FORCE2017-53
FORCE2017

FAIR principles in practice at the ENCODE data portal

Esther Chan (1) and Idan Gabdank (2)
(1) Stanford University, Orcid 0000-0002-2406-2623, (2) Stanford University

The ENCODE project (Encyclopedia of DNA Elements) is a NIH funded, public consortium that has been mapping and annotating regions across the human genome using assays that measure biochemical activities, from which biological function can then be inferred. This effort, ongoing since 2003, has produced over 500 terabytes of raw experimental, processed and analysis data files and corresponding metadata. This large data corpus also includes more recent efforts to integrate data and metadata from related projects such as mouseENCODE, modENCODE (model organism), Roadmap Epigenomics (REMC) and Genetics of Gene Regulation (GGR), all freely accessible through the ENCODE data portal (https://www.encodeproject.org).

The ENCODE Data Coordinating Center (DCC) is tasked with managing the intake, organization and curation of the consortium-produced data, as well as providing community access to the data resource and outreach. Beginning in 2012, the ENCODE DCC has overhauled the data portal to facilitate better access as the ENCODE data collection rapidly grows. Our design and implementation of the database organization infrastructure, data modeling and metadata standardization efforts share many of the guiding principles outlined in the Force11 FAIR (Finding, Accessible, Interoperable and Re-usable) Data Publishing document (https://www.force11.org/fairprinciples). Here, we present some vignettes illustrating our implementation of these principles, including the use of persistent unique identifiers to guard against data duplication (see abstract by Gabdank et al.), the organization of detailed metadata in structured, machine and human-readable formats, and the tracking of provenance of data elements.