Working Group 4: NLP, Published Literature & Data Mining

Juan de Pablo, Ankit Agrawal,
Kenneth Kroenlein, Alden Dima,
Adele Peskin, Roselyne Tchoua

NIST National Institute of Standards and Technology U.S. Department of Commerce

THE UNIVERSITY OF CHICAGO

NORTHWESTERN UNIVERSITY

Argonne NATIONAL LABORATORY

CHiMaD

Significance of WG's Focus

- Explosion of available data in literature that must be made available, organized, analyzed, utilized
- Provide design tools that at continuously up to date, verify data, avoid duplication of efforts
- Identify requirements to created fully automated data bases for new classes of materials
- Could impact multiple disciplines and provide a basis for cross fertilization

Summary of WG's Goals

- A fully automated data base for polymer alloys that includes phase diagram (x vs. T), Tg, DH, G', G"
- A fully automated data base for polymer composites that includes mixing, interfacial energy, G', G"
- Gain perspective of how data are generated, interpreted, used, in different fields and identify best practices
- Understand challenges when individual fields are looked at from a multidisciplinary perspective
- Detect mistakes and learn how to avoid them

Technical Requirements/Needs

- Access to journals
- Where to we get the data set from (e.g. publications). Use same data set (structured test) and apply different methods (different groups) and evaluate outcomes
- Can we get journals to provide structured test (e.g. HTML)
- Identify relevance (classification)
- Extract data from relevant publications
- How to handle text, tables, figures
- Output should be structured text (xml format) suitable for the curator, processable (e.g. data curator)

Collaborations/Synergies

- Other Working Groups? *{Who in CHiMaD might use/need parts of the WG focus?}*
- *DSA Copolymer working group*
- *Polymer composite working group*
- *Charge driven assembly working group*
- *General applicability to all working groups in need of new data bases – new methods and tools of general use*

Collaborations/Synergies

- Others in NIST? (*How does/might the WG focus connect with NIST materials informatics efforts?*)
- *Use knowledge from TRC and apply it to challenge problems*
- *Lessons learned will benefit TRC*
- *Other groups at NIST, e.g. Carelyn Campbell, Bob Hanisch, Polymers Division*