Big Data, Big Metadata & Byte sized pieces: Big Picture

Chris Little
IT Fellow, Met Office
OGC Architecture Board Member
Chair: Met Ocean DWG, EDR API SWG, Temporal DWG

6 December 2022
Tim Berners-Lee’s 5-star scheme for quality of web open data

Highest ranking is Linked Open Data:[10]

• 1 star: data is openly available in some format
• 2 stars: data available in structured format, e.g. MSExcel (.xls)
• 3 stars: data available in a non-proprietary structured format, such as Comma-separated values (.csv)
• 4 stars: data follows W3C standards, like using RDF and URIs
• 5 stars: all above, and links to other Linked Open Data sources.
Step Away from the Biscuit Tin
Domain Specific containers/formats will continue

• Optimized
• Efficient
• Installed base
• Ecosystem of supporting tools
• Specialized Controlled Vocabularies
• Captures domain expertise and specialities
• Generic alternatives not convincing for Big Data. E.g:
  – XML,
  – RDF,
  – JSON, ....
  – Or even EXI, BinaryXML, Fast InfoSet, ProtoBuf, Thrift/Avro, DDS, etc
Metadata not quite good enough

• Metadata is very open ended
• Metadata should be cross-domain
• Rigid containers with fixed slots e.g. ISO19115 file
• Still evolving, not yet well established:
• Few end user tools, usually domain specific
• No user defined extensions
• No data structural metadata
• No valid query metadata
• How does it support conceptual models?
Big Picture – 3: How to find our stuff?

• Catalogues:
  – Domain specific usually
  – Controlled vocabularies very helpful
  – Not very comprehensive
  – How do you find the Catalogue?

• Portals
  – Domain specific
  – All different “I am the one & only true portal (for your domain)”
  – Very little federation, though technically feasible
  – How do you find the Portal?

• Search Engines
  – Proprietary “magic”
  – Controlled vocabularies help
  – More dynamic than catalogues
  – Knowledge Graphs and ontologies
Big Picture - 4

- Semantic Web / Web 2.0
- Lots of relevant metadata outside of containers
- Highly scalable
- Highly flexible
- Resolvable registries of controlled vocabularies happening
- Conceptual models can be stored as ontologies
  - Abstract (IDs only) vs Logical (attributes) vs Implementation
- Ontologies allow valid machine reasoning
- Semantic formats much too verbose for most data

Copyright © 2017 Open Geospatial Consortium
“Data Gravity” needs apps to move to data
Metadata generally does not enable access to data (FAIR)
Web APIs help (for non-expert or expert users)
Data and Metadata should be cacheable for use at scale
XML good for stable environments: slow change, well-defined
XML not good for dynamic environments, flexibility, multiple ownership, fragments
Industry moving to JSON schemas, key value pairs, etc
Ontologies – KV pairs- Schema fragments – XML schemas – Controlled Vocabularies ??
Summary - Data

Cross-domain information becoming the norm, but

- Domain specific (big) data formats here to stay:
  - (NetCDF, HDF, GRIB, BUFR, FITS, BAM, VCF, GF3, …)
  - Binary, efficient, optimised
  - Established eco-systems of access software & tools
  - Established domain expertise, controlled vocabularies

- Generic data formats not much traction, but

- Web friendly transfer formats for browsers and apps
  - (CSV, JSON, XML)
  - More verbose – small amounts OK
Summary - Metadata

- Metadata should be cross-domain for discovery
  - Metadata fixed format/containers a good start
  - Only for discovery, not usage and management
  - Metadata is open-ended, not in containers/catalogues/portals
- Controlled Vocabularies -> Taxonomies -> Ontologies
- Semantic Web / Web 2.0: standards action in W3C
- Should be highly scalable, highly flexible
- Resolvable registries of controlled vocabularies happening
- Conceptual models can be stored as ontologies
  - Abstract (IDs only) vs Logical (attributes) vs Implementation
- Ontologies allow valid machine reasoning
- Semantic formats too verbose for data but metadata OK
Big Picture Summary

• Domain specific binary formats & access software to stay
  – Cross-domain access patterns need to be identified

• Metadata containers not good enough
  – Mainly for discovery not access or management

• Persistent cross-domain data access APIs needed
  – Authentication, authorisation, management separate, generic

• Cross-domain conceptual data models needed
  – XML Schema too limited
  – UML more general
  – Ontologies becoming useful
  – Schema fragments are useful

• Cross-domain, global solutions are essential
What am I (and OGC) doing about it?

- Primary use case: non-expert user, cross-domain, at scale
- Published OGC API-EDR for spatiotemporal data queries
  - Queries: point, radius, area, cube, trajectory, corridor
  - 1,2,3,4 or n-D, continuous or categorical
  - Also “items” query for API-Features compatibility
- Publishing CoverageJSON as OGC Community Standard
  - Web and web-developer friendly
  - JSON data for browser consumption
  - Supports JSON-LD
  - Established user base
- OGC register of Building Blocks
  - Registering schema fragments
  - Generate automatically “semantic uplift” of metadata to RDF
Questions, Discussion, Answers ?
Spare slides