**Use Case: Biodiversity**

**NBD(NIST Big Data) Requirements WG**

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

**Contents**

1. Blank Template
2. Biodiversity and LifeWatch – Wouter Los, Yuri Demchenko, University of Amsterdam

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | LifeWatch – E-Scienc European Infrastructure for Biodiversity and Ecosystem Research | |
| **Vertical (area)** | | Life Science | |
| **Author/Company/Email** | | Wouter Los, Yuri Demchenko (y.demchenko@uva.nl), University of Amsterdam | |
| **Actors/Stakeholders and their roles and responsibilities** | | End-users (biologists, ecologists, field researchers)  Data analysts, data archive managers, e-Science Infrastructure managers, EU states national representatives | |
| **Goals** | | Research and monitor different ecosystems, biological species, their dynamics and migration. | |
| **Use Case Description** | | LifeWatch project and initiative intends to provide integrated access to a variety of data, analytical and modeling tools as served by a variety of collaborating initiatives. Another service is offered with data and tools in selected workflows for specific scientific communities. In addition, LifeWatch will provide opportunities to construct personalized ‘virtual labs', also allowing to enter new data and analytical tools.  New data will be shared with the data facilities cooperating with LifeWatch.  Particular case studies: Monitoring alien species, monitoring migrating birds, wetlands  LifeWatch operates Global Biodiversity Information facility and Biodiversity Catalogue that is Biodiversity Science Web Services Catalogue | |
| **Current**  **Solutions** | **Compute(System)** | | Field facilities TBD  Datacenter: General Grid and cloud based resources provided by national e-Science centers |
| **Storage** | | Distributed, historical and trends data archiving |
| **Networking** | | May require special dedicated or overlay sensor network. |
| **Software** | | Web Services based, Grid based services, relational databases |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Ecological information from numerous observation and monitoring facilities and sensor network, satellite images/information, climate and weather, all recorded information.  Information from field researchers |
| **Volume (size)** | | Involves many existing data sets/sources  Collected amount of data TBD |
| **Velocity**  **(e.g. real time)** | | Data analysed incrementally, processes dynamics corresponds to dynamics of biological and ecological processes.  However may require real time processing and analysis in case of the natural or industrial disaster.  May require data streaming processing. |
| **Variety**  **(multiple datasets, mashup)** | | Variety and number of involved databases and observation data is currently limited by available tools; in principle, unlimited with the growing ability to process data for identifying ecological changes, factors/reasons, species evolution and trends.  See below in additional information. |
| **Variability (rate of change)** | | Structure of the datasets and models may change depending on the data processing stage and tasks |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | In normal monitoring mode are data are statistically processed to achieve robustness.  Some biodiversity research are critical to data veracity (reliability/trustworthiness).  In case of natural and technogenic disasters data veracity is critical. |
| **Visualization** | | Requires advanced and rich visualization, high definition visualisation facilities, visualisation data   * 4D visualization * Visualizing effects of parameter change in (computational) models * Comparing model outcomes with actual observations (multi dimensional) |
| **Data Quality** | | Depends on and ensued by initial observation data.  Quality of analytical data depends on used mode and algorithms that are constantly improved.  Repeating data analytics should be possible to re-evaluate initial observation data.  Actionable data are human aided. |
| **Data Types** | | Multi-type.  Relational data, key-value, complex semantically rich data |
| **Data Analytics** | | Parallel data streams and streaming analytics |
| **Big Data Specific Challenges (Gaps)** | | Variety, multi-type data: SQL and no-SQL, distributed multi-source data.  Visualisation, distributed sensor networks.  Data storage and archiving, data exchange and integration; data linkage: from the initial observation data to processed data and reported/visualised data.   * Historical unique data * Curated (authorized) reference data (i.e. species names lists), algorithms, software code, workflows * Processed (secondary) data serving as input for other researchers * Provenance (and persistent identification (PID)) control of data, algorithms, and workflows | |
| **Big Data Specific Challenges in Mobility** | | Require supporting mobile sensors (e.g. birds migration) and mobile researchers (both for information feed and catalogue search)   * Instrumented field vehicles, Ships, Planes, Submarines, floating buoys, sensor tagging on organisms * Photos, video, sound recording | |
| **Security & Privacy**  **Requirements** | | Data integrity, referral integrity of the datasets.  Federated identity management for mobile researchers and mobile sensors  Confidentiality, access control and accounting for information on protected species, ecological information, space images, climate information. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | * Support of distributed sensor network * Multi-type data combination and linkage; potentially unlimited data variety * Data lifecycle management: data provenance, referral integrity and identification * Access and integration of multiple distributed databases | |
| **More Information (URLs)** | | <http://www.lifewatch.eu/web/guest/home>  <https://www.biodiversitycatalogue.org/> | |
| **Note:** <additional comments>  **Variety of data used in Biodiversity research**  Genetic (genomic) diversity   * DNA sequences & barcodes * Metabolomics functions   Species information   * -species names * occurrence data (in time and place) * species traits and life history data * host-parasite relations * collection specimen data   Ecological information   * biomass, trunk/root diameter and other physical characteristics * population density etc. * habitat structures * C/N/P etc molecular cycles   Ecosystem data   * species composition and community dynamics * remote and earth observation data * CO2 fluxes * Soil characteristics * Algal blooming * Marine temperature, salinity, pH, currents, etc.   Ecosystem services   * productivity (i.e biomass production/time) * fresh water dynamics * erosion * climate buffering * genetic pools   Data concepts   * conceptual framework of each data * ontologies * provenance data   Algorithms and workflows   * software code & provenance * tested workflows   **Multiple sources of data and information**   * Specimen collection data * Observations (human interpretations) * Sensors and sensor networks (terrestrial, marine, soil organisms), bird etc tagging * Aerial & satellite observation spectra * Field \* Laboratory experimentation * Radar & LiDAR * Fisheries & agricultural data * Deceases and epidemics | | | |