**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Enabling Face-Book like Semantic Graph-search on Scientific Chemical and Text-based Data | |
| **Vertical (area)** | | Management of Information from Research Articles | |
| **Author/Company/Email** | | bhat@nist.gov | |
| **Actors/Stakeholders and their roles and responsibilities** | | Chemical structures, Protein Data Bank, Material Genome Project, Open-GOV initiative, Semantic Web, Integrated Data-graphs, Scientific social media | |
| **Goals** | | Establish infrastructure, terminology and semantic data-graphs to annotate and present technology information using ‘root’ and rule-based methods used primarily by some Indo-European languages like Sanskrit and Latin. | |
| **Use Case Description** | | * **Social media hype**   + Internet and social media play a significant role in modern information exchange. Every day most of us use social-media both to distribute and receive information. Two of the special features of many social media like Face-Book are     - the community is both data-providers and data-users     - they store information in a pre-defined ‘data-shelf’ of a data-graph     - Their core infrastructure for managing information is reasonably language free * **What this has to do with managing scientific information?**   During the last few decades science has truly evolved to become a community activity involving every country and almost every household. We routinely ‘tune-in’ to internet resources to share and seek scientific information.   * **What are the challenges in creating social media for science**   + Creating a social media of scientific information needs an infrastructure where many scientists from various parts of the world can participate and deposit results of their experiment. Some of the issues that one has to resolve prior to establishing a scientific social media are:     - How to minimize challenges related to local language and its grammar?     - How to determining the ‘data-graph’ to place an information in an intuitive way without knowing too much about the data management?     - How to find relevant scientific data without spending too much time on the internet?   **Approach:** Most languages and more so Sanskrit and Latin use a novel ‘root’-based method to facilitate the creation of on-demand, discriminating words to define concepts. Some such examples from English are Bio-logy, Bio-chemistry. Youga, Yogi, Yogendra, Yogesh are examples from Sanskrit. Genocide is an example from Latin. These words are created on-demand based on best-practice terms and their capability to serve as node in a discriminating data-graph with self-explained meaning. | |
| **Current**  **Solutions** | **Compute(System)** | | Cloud for the participation of community |
| **Storage** | | Requires expandable on-demand based resource that is suitable for global users location and requirements |
| **Networking** | | Needs good network for the community participation |
| **Software** | | Good database tools and servers for data-graph manipulation are needed |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed resource with a limited centralized capability |
| **Volume (size)** | | **Undetermined. May be few terabytes at the beginning** |
| **Velocity**  **(e.g. real time)** | | **Evolving with time to accommodate new best-practices** |
| **Variety**  **(multiple datasets, mashup)** | | **Wildly varying depending on the types available technological information** |
| **Variability (rate of change)** | | **Data-graphs are likely to change in time based on customer preferences and best-practices** |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | **Technological information is likely to be stable and robust** |
| **Visualization** | | **Efficient data-graph based visualization is needed** |
| **Data Quality** | | **Expected to be good** |
| **Data Types** | | **All data types, image to text, structures to protein sequence** |
| **Data Analytics** | | **Data-graphs is expected to provide robust data-analysis methods** |
| **Big Data Specific Challenges (Gaps)** | | This is a community effort similar to many social media. Providing a robust, scalable, on-demand infrastructures in a manner that is use-case and user-friendly is a real-challenge by any existing conventional methods | |
| **Big Data Specific Challenges in Mobility** | | A community access is required for the data and thus it has to be media and location independent and thus requires high mobility too. | |
| **Security & Privacy**  **Requirements** | | None since the effort is initially focused on publicly accessible data provided by open-platform projects like open-gov, MGI and protein data bank. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | This effort includes many local and networked resources. Developing an infrastructure to automatically integrate information from all these resources using data-graphs is a challenge that we are trying to solve. | |
| **More Information (URLs)** | | <http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php>  <http://xpdb.nist.gov/chemblast/pdb.pl>  <http://xpdb.nist.gov/chemblast/pdb.pl> | |
| **Note:** <additional comments> | | | |

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