Harmonising data description across the social and natural sciences

Cox, Hodson, Wyborn, Archer

The Data Integration Initiative coordinated through CODATA is examining the potential and barriers to the use of science data across the conventional disciplinary boundaries. As grounding for this, the initiative is conducting pilot studies in three application areas: resilient cities, infectious disease tracking, and disaster recovery. Each of these requires the integration of multiple datasets from various science repositories, together with routine data curated through monitoring and forecasting agencies, administrative data, and informal sources including live news feeds, citizen science, and social media. While the science data is increasingly FAIR, the source data from most of these sources may be only weakly structured and informally described and curated.

Discovery is the first step in accessing the required data. This is typically enabled through metadata available from catalogues and indexes, which may be hand-created or automatically populated. Standards for metadata have been developed and deployed in a number of communities for description of data in different application domains. For example, the ISO 19115 suite applies to geospatial data, particularly remote sensing imagery and map series; DDI applies to statistical data in the social sciences; EML to ecological data; DATS for health data; and Schema.org for web-page content primarily around commerce and cultural events. The level of detail of these various technologies varies, with some providing general discovery-level description, and others more detailed information supporting data access and re-use. Discovery of heterogeneous data requires use-agents to be capable of understanding and processing all the different metadata representations, or that a harmonized view is presented through some portal or mediator. In support of this goal, crosswalks convert between different representations. However, harmonized metadata is often lowest-common-denominator metadata, in which the detail that is critical to discovery and evaluation is lost.

An alternative approach is to use a common base standard for the general structure, and construct specialized metadata through well identified extension points, in order to support detailed characterization on a discipline-by-discipline basis. This approach is being assessed through CODATA in conjunction with the DDI initiative, particularly utilizing the work of the Data eXchange Working Group (DXWG) of the World Wide Web Consortium (W3C). The DXWG is focussing on two main deliverables: (i) an update to the Data Catalog Vocabulary (DCAT) (ii) a formal approach to defining metadata ‘profiles’.

DCAT extends the well known Dublin Core metadata element set to support high-level description of data representations and services in catalogues or repositories. It is formalized as an RDF vocabulary, which allows it to be easily combined syntactically with other RDF vocabularies, as well as easy integration with community vocabularies provided externally as linked-open-data resources, such as observable properties and traits, or units of measure. Coherent scientific metadata is likely to make use of specialized RDF vocabularies, such as SOSA/SSN (observations and sampling), QB (statistical data-cubes) and PROV (detailed provenance and workflows), as well as the significant ontologies for the health and life-sciences developed through the OBO foundation.