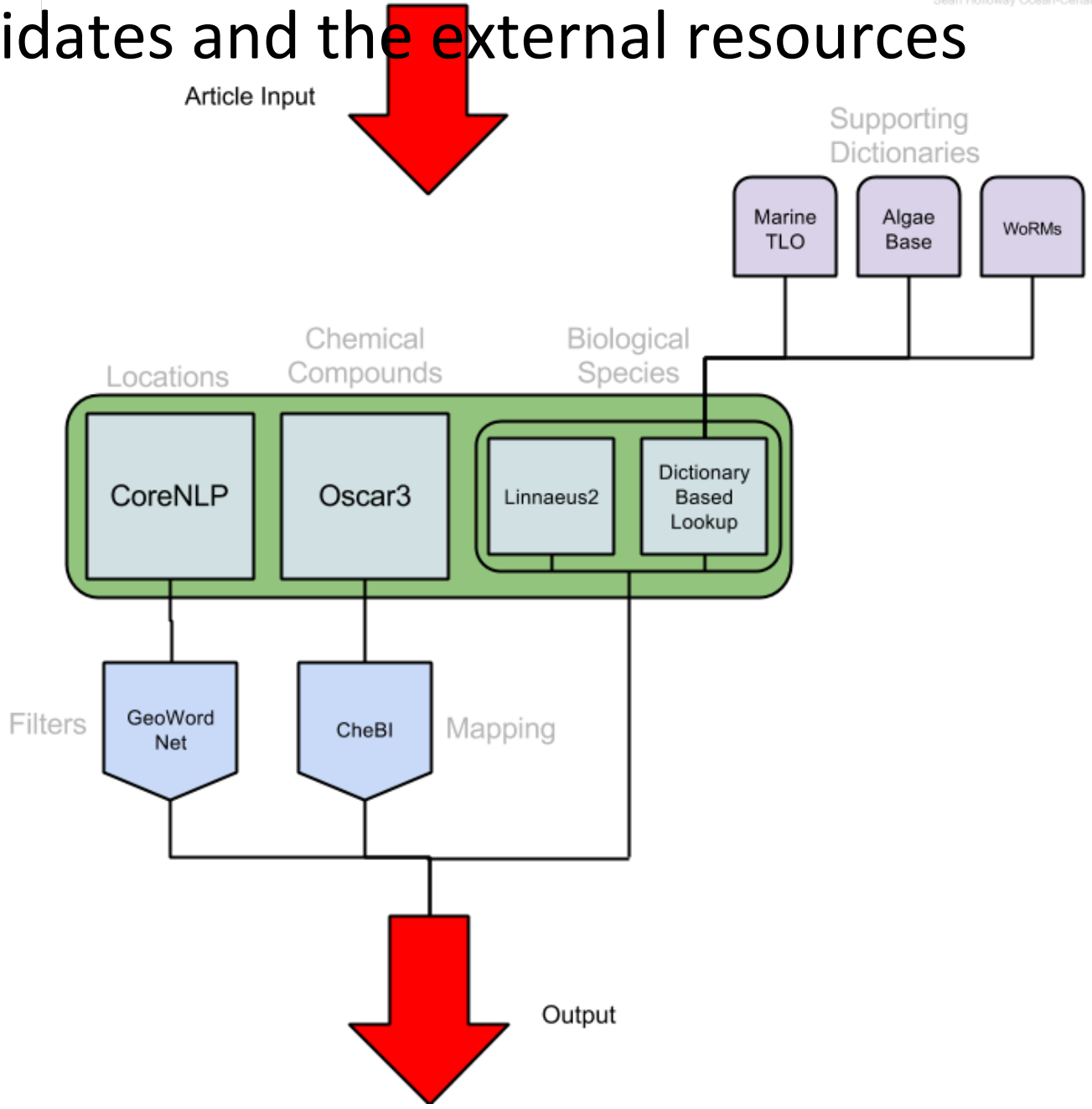
Example: Identify tools &external resources for named-entity-recognition (NER) in Ocean-Certain

- Named entities are the entities of interest
  - Examples in news text: people names, organisations, places
  - Examples in Ocean-certain: chemical compounds, biological species, locations
- A lot of NER systems but mostly built for other domains (e.g, news, humanities or biomedicine)
- Check whether/which existing NER systems can be used for processing papers in the climate change domain
- In particular, we are evaluating :
  - CoreNLP (for geographical locations)
  - Linnaeus2 (species)
  - Oscar3 (chemical compounds)

Evaluation of existing NER tools

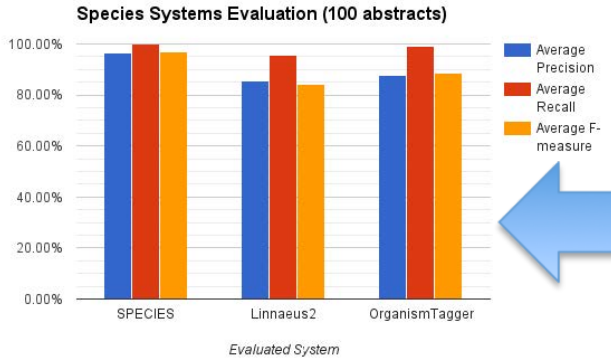


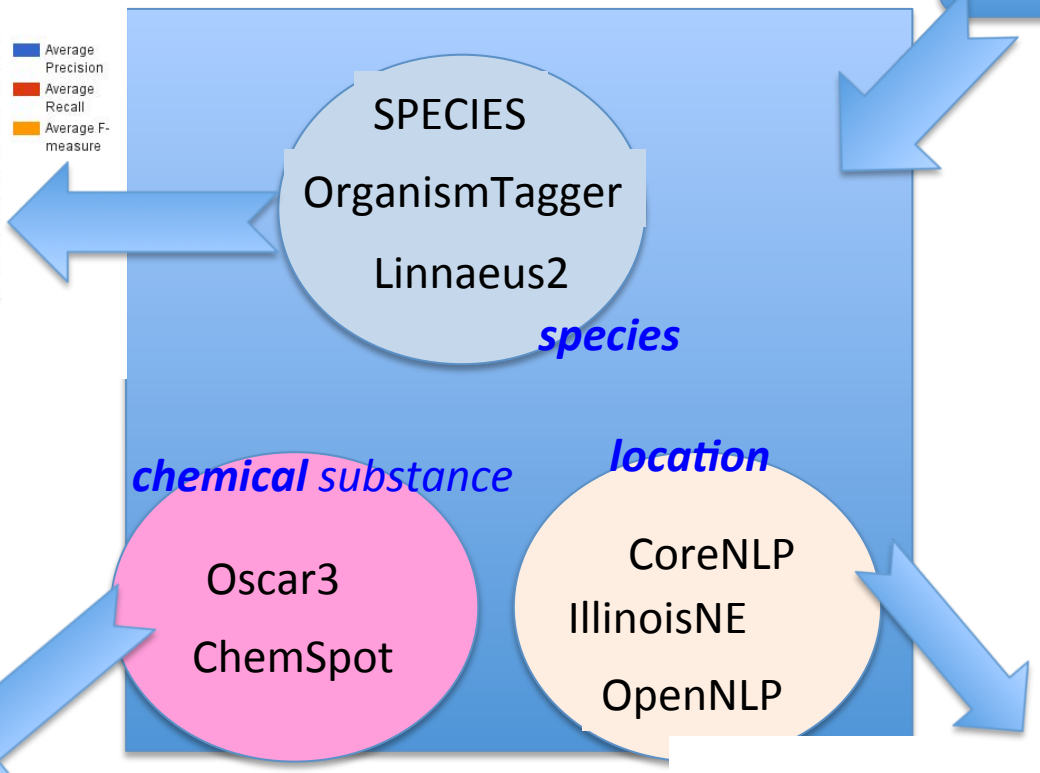
NER candidates and the external resources

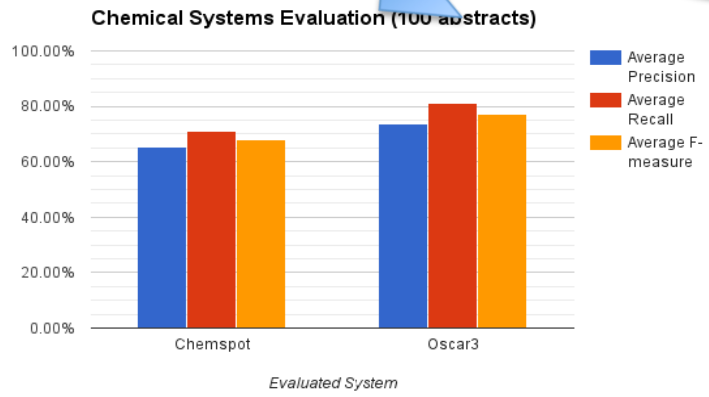


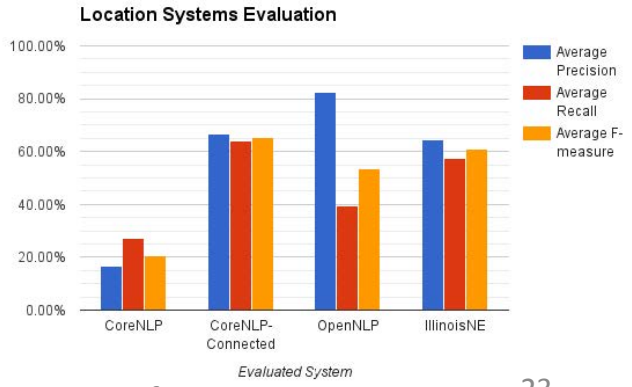
This slide is prepared by Sean Holloway – MSc student(NTNU), 2015

NER System Results









Evaluation experiments are run by Sean Holloway – MSc student, 2015

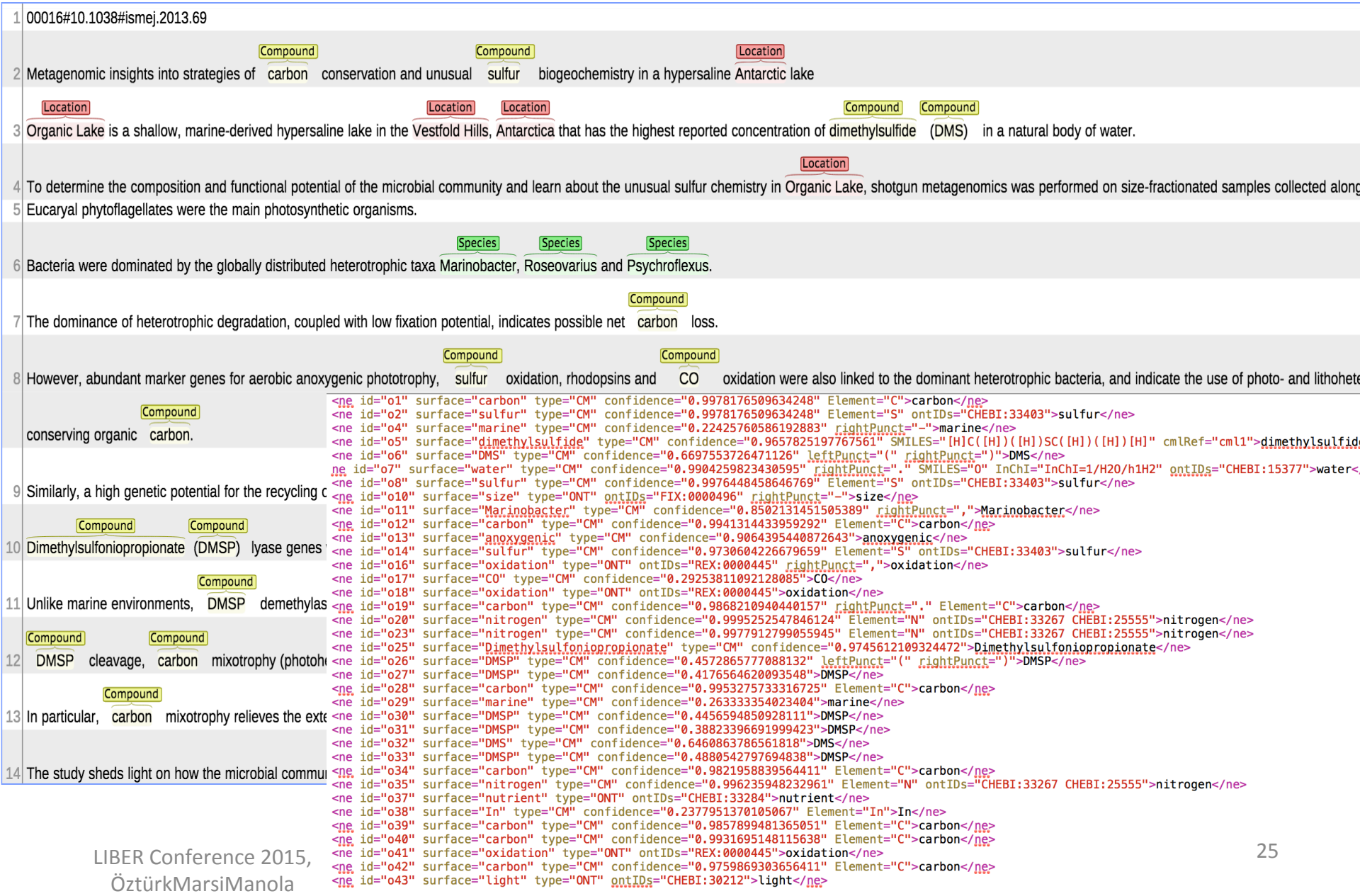
Sharing extended resources?

- Preprocessed scientific papers in machine readable format
  - 10 K full papers from Nature but we cannot share them

- Annotated papers
  - Two types of annotations
    - For Entity recognition
    - For relation and event recognition

- Currently crawling open access (PLOS first) publications- aiming to prepare and share a large volume corpus for CC domain

Annotated gold standard – not shared


00016#10.1038#ismej.2013.69
Metagenomic insights into strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline Antarctic lake
Organic Lake is a shallow, marine-derived hypersaline lake in the Vestfold Hills, Antarctica that has the highest reported concentration of dimethylsulfide (DMS) in a natural body of water.
To determine the composition and functional potential of the microbial community and learn about the unusual sulfur chemistry in Organic Lake, shotgun metagenomics was performed on size-fractionated samples collected along
Eucaryal phytoflagellates were the main photosynthetic organisms.
Bacteria were dominated by the globally distributed heterotrophic taxa Marinobacter, Roseovarius and Psychroflexus.
The dominance of heterotrophic degradation, coupled with low fixation potential, indicates possible net carbon loss.
However, abundant marker genes for aerobic anoxygenic phototrophy, sulfur oxidation, rhodopsins and CO oxidation were also linked to the dominant heterotrophic bacteria, and indicate the use of photo- and lithohete
conserving organic carbon.
Similarly, a high genetic potential for the recycling c
Dimethylsulfoniopropionate (DMSP) lyase genes
Unlike marine environments, DMSP demethylas
DMSP cleavage, carbon mixotrophy (photoh
In particular, carbon mixotrophy relieves the exte
The study sheds light on how the microbial commu
<ne id="o1" surface="carbon" type="CM" confidence="0.9978176509634248" Element="C">carbon</ne>
<ne id="o2" surface="sulfur" type="CM" confidence="0.9978176509634248" Element="S" ontIDs="CHEBI:33403">sulfur</ne>
<ne id="o4" surface="marine" type="CM" confidence="0.22425760586192883" rightPunct="-">marine</ne>
<ne id="o5" surface="dimethylsulfide" type="CM" confidence="0.9657825197767561" SMILES="[H]C([H])([H])SC([H])([H])[H]" cmlRef="cml1">dimethylsulfid
<ne id="o6" surface="DMS" type="CM" confidence="0.6697553726471126" leftPunct="(" rightPunct=")">DMS</ne>
ne id="o7" surface="water" type="CM" confidence="0.9904259823430595" rightPunct="," SMILES="O" InChI="InChI=1/H2O/h1H2" ontIDs="CHEBI:15377">water<
<ne id="o8" surface="sulfur" type="CM" confidence="0.9976448458646769" Element="S" ontIDs="CHEBI:33403">sulfur</ne>
<ne id="o10" surface="size" type="ONT" ontIDs="FIX:0000496" rightPunct="-">size</ne>
<ne id="o11" surface="Marinobacter" type="CM" confidence="0.85021314515105389" rightPunct=",">Marinobacter</ne>
<ne id="o12" surface="carbon" type="CM" confidence="0.9941314433959292" Element="C">carbon</ne>
<ne id="o13" surface="anoxygenic" type="CM" confidence="0.9064395440872643">anoxygenic</ne>
<ne id="o14" surface="sulfur" type="CM" confidence="0.9730604226679659" Element="S" ontIDs="CHEBI:33403">sulfur</ne>
<ne id="o16" surface="oxidation" type="ONT" ontIDs="REX:0000445" rightPunct=",">oxidation</ne>
<ne id="o17" surface="CO" type="CM" confidence="0.29253811092128085">CO</ne>
<ne id="o18" surface="oxidation" type="ONT" ontIDs="REX:0000445">oxidation</ne>
<ne id="o19" surface="carbon" type="CM" confidence="0.9868210940440157" rightPunct="." Element="C">carbon</ne>
<ne id="o20" surface="nitrogen" type="CM" confidence="0.9995252547846124" Element="N" ontIDs="CHEBI:33267 CHEBI:25555">nitrogen</ne>
<ne id="o23" surface="nitrogen" type="CM" confidence="0.9977912799055945" Element="N" ontIDs="CHEBI:33267 CHEBI:25555">nitrogen</ne>
<ne id="o25" surface="Dimethylsulfoniopropionate" type="CM" confidence="0.9745612109324472">Dimethylsulfoniopropionate</ne>
<ne id="o26" surface="DMSP" type="CM" confidence="0.4572865777088132" leftPunct="(" rightPunct=")">DMSP</ne>
<ne id="o27" surface="DMSP" type="CM" confidence="0.4176564620093548">DMSP</ne>
<ne id="o28" surface="carbon" type="CM" confidence="0.9953275733316725" Element="C">carbon</ne>
<ne id="o29" surface="marine" type="CM" confidence="0.263333354023404">marine</ne>
<ne id="o30" surface="DMSP" type="CM" confidence="0.4456594850928111">DMSP</ne>
<ne id="o31" surface="DMSP" type="CM" confidence="0.3882339669199423">DMSP</ne>
<ne id="o32" surface="DMS" type="CM" confidence="0.6460863786561818">DMS</ne>
<ne id="o33" surface="DMSP" type="CM" confidence="0.4880542797694838">DMSP</ne>
<ne id="o34" surface="carbon" type="CM" confidence="0.9821958839564411" Element="C">carbon</ne>
<ne id="o35" surface="nitrogen" type="CM" confidence="0.996235948232961" Element="N" ontIDs="CHEBI:33267 CHEBI:25555">nitrogen</ne>
<ne id="o37" surface="nutrient" type="ONT" ontIDs="CHEBI:33284">nutrient</ne>
<ne id="o38" surface="In" type="CM" confidence="0.2377951370105067" Element="In">In</ne>
<ne id="o39" surface="carbon" type="CM" confidence="0.9857899481365051" Element="C">carbon</ne>
<ne id="o40" surface="carbon" type="CM" confidence="0.9931695148115638" Element="C">carbon</ne>
<ne id="o41" surface="oxidation" type="ONT" ontIDs="REX:0000445">oxidation</ne>
<ne id="o42" surface="carbon" type="CM" confidence="0.9759869303656411" Element="C">carbon</ne>
<ne id="o43" surface="light" type="ONT" ontIDs="CHEBI:30212">light</ne>

Summary

- Text mining as a support to scientific discovery
  - The preliminary results promising for extraction of
    - entities/variables,
    - change events and
    - relations between events

Conclusion

- Some of the existing tools (general and specific to other domains) may be useful
- However, we need to adapt and extend these for the CC domain
- Corpus is an important problem
  - We cannot share the preprocessed and annotated corpus we create
  - We would not possible use others' resources because of the same reasons
  - Repetition of task (inefficient use of money and time)
  - Slows down our own work as well as the knowledge discovery research in CC domain, because of

Future work

- Preparation of a corpus for the CC domain – larger volume and sharable
- Currently working on automated crawling&preprocessing that fits to variations in various publishers
- We need more annotation, meaning more people, more funding
- Planning to apply EU and Norwegian Research council for funding
- Organizing a workshop (in connection with the OC project) to gather people working in text mining in Earth science
- Want/need collaboration with other people/universities

- The work presented here is partly reported in :

Marsi, Erwin; Özturk, Pinar; Aamot, Elias; Sizov, Gleb Valerjevich; Ardelan, Murat Van. (2014) Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*

Demo

- http://www.idi.ntnu.no/~emarsi/ocwp1/chavarex