**Use Cases from NBD(NIST Big Data) Requirements WG**

<http://bigdatawg.nist.gov/home.php>

**NBD(NIST Big Data) Requirements WG Use Case Template Aug 22 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Materials Data | |
| **Vertical (area)** | | Manufacturing, Materials Research | |
| **Author/Company/Email** | | John Rumble, R&R Data Services; jumbleusa@earthlink.net | |
| **Actors/Stakeholders and their roles and responsibilities** | | Product Designers (Inputters of materials data in CAE)  Materials Researchers (Generators of materials data; users in some cases)  Materials Testers (Generators of materials data; standards developers)  Data distributors ( Providers of access to materials, often for profit) | |
| **Goals** | | Broaden accessibility, quality, and usability; Overcome proprietary barriers to sharing materials data; Create sufficiently large repositories of materials data to support discovery | |
| **Use Case Description** | | Every physical product is made from a material that has been selected for its properties, cost, and availability. This translates into hundreds of billion dollars of material decisions made every year.  In addition, as the Materials Genome Initiative has so effectively pointed out, the adoption of new materials normally takes decades (two to three) rather than a small number of years, in part because data on new materials is not easily available.  All actors within the materials life cycle today have access to very limited quantities of materials data, thereby resulting in materials-related decision that are non-optimal, inefficient, and costly. While the Materials Genome Initiative is addressing one major and important aspect of the issue, namely the fundamental materials data necessary to design and test materials computationally, the issues related to physical measurements on physical materials ( from basic structural and thermal properties to complex performance properties to properties of novel (nanoscale materials) are not being addressed systematically, broadly (cross-discipline and internationally), or effectively (virtually no materials data meetings, standards groups, or dedicated funded programs).  One of the greatest challenges that Big Data approaches can address is predicting the performance of real materials (gram to ton quantities) starting at the atomistic, nanometer, and/or micrometer level of description.  As a result of the above considerations, decisions about materials usage are unnecessarily conservative, often based on older rather than newer materials R&D data, and not taking advantage of advances in modeling and simulations. Materials informatics is an area in which the new tools of data science can have major impact. | |
| **Current**  **Solutions** | **Compute(System)** | | None |
| **Storage** | | Widely dispersed with many barriers to access |
| **Networking** | | Virtually none |
| **Software** | | Narrow approaches based on national programs (Japan, Korea, and China), applications (EU Nuclear program), proprietary solutions (Granta, etc.) |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Extremely distributed with data repositories existing only for a very few fundamental properties |
| **Volume (size)** | | **It is has been estimated (in the 1980s) that there were over 500,000 commercial materials made in the last fifty years. The last three decades has seen large growth in that number.** |
| **Velocity**  **(e.g. real time)** | | **Computer-designed and theoretically design materials (e.g., nanomaterials) are growing over time** |
| **Variety**  **(multiple datasets, mashup)** | | **Many data sets and virtually no standards for mashups** |
| **Variability (rate of change)** | | **Materials are changing all the time, and new materials data are constantly being generated to describe the new materials** |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | **More complex material properties can require many (100s?) of independent variables to describe accurately. Virtually no activity no exists that is trying to identify and systematize the collection of these variables to create robust data sets.** |
| **Visualization** | | **Important for materials discovery. Potentially important to understand the dependency of properties on the many independent variables. Virtually unaddressed.** |
| **Data Quality** | | **Except for fundamental data on the structural and thermal properties, data quality is poor or unknown. See Munro’s NIST Standard Practice Guide.** |
| **Data Types** | | **Numbers, graphical, images** |
| **Data Analytics** | | **Empirical and narrow in scope** |
| **Big Data Specific Challenges (Gaps)** | | 1. Establishing materials data repositories beyond the existing ones that focus on fundamental data 2. Developing internationally-accepted data recording standards that can be used by a very diverse materials community, including developers materials test standards (such as ASTM and ISO), testing companies, materials producers, and R&D labs 3. Tools and procedures to help organizations wishing to deposit proprietary materials in data repositories to mask proprietary information, yet to maintain the usability of data 4. Multi-variable materials data visualization tools, in which the number of variables can be quite high | |
| **Big Data Specific Challenges in Mobility** | | Not important at this time | |
| **Security & Privacy**  **Requirements** | | Proprietary nature of many data very sensitive. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Development of standards; development of large scale repositories; involving industrial users; integration with CAE (don’t underestimate the difficulty of this – materials people are generally not as computer savvy as chemists, bioinformatics people, and engineers) | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

**Note: No proprietary or confidential information should be included**