Devising an Infrastructure for Data Interoperability

Alberto Pepe, Jillian C. Wallis, Matthew Mayernik & Christine L. Borgman
CENS Data Management

Introduction

Data handling practices

- The ways in which sensor data are encoded, analyzed and shared vary widely across different CENS research groups and projects
- The entire data production ***life cycle***, from data acquisition to data dissemination is affected by the ***lack of standard procedures***
  - Different metadata formats are used to encode sensor data
  - Internal circulation of data is mostly based on the use of email, FTP or shared server space
- This data management setting ***does not stimulate data exchange***

Data interoperability

- In order to incite data sharing and guarantee data consistency across the CENS research spectrum, it is vital to implement an ***underlying common infrastructure*** for ***data handling***
- Such an infrastructure would 1) enforce adoption of ***XML-based formats***, such as ***SensorML*** and ***EML***, to encode sensor data as well as detector and processes; 2) function on top of an application-independent framework for interoperability, such as the ***OAI-PMH***
- The utilization of standards would encourage ***data sharing***, both internally and externally, and enhance ***data visibility and re-use*** while ensuring long-term preservation.

Problem: Data handling practices vary widely across CENS community

The entire data production life cycle is affected

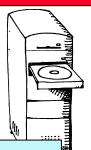
Data acquisition
- Raw sensor data instances and sensor system specifications are encoded and stored in ***different formats***
- Formats adopted might be ***incompatible*** with each other and might ***lack appropriate metadata description***

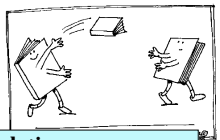
Data analysis
- ***Ad-hoc procedures*** are employed to process different kinds of sensor data
- Instructions and procedures concerning the process of measurement and analysis are ***not defined using a standard model***

Data circulation
- Raw and processed data are circulated internally mostly via ***shallow methods*** (e.g. email attachment, shared server space, static FTP)
- Data circulated is often in ***proprietary formats*** (e.g. Excel tables)

Data sharing
- Sensor data is made publicly available by ***static deposit*** on the CENS website or on specialized repositories.
- The means and extent of data sharing is not made clear by a ***CENS-wide policy***

Adopting standards to enforce interoperability across the entire data life-cycle






Data acquisition
- Detectors, sensor systems and raw data instances are defined using open standard XML encoding, e.g. ***SensorML***
- All raw and descriptive data is stored in a searchable archive for later ***discovery and re-use***

Data analysis
- The instructions and descriptions of any analytical process are documented using standard encoding, e.g. ***SensorML***
- As for the raw data, process data is stored in a searchable archive for later ***retrieval, analysis and/or visualization***

Data circulation
- Specific datasets might be specified in high-level metadata languages, e.g. ***EML***
- All raw, descriptive and process data are circulated internally via a ***shared, searchable, archive*** in which data is encoded in ***open formats***

Data sharing
- Sensor data and publications are shared with external institutions and individuals via a ***web portal*** built on top of the central archive
- Automated harvesting and provision of content in the archive is made possible by ***compliance to standard protocols for interoperability***, e.g. ***OAI-PMH***.

SensorML ***Sensor Model Language***
- Provides standard models and XML encodings for describing sensor resources (*"how were the observations obtained?"*) as well as higher-level sensor information (*"what can we derive from these observations?"*)
- makes sensors and processes used visible and traceable
- "A means of countering the proliferation of disparate, stovepipe systems for processing sensor data within various sensor communities"

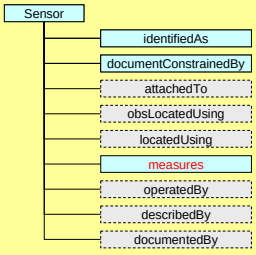


Image adapted from sensorsmag.com (Apr 1, 2003)

EML ***Ecological Metadata Language***
- A metadata specification to describe high-level ecological data (other disciplines in the CENS community would use other appropriate, specific, data formats)
- Could be used on top of data repositories, such as Sensorbase.org
- Standard syntax and structure allow data retrieval and manipulation

OAI-PMH
Open Archive Initiative Protocol for Metadata Harvesting
- Provides an application-independent interoperability framework based on metadata harvesting
- An 'envelope' to wrap SensorML, EML or any other XML-based metadata instance in a common archive
- A web portal can be built on top of the repository allowing search and retrieval directly from the metadata archive or via cross-searching (e.g. using Z39.50 protocol)