FORCE2017-27-1
FORCE2017

Preventing unintended data duplication: ENCODE DCC showcase

Idan Gabdank, Esther T. Chan, Jason A. Hilton, Jean M. Davidson, Carrie A. Davis, J. Seth Strattan, Aditi K. Narayanan, Kathrina C. Onate, Timothy R. Dreszer, Ulugbek K. Baymuradov, Keenan Graham, Otto Jolanki, Stuart R. Miyasato, Forrest Y. Tanaka, Matt Simison, Benjamin C. Hitz, Cricket A. Sloan, and J. Michael Cherry
ENCODE DCC, Department of Genetics, Stanford University School of Medicine, Stanford, California, USA

According to FAIR Guiding Principles for a data object to be (F)indable it should be assigned a globally unique and persistent identifier. A data object is allowed to have multiple identifiers, as long as these identifiers are unique and are referring to a single concept. These guidelines do not ensure unambiguous representation of a concept by data objects. Creation of multiple data objects representing a single concept is problematic for multiple reasons: (1) the existence of duplicated objects is misleading for the resource users that get the false notion of multiple data objects present in the repository, while in truth there is only one single entity represented by duplicated objects, (2) since the duplication is unintended and many times remains undetected, the maintenance of these duplicated objects proves to be challenging from the standpoint of synchronization and consistency between the objects, (3) sub-optimal storage efficiency due to the presence of duplicated objects or data files (gigabytes of storage capacity in case of the files produced by NGS technologies). Conversely, having data object duplicates representing multiple distinct concepts is also problematic because it erroneously represents different concepts as identical data objects. Here we present the principles and methods ENCODE Data Coordination Center (DCC) applies during the data submission process to the ENCODE portal (https://www.encodeproject.org/) in order to prevent cases of unintended duplication of the submitted data.