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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Crowd Sourcing in the Humanities as Source for Big and Dynamic Data | |
| **Vertical (area)** | | Humanities, Social Sciences | |
| **Author/Company/Email** | | Sebastian Drude <Sebastian.Drude@mpi.nl>, Max Planck Institute for Psycholinguistics (MPI) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Scientists (Sociologists, Psychologists, Linguists, Politic Scientists, Historians, etc.), data managers and analysts, data archives  The general public as data providers and participants | |
| **Goals** | | Capture information (manually entered, recorded multimedia, reaction times, pictures, sensor information) from many individuals and their devices.  Thus capture wide ranging individual, social, cultural and linguistic variation among several dimensions (space, social space, time). | |
| **Use Case Description** | | Many different possible use cases: get recordings of language usage (words, sentences, meaning descriptions, etc.), answers to surveys, info on cultural facts, transcriptions of pictures and texts -- correlate these with other phenomena, detect new cultural practices, behavior, values and believes, discover individual variation | |
| **Current**  **Solutions** | **Compute(System)** | | Individual systems for manual data collection (mostly Websites) |
| **Storage** | | Traditional servers |
| **Networking** | | barely used other than for data entry via web |
| **Software** | | XML technology, traditional relational databases for storing pictures, not much multi-media yet. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed, individual contributors via webpages and mobile devices |
| **Volume (size)** | | Depends dramatically, from hundreds to millions of data records.  Depending on data-type: from gigabytes (text, surveys, experiment values) to hundreds of terabytes (multimedia) |
| **Velocity**  **(e.g. real time)** | | Depends very much on project: dozens to thousands of new data records per day  Data has to be analyzed incrementally. |
| **Variety**  **(multiple datasets, mashup)** | | so far mostly homogeneous small data sets; expected large distributed heterogeneous datasets which have to be archived as primary data |
| **Variability (rate of change)** | | Data structure and content of collections are changing during data lifecycle.  There is no critical variation of data producing speed, or runtime characteristics variations. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Noisy data is possible, unreliable metadata, identification and pre-selection of appropriate data |
| **Visualization** | | important for interpretation, no special visualization techniques |
| **Data Quality** | | validation is necessary; quality of recordings, quality of content, spam |
| **Data Types** | | individual data records (survey answers, reaction times);  text (e.g., comments, transcriptions,…);  multi-media (pictures, audio, video) |
| **Data Analytics** | | pattern recognition of all kind (e.g., speech recognition, automatic A&V analysis, cultural patterns), identification of structures (lexical units, linguistic rules, etc) |
| **Big Data Specific Challenges (Gaps)** | | Data management (metadata, provenance info, data identification with PIDs)  Data curation  Digitising existing audio-video, photo and documents archives | |
| **Big Data Specific Challenges in Mobility** | | Include data from sensors of mobile devices (position, etc.);  Data collection from expeditions and field research. | |
| **Security & Privacy**  **Requirements** | | Privacy issues may be involved (A/V from individuals), anonymization may be necessary but not always possible (A/V analysis, small speech communities)  Archive and metadata integrity, long term preservation | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Many individual data entries from many individuals, constant flux of data entry, metadata assignment, etc.  Offline vs. online use, to be synchronized later with central database.  Giving significant feedback to contributors. | |
| **More Information (URLs)** | | --- | |
| **Note:** Crowd sourcing has been barely started to be used on a larger scale. With the availability of mobile devices, now there is a huge potential for collecting much data from many individuals, also making use of sensors in mobile devices. This has not been explored on a large scale so far; existing projects of crowd sourcing are usually of a limited scale and web-based. | | | |

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