RCN Open Data Formats and Standards

Sandy Williams, James Gallagher, Fred Maltz,
Christoph Waldmann, and Mike McCann

Why do formats and standards matter?

- Access to data requires that the format be recognized
- Standards ensure these formats remain available
- Many standards are defined by a standards association
  - ASCII, ANSI, ISO
- Other standards are ad hoc or commercial
  - PDF, MS Word, printing on archival paper
- Geological and biological samples are also "data"
  - Unprocessed biological samples are saved and archived
    - Biological whole animal storage "standards" include buffered formalin, alcohol, non-formalin EPA aqueous solution, liquid nitrogen, dried, mounted on microscope slides

Accessibility - Longevity

- Cuneiform is still accessible
- Text representations of numeric data can be accessed with ASCII standard format
- Data transcends media of storage so that migration from handwriting to printing to digital encoding on cards, tape, disc, and flash memory, to cloud (unknown medium to me and others)
- Once common formats may become inaccessible encouraging backup copies in more primitive format
  - Multiplan format data may not be readable in Excel but text data can be imported to current spreadsheet applications

Open Data Requires Accessibility

- An article of faith is that sharing observational data benefits society
  - Short term exceptions may require delay of access for reasons of security, personal benefit, corporate profit
- Impediments to data exchange and sharing that are artificial due to incompatibility of formats should be removed
  - Conversion applications can improve interoperability
- Established or de facto standard formats reduce inaccessibility

Standard Data Formats Are Many

- **Glossary**
  - ANSI American National Standards Institute
  - API application programming interface
  - ArcIMS Arc internet map server, a web map server produced by Esri
  - ArcSDE spatial database engine produced and marketed by Esri
  - ASCII American Standard Code for Information Interchange
  - CEN European committee for standardization
  - CEOS committee on earth observations satellites
  - CEOS QA4EO GEO/CEOS Workshop on Quality Assurance of Calibration & Validation Processes
- CSV comma separated values
- DAP data access protocol
- DBMS database management system
- DGIWG digital geographic information working group
- DITA Darwin information typing architecture
- DLG digital line graph
- DMAC data management and communication
- DXF drawing interchange format
- FGDC federal geographic data committee
- GeoTiff a public domain metadata standard which allows georeferencing information to be embedded within a TIFF file

How Can Use of Standard Formats Be Encouraged?

- Sponsors of observational science programs can encourage/ require standard formats for data storage
- Scientific journals can require that data in support of publication be deposited in a standard format
- Communities of scientists can recognize observational or model data sets are professional contributions deserving citation and therefore needing to be in a standard format for accessibility

Digital Data Stored Electronically

- Streaming data - dynamic
  - Real-time
  - Satellite
  - Minimal quality control, minimal preprocessing
- Archived voluminous data - static
  - Quality flag metadata stored with data
- Hand acquired data – static, small volume
  - Notebook, paper, few data, possibly preprocessed
  - Powerful data exchange applications unnecessary

File vs. Web Data Exchange

- File transfers make up most of the networked data access
  - Users must first find the files they want
  - Users must figure out how to read them
  - Files are generally larger than the user needs so they must be broken apart and combined as needed
- Web services address the problem with file transfers by hiding the actual file format
  - Transfer may be proprietary
  - Technically challenging for users
    - Help is available
  - DAP (data access protocol) and OPeNDAP
    - Open source, non-profit

Other Data Types

- GIS data are handled through data converters like DLG, MOSS, and GIRAS
- Standard interchange formats like SDTS, DXF, and GML
- Open file formats like VPF, shapefiles
- Direct read application programming interfaces like ArcSIDE API, CAD Reader, ArcSide CAD Client
- Common features in a database management system (DBMS) like OGC Simple Feature Specification for SQL
- Integration of standard GIS Web services like WMS, WFS, ArcIMS

High Level Proprietary Data Formats

- Google Maps from satellite downloads of data in a NASA format
- Delivery of image files for display of Google Maps on a PC or iPad
- GPS location data generated by Magellan or Garmin processing of GPS satellite transmissions in a DoD format
- Combination of GPS and GIS map data for navigation by iPad or equivalent through third party applications like Map GPS using Google Imagery, USDA Farm Service Agency, DigitalGlobe, GeoEye, U.S. Geological Survey

OpenData Formats vs. Proprietary Formats for High End Applications

- Commercial applications can provide processed data in a highly effective product like the GPS navigation and Google map products
- The format of these products is not open and is difficult or impossible to decode for other purposes
- The raw data used in generating these products is available but not easily so and the processing is so complex that very little of the final product might be achieved economically
- Since the original raw data are still available, there is no loss of accessibility to the OpenData through subsequent processing

Recommendations

- National policies must ensure formats of data acquired under their federal sponsorship are OpenData formats, complying with known and available standards
  - Exceptions should be limited to national security restrictions but of limited duration
- Professional societies should require that data substantiating published papers in their journals should be available in OpenData formats, complying with known and available standards
- Academic Institutions should require that research performed by faculty and employees that rely on collected data have these data made available in OpenData formats, complying with known and available standards
  - Exceptions should be limited to short intervals for professional benefits of the researcher

Conclusions

- OpenData formats derive from common practice in many scientific communities
  - These standards are ad hoc
  - Some communities have established format standards through agencies such as ASCII, ANSI, and ISO
- Major commercial entities have established ad hoc standard formats
  - Adobe has established PDF
  - Microsoft has established Word and Excel
- Web based products at high levels use proprietary formats and these are not open but the products are derived from OpenData formats, complying with known and available standards