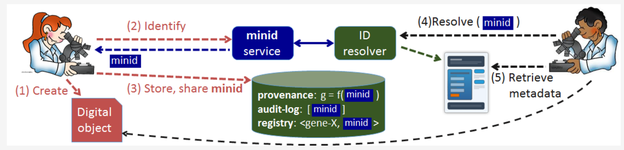
minids and BD Bags: New Tools for Working with Big Data

Big Data for Discovery Science, a BD2K Center of Excellence

BIG DATA FOR DISCOVERY SCIENCE

INI USC Stevens Neuroimaging and Informatics Institute | USC Laboratory of Neuro Imaging | NIH National Institutes of Health | Computation Institute | School of Nursing Statistics Online Computational Resource (SOCR) | USC Viterbi School of Engineering Information Sciences Institute | Institute for Systems Biology

minids: Lightweight Identifiers for Digital Objects



Minids are:

- Globally unique identifiers for digital objects
- Lightweight identifiers for intermediate data products that aren't intended for long-term archiving or formal publication
- Supported by software that makes them easy to generate, resolve, and query

A minid:

ark:/88120/r8059v

minid landing page:

Minids make it easy to:

- share or acquire datasets in a verifiable way
- to associate data objects with indexable, simple metadata
- to look up metadata corresponding to an identified digital object
- to pass data objects between researchers or analysis software tools, pipelines, and platforms

BD Bags: Structured Containers for Sets of Digital Objects

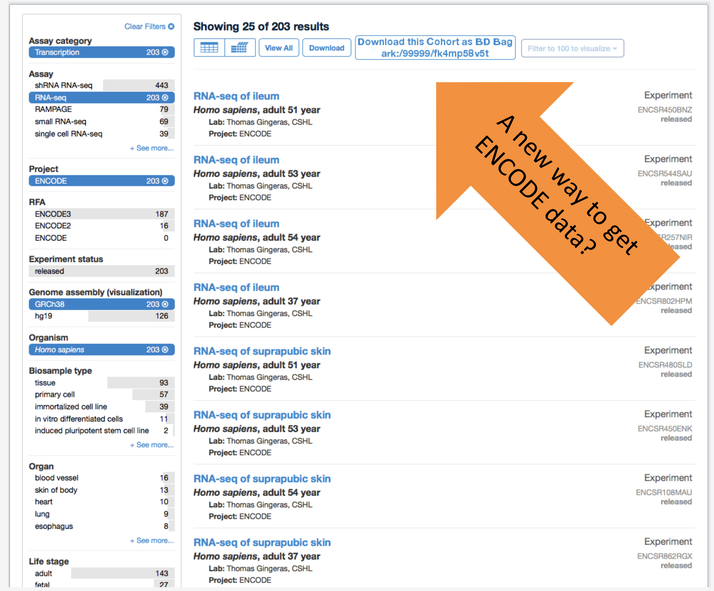
BD Bags are:

- Designed to support Discovery Science in the life sciences by supporting the ability to manage and process large sets of related data sets
- An extension of the Library of Congress Bagit specification for a hierarchical file packaging format designed to support disk-based storage and network transfer of arbitrary digital content.

BD Bags make it easy to:

- bundle a set of data files into a single downloadable package with a manifest and checksums
- assess what types of files are included in a BD Bag for automated acquiring software without complex file type discovery code
- exchange a set of files relating to a single research cohort between software services and pipelines without moving large volumes of data
- Operate in the emerging ecosystem of structured research data formats (e.g. the Library of Congress BagIt specification, the Research Object Bundle specification)

Future Capabilities?



For More Information

Email: Ben Heavner: ben.heavner@systemsbiology.org

BDDS website: http://bd2k.ini.usc.edu/

minids: http://minid.bd2k.org/

BD Bags repostiroy: https://github.com/ini-bdds/bagit-python