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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | DataNet Federation Consortium (DFC) | |
| **Vertical (area)** | | Collaboration Environments | |
| **Author/Company/Email** | | Reagan Moore / University of North Carolina at Chapel Hill / rwmoore@renci.org | |
| **Actors/Stakeholders and their roles and responsibilities** | | National Science Foundation research projects: Ocean Observatories Initiative (sensor archiving); Temporal Dynamics of Learning Center (Cognitive science data grid); the iPlant Collaborative (plant genomics); Drexel engineering digital library; Odum Institute for social science research (data grid federation with Dataverse). | |
| **Goals** | | Provide national infrastructure (collaboration environments) that enables researchers to collaborate through shared collections and shared workflows. Provide policy-based data management systems that enable the formation of collections, data grid, digital libraries, archives, and processing pipelines. Provide interoperability mechanisms that federate existing data repositories, information catalogs, and web services with collaboration environments. | |
| **Use Case Description** | | Promote collaborative and interdisciplinary research through federation of data management systems across federal repositories, national academic research initiatives, institutional repositories, and international collaborations. The collaboration environment runs at scale: petabytes of data, hundreds of millions of files, hundreds of millions of metadata attributes, tens of thousands of users, and a thousand storage resources. | |
| **Current**  **Solutions** | **Compute(System)** | | Interoperability with workflow systems (NCSA Cyberintegrator, Kepler, Taverna) |
| **Storage** | | Interoperability across file systems, tape archives, cloud storage, object-based storage |
| **Networking** | | Interoperability across TCP/IP, parallel TCP/IP, RBUDP, HTTP |
| **Software** | | Integrated Rule Oriented Data System (iRODS) |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Manage internationally distributed data |
| **Volume (size)** | | Petabytes, hundreds of millions of files |
| **Velocity**  **(e.g. real time)** | | Support sensor data streams, satellite imagery, simulation output, observational data, experimental data |
| **Variety**  **(multiple datasets, mashup)** | | Support logical collections that span administrative domains, data aggregation in containers, metadata, and workflows as objects |
| **Variability (rate of change)** | | Support active collections (mutable data), versioning of data, and persistent identifiers |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Provide reliable data transfer, audit trails, event tracking, periodic validation of assessment criteria (integrity, authenticity), distributed debugging |
| **Visualization** | | Support execution of external visualization systems through automated workflows (GRASS) |
| **Data Quality** | | Provide mechanisms to verify quality through automated workflow procedures |
| **Data Types** | | Support parsing of selected formats (NetCDF, HDF5, Dicom), and provide mechanisms to invoke other data manipulation methods |
| **Data Analytics** | | Provide support for invoking analysis workflows, tracking workflow provenance, sharing of workflows, and re-execution of workflows |
| **Big Data Specific Challenges (Gaps)** | | Provide standard policy sets that enable a new community to build upon data management plans that address federal agency requirements | |
| **Big Data Specific Challenges in Mobility** | | Capture knowledge required for data manipulation, and apply resulting procedures at either the storage location, or a computer server. | |
| **Security & Privacy**  **Requirements** | | Federate across existing authentication environments through Generic Security Service API and Pluggable Authentication Modules (GSI, Kerberos, InCommon, Shibboleth). Manage access controls on files independently of the storage location. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Currently 25 science and engineering domains have projects that rely on the iRODS policy-based data management system:  Astrophysics Auger supernova search  Atmospheric science NASA Langley Atmospheric Sciences Center  Biology Phylogenetics at CC IN2P3  Climate NOAA National Climatic Data Center  Cognitive Science Temporal Dynamics of Learning Center  Computer Science GENI experimental network  Cosmic Ray AMS experiment on the International Space Station  Dark Matter Physics Edelweiss II  Earth Science NASA Center for Climate Simulations  Ecology CEED Caveat Emptor Ecological Data  Engineering CIBER-U  High Energy Physics BaBar  Hydrology Institute for the Environment, UNC-CH; Hydroshare  Genomics Broad Institute, Wellcome Trust Sanger Institute  Medicine Sick Kids Hospital  Neuroscience International Neuroinformatics Coordinating Facility  Neutrino Physics T2K and dChooz neutrino experiments  Oceanography Ocean Observatories Initiative  Optical Astronomy National Optical Astronomy Observatory  Particle Physics Indra  Plant genetics the iPlant Collaborative  Quantum Chromodynamics IN2P3  Radio Astronomy Cyber Square Kilometer Array, TREND, BAOradio  Seismology Southern California Earthquake Center  Social Science Odum Institute for Social Science Research, TerraPop | |
| **More Information (URLs)** | | The DataNet Federation Consortium: <http://www.datafed.org>  iRODS: http://www.irods.org | |
| **Note:** <additional comments>A major challenge is the ability to capture knowledge needed to interact with the data products of a research domain. In policy-based data management systems, this is done by encapsulating the knowledge in procedures that are controlled through policies. The procedures can automate retrieval of data from external repositories, or execute processing workflows, or enforce management policies on the resulting data products. A standard application is the enforcement of data management plans and the verification that the plan has been successfully applied. | | | |

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