Assembling Ontologies for the Discovery of New Materials

NSF-HDR# 1940239

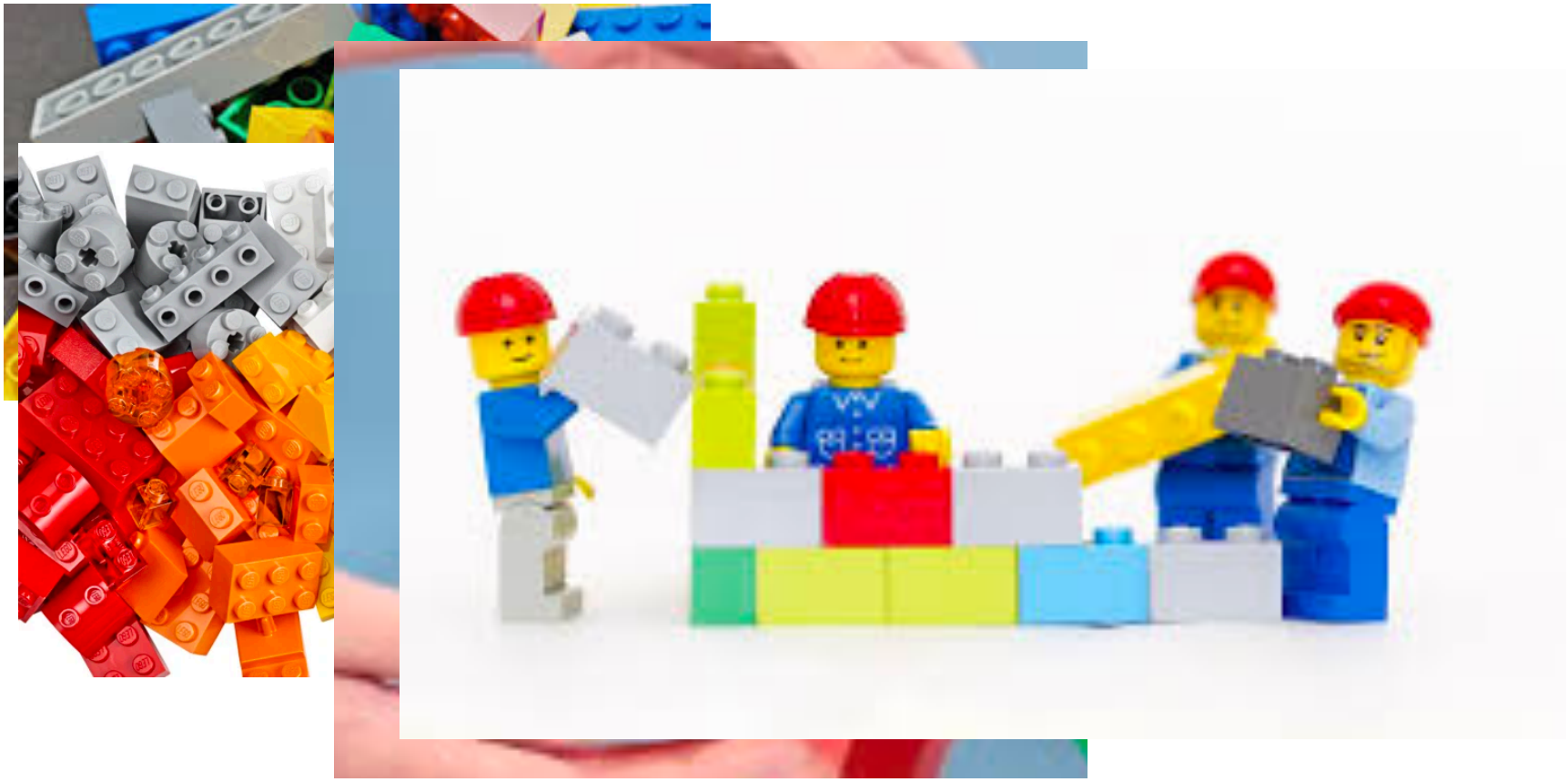
NKOS: Networked Knowledge Organization Systems Workshop
September 9, 2020

Jane Greenberg, Xintong Zhao, Xiaohua Tony Hu,
Metadata Research Center, Drexel University
Vanessa Meschke, *Eric Toberer*, Colorado School of Mines
Jordan Cox, Steven Lopez, Northeastern University
Semion K. Saikin, Kebotix
Remco Chang, Tufts University
Roman Garnett, Washington University, St. Louis

Outline

- Background and Motivation
- Current Progress, Workplan
- Conclusion/Future Plans
- Q&A

Assembling ontologies



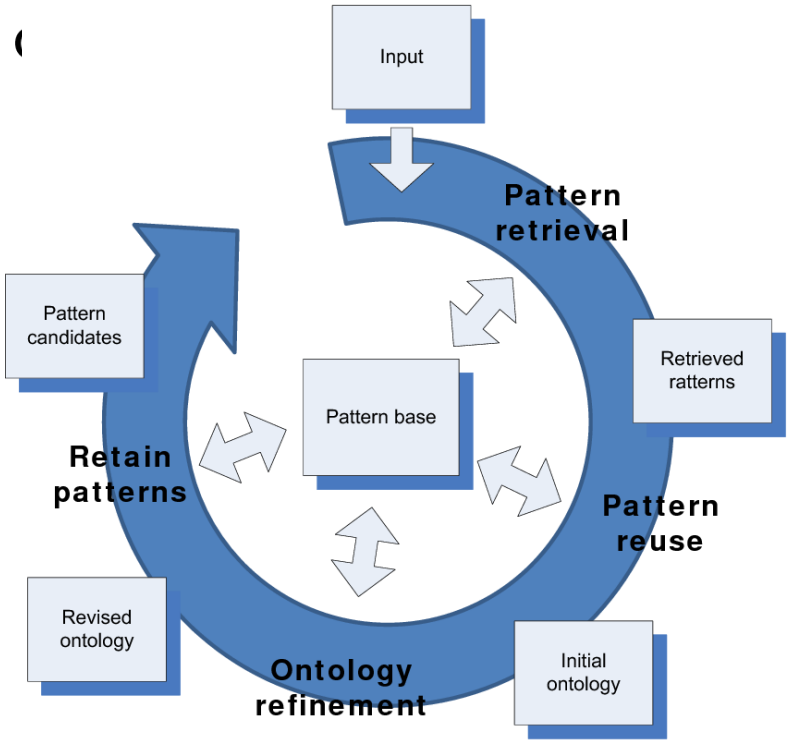
Oversight, development, and commitment

- **Institutional:** Library of Congress, USGS, FAO(!)
- **Community:** Phenoscape (https://wiki.phenoscape.org/wiki/Ontologies)
  - Database of phenotype data for teleost fish, set of ontologies, Vo-camp

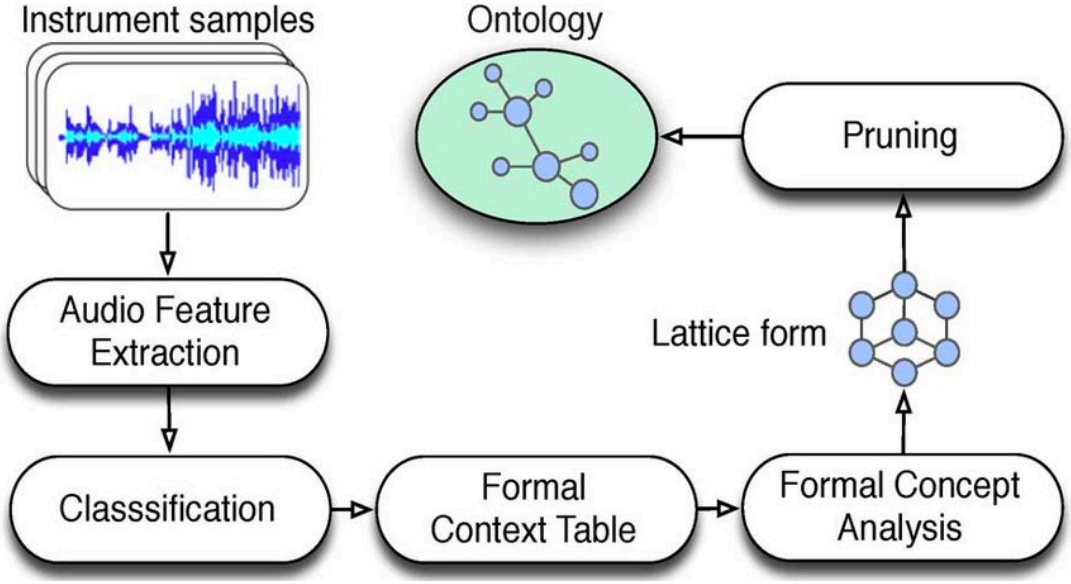
Uberon Anatomy Ontology	**Taxonomy Ontologies**
Anatomy ontologies merged with Uberon • Vertebrate Skeletal Anatomy Ontology (VSAO) • Teleost Anatomy Ontology (TAO) • Amphibian Anatomy Ontology (AAO) • Mouse Adult Gross Anatomy (MA) • Xenopus Anatomy Ontology (XAO) • Zebrafish Anatomical Ontology (ZFA)	• Vertebrate Taxonomy Ontology (VTO) • Teleost Taxonomy Ontology (TTO) • Taxonomic Rank Vocabulary (TAXRANK) • Fish Collection Codes Vocabulary • Amphibian Taxonomy Ontology (ATO) • Additional documentation

- **Resources:** Bioportal, Ontobee, OBO Foundry, LC linked data services, Fairsharing.org, FAO

Method/Approach: Manual (Protégé), semi-automatic, and automatic (NLP, Named Entity Recognition, and RE-Relation Extraction)



Blomqvist (2009). Semi-automatic Ontology Construction based on Patterns



Kolozali, Şefki & Barthet, Mathieu & Fazekas, György & Sandler, Mark. (2013). Automatic Ontology Generation for Musical Instruments Based on Audio Analysis. Audio, Speech, and Language Processing, IEEE Transactions...

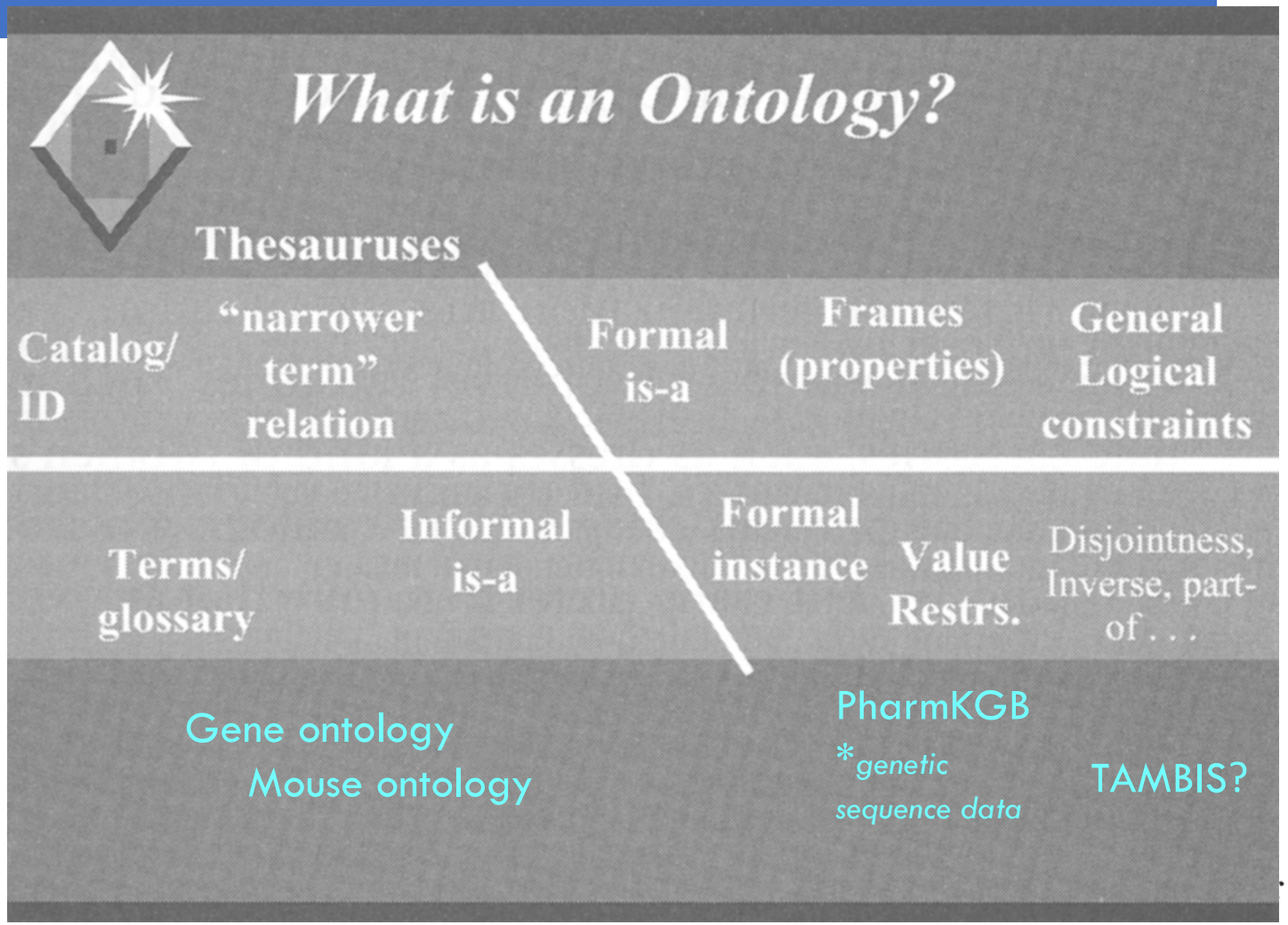
(McGuinness, D. L. (2003). Ontologies Come of Age. In Fensel, et al, *Spinning the Semantic Web.* (Cambridge, MIT Press)

Why this review?

Status of ontologies

- Where we can go to better support development of knowledge graphs, deep learning/AI

**our case, Materials Science



Materials Science

- NSF-HDR: *Accelerating the Discovery of Electronic Materials through Human-Computer Active Search*
- Interdisciplinary field, engineering, chemistry, and physics
- Aim: Discovery of new materials; develop functional, benign, less costly materials
- Study properties of materials, observation and measurement

Property	Description	Observations
Thermal conductivity	Ability to conduct heat	Aluminum conducts heat at a much higher rate (more rapidly) than than of bronze or steel
Buoyancy	Ability to float in water	Sea water: Polyethylene terephthalate (water bottles) will sink, and polypropylene (water bottle caps) will float at least for a time period, due to density

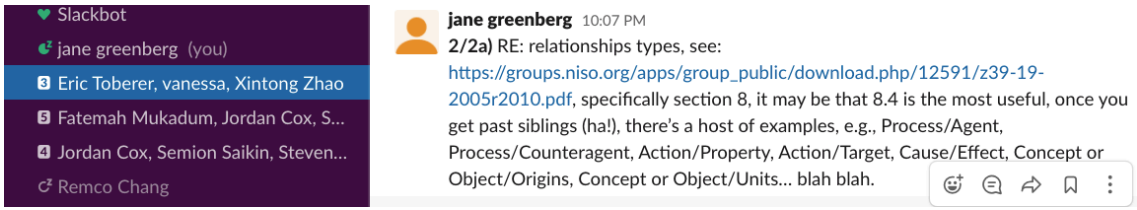
- Relationships between material entities is ontological
  - Steel/stainless steel; OR steel ←→ iron and carbon

- Focus: Thermoelectric and photocatalytic materials
- Goal: Undiscovered public knowledge (Swanson, 1986), find knowledge buried in research literature.
- Enable: Prediction, of materials synthesis and characterization
- High-efficiency thermoelectric materials - heat transformation to electrical power (energy conversions). (Colorado/MINES)
- Earth-abundant light-responsive catalysts - less costly to store solar energy (Northeastern U.)

Interest in/value of sophisticated, more granular relationships with materials research

▪ Process/Agent ▪ Process/Counteragent ▪ Action/Property ▪ Action/Target	▪ Cause/Effect ▪ Concept or Object/Property ▪ Concept or Object/Units ▪ Raw Material/Product

Table 2: Selected associative relationships from ANSI/NISO Z39.10-2005(R2010)



Challenge → solution

- Challenge: The volume of academic articles is too large for researcher to fully read even a portion of papers in their lifetime;
  - The use of time becomes inefficient
  - Hard to accurately retrieve needed information in short time
- Ontologies as a solution
  - Material properties, processing methods and structures can support discovery
- Materials Science (Ashino (2010); Cheung (2009)); inspired also by biomedicine

- Property - Structure (atom → molecules)/or Structure - Property
- Structure - Process

- Property - Structure (atom → molecules)/or Structure - Property
- Structure - Process

These materials were used to form thin transparent films by a spin-coating technique. Relation (RE): thin film - spin-coating ; structure-process

Then the ability of thin hybrid films to reversible trans-cis photoisomerization under illumination was investigated using ellipsometry and UV-Vis spectroscopy. RE: thin-film - reversible trans-cis photoisomerization ; structure-property

The reversible changes of refractive index of the films under illumination were in the range of 0.005-0.056.

RE:refractive index - thin film; property-structure

Refractive index - 0.005-0.056 ; has-value

The maximum absorption of these materials was located at 462-486 nm. HARD

RE:thin-film - absorption; structure-property;

Absorption - 462-486nm ; has-value

Moreover, the organic-inorganic azobenzene materials were used to form nanofibers by electrospinning using various parameters of the process.

nanofibers - electrospinning ; structure-process

The microstructure of electrospun fibers depended on sols properties (e.g. concentration and viscosity of the sols) and process conditions (e.g. the applied voltage, temperature or type of the collector) at ambient conditions.

electrospun fibers - sols properties ; structure - process

Electrospun fibers - process conditions; structure - process

MATScholar (NER-Named Entity Recognition)

Lawrence Berkeley National Lab: https://www.matscholar.com/

Extracted Entity Tags:

In this study , we attempted to reduce **firing** voltage of **ac - PDPs** by **alloying** **MgO** **electron emission** material with **OZn** . this approach was aimed to reduce **band gap energy** of **MgO** by the **alloying** and thereby promote the auger neutralization reaction of xe+ ions on **MgO** **surface** . **pellets** were prepared by **sintering** **MgO** and **OZn** **powder** mixture at 1300 Â° C for 8 h under nitrogen atmosphere . test panels with such **alloyed** **MgO** **films** showed significantly reduced **firing** voltages , especially when the discharge gas is of high **Xe content** . these results represent a new way of approaching in the development of **electron emission** materials for **ac - PDPs** .

Labels:

material **application** **property** **phase** **synthesis** **characterization**

descriptor **property value** **property unit**

Work activity underway (NER)

Inorg./Organic	Label, description
▪ MAT = MOL	▪ Molecules or fragments
▪ SPL = None	▪ Symmetry phase label/NA
▪ None = POLY	▪ NA/Polymers, general organic matter
▪ DSC = DSC	▪ Sample descriptors
▪ PRO = PRO	▪ Material Property
▪ APL = APL	▪ Material application
▪ SMT = RXN	▪ Synthesis method/Reaxtion
▪ CMT = CCMT	▪ Characterization method

NA = not applicable.

Overall workflow/plan

Gathering data (abstracts)

NER and RE (Use relation extraction to construct knowledge base)

Develop ontologies to underlie knowledge graphs

Help researchers locate key information from textual data

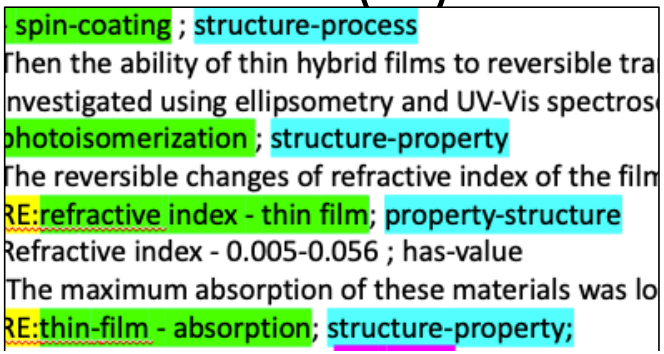
Driving idea....Ideal outcome

Input: "Hey system, what are common materials that have thermoelectric property?"

Output: return integrated information containing the N most frequent materials + related properties/applications + list of papers

Involved Methods

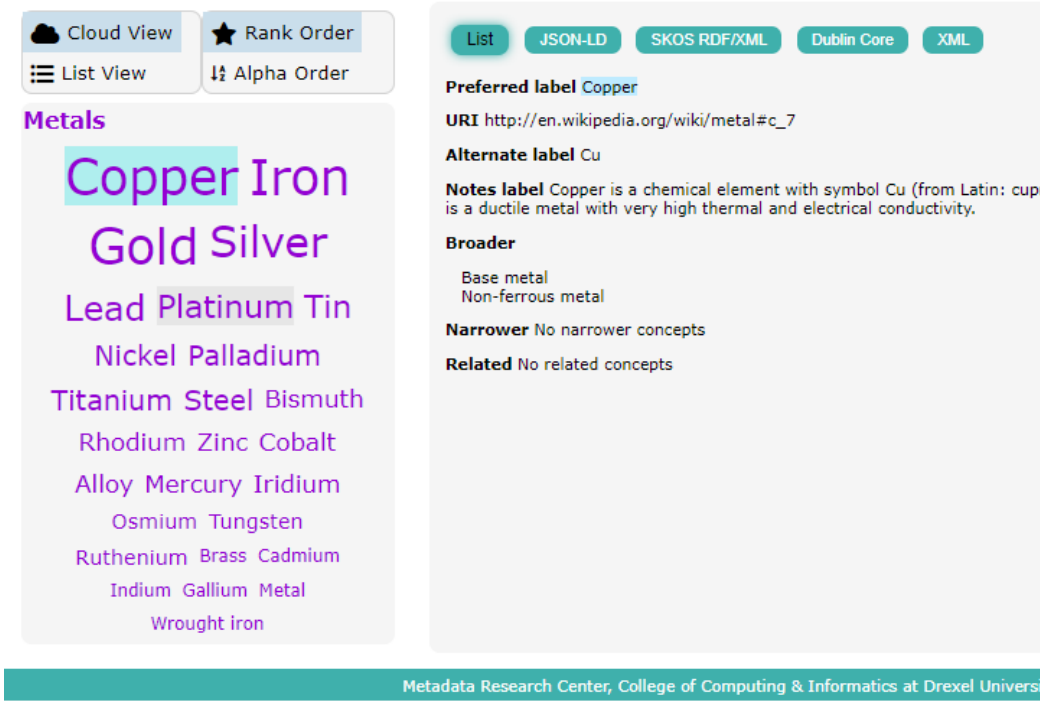
- Named Entity Recognition (NER)
  - NER is a subtask of Information extraction (IE) that can support semantic labeling. NER involves deep learning to detect named entities and their type in a sentence.
  - Since it's supervised learning, a large training set is required
- Relation Extraction (RE) follows next



- **Traditional Machine Learning Algorithms for Keyword Extraction**
  - The process involves automatic indexing to extract key terms from a document; followed by matching these initial results to terms encoded in a knowledge structure, such as an ontology.
  - There are multiple algorithms, we take the **RAKE** (Rapid Keyword Extraction) as an example
  - Un-supervised learning, which does not require training set.

Methods and Procedures (Cont'd)

- HIVE-4-MAT: we design it as a linked data automatic indexing application, and it is still under construction; it builds off the original HIVE (Zhang et al., 2015) system developed earlier at Metadata Research Center of Drexel University.
  - Original HIVE: http://hive2.cci.drexel.edu:8080/



(Zhang et al., 2015)

Conclusion and next steps

- New work for materials science, can learn biomedicine
- Great team, a good bit of encoding, encouraging results (Xintong Zhao, 2020, JCDL workshop, **Organizing Data, Information, and Knowledge in Big Data Environments**)
- Continue to expand the text data from inorganic materials (MATScholar) to both organic and inorganic materials.
  - Also, keeping track more specifically of thermoelectric and photocatalytic materials
- Continue to create a dataset for relation extraction
- Evaluate and refine
- Run tests with questions, e.g. the ideal....(t.b.d.)

References

- Ashino, T. (2010). Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. Data Science Journal, 9, 54-61. doi:10.2481/dsj.008-041
- Cheung, K., Hunter J., and Drennan, J. (2009). "MatSeek: An Ontology-Based Federated Search Interface for Materials Scientists. IEEE Intelligent Systems, 24(1): 10.1109/MIS.2009.13.
- Swanson, D. R. (1986). Undiscovered public knowledge. The Library Quarterly, 56(2), 103-118.
- Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., Ceder, G., & Jain, A. (2019). Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of Chemical Information and Modeling, 59(9), 3692–3702. https://doi.org/10.1021/acs.jcim.9b00470
- Wei, C. H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., ... & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database
- **Zhao, X., Greenberg, J., Hu, X., Meschke, V., & Toberer, E. (2020, August 1-5). Scholarly Big Data: Computational Approaches to Semantic Labeling in Materials Science. Organizing Data, Information, and Knowledge in Big Data Environments workshop. ACM/IEEE Joint Conference on Digital Libraries, Wuhan, Hubei, P. R. China: Paper and slides: https://cci.drexel.edu/mrc/publications/
- Zhang, Y., Greenberg, J., Ogletree, A., and Tucker, G. (2015). Advancing Materials Science Semantic Metadata via HIVE. International Conference on Dublin Core and Metadata Applications, p. 209-211: https://dcpapers.dublincore.org/pubs/article/view/3783