SHEDDING LIGHT ON FAIR DATA AND OPEN SCIENCE

Hodson, S.[1]
[1] CODATA, Paris, FRANCE
simon@codata.org

Abstract

First published in 2016, in an article that has now received over 6000 citations, the FAIR (Findability, Accessibility, Interoperability, Reusability) Principles have had a significant influence in policy, practice and thinking about research data management and stewardship. The catchy mnemonic has been effective, but the article and related work convey an important message. The fundamental purpose of the FAIR principles is to provide guidelines such that data and metadata relevant to all kinds of research outputs are machine readable and machine actionable. The vision is one in which research outputs can be visited online and at vast scale and the potential of machine assisted analysis can be realised, but with data and metadata that are sufficiently reliable so as to reduce error and quantify uncertainty. Another catchphrase is that FAIR also means Fully AI Ready.

The FAIR principles do not insist that data or metadata be fully open, and make the case that the principles apply just as well to cases where data can only be accessed in controlled circumstances. Yet, the interest in FAIR is related to another major development and movement in early 21st century research: that of Open Science. Recognising, of course, that in some circumstances access restrictions are necessary, the Open Science movement aims to maximise the public benefit of research (particularly publicly funded research) by insisting that scientific outputs should be as open as possible, and that access restrictions should be "proportionate and justified". he nuances of these arguments are frequently lost, and sometimes obfuscated by vested interests. Nevertheless, the UNESCO Recommendation on Open Science, supported by the International Science Council, represents another major step in the transformation of science policy.

Why does this matter? Why should we care about these developments in high level policy and in data practice? The major global scientific and human challenges of the 21st century (including climate mitigation and adaptation, sustainable development, disaster risk research and reduction, smart cities and energy systems, precision medicine or agriculture) can only be addressed through cross-domain research that seeks to understand complex systems through machine-assisted analysis at scale. Our capacity for such analysis is currently constrained by the limitations in our ability to access and combine heterogenous data within and across domains. The FAIR principles and the frameworks set by Open Science provide a significant part of the solution. Attention needs to be paid to the interfaces where data is used between disciplines, but the engagement with key research disciplines around data (re-)usability is essential.

To help address these issues, CODATA has been entrusted by the International Science Council (ISC) to develop a programme of activity: 'Making Data Work for Cross-Domain Grand Challenges'. After some exploratory work, the flagship activity is the WorldFAIR project which focuses on the implementation of the FAIR principles both within and across 11 different domain and cross-domain case studies, with a central effort to understand and guide cross-domain FAIR. It is the first broad-based effort to understand the issues around cross-domain and cross-infrastructure FAIR implementation through a case study driven methodology. Ultimately, it hopes to provide guidance for FAIR implementation both within specific domains and infrastructures and across them.

The I and the R of FAIR pose particular challenges but are particularly important in addressing complex issues where datasets need to be combined and in enhancing scientific rigour and reproducibility. Consequently, increasing attention is being paid to semantics, the maintenance of referenceable vocabularies, terminologies and ontologies and to metadata profiles—and to tools that facilitate the tracking of provenance and process, or that use variable level metadata and semantics to facilitate data integration. The outputs will include recommendations for FAIR implementation for each

case study, as well as a proposed Cross-Domain Interoperability Framework (CDIF). The CDIF Identifies a set of functional requirements for interoperability, particularly for steps in data combination, and recommends good practices for each of these requirements, in relation to the use of existing or emerging standards and specifications. The CDIF is categorically not a new standard, but is intended to act as a lingua franca across domain data practices and encourage the incorporation of a number of standards that perform important and specific functions across domains. We are keen to test this approach with colleagues from as many disciplines and application areas as possible.

This talk will explore these developments in detail and make a case for the importance of the FAIR principles and related initiatives for the CIE community.