How Helium Team's Stack Uses Minids and BDBags

May 24, 2018

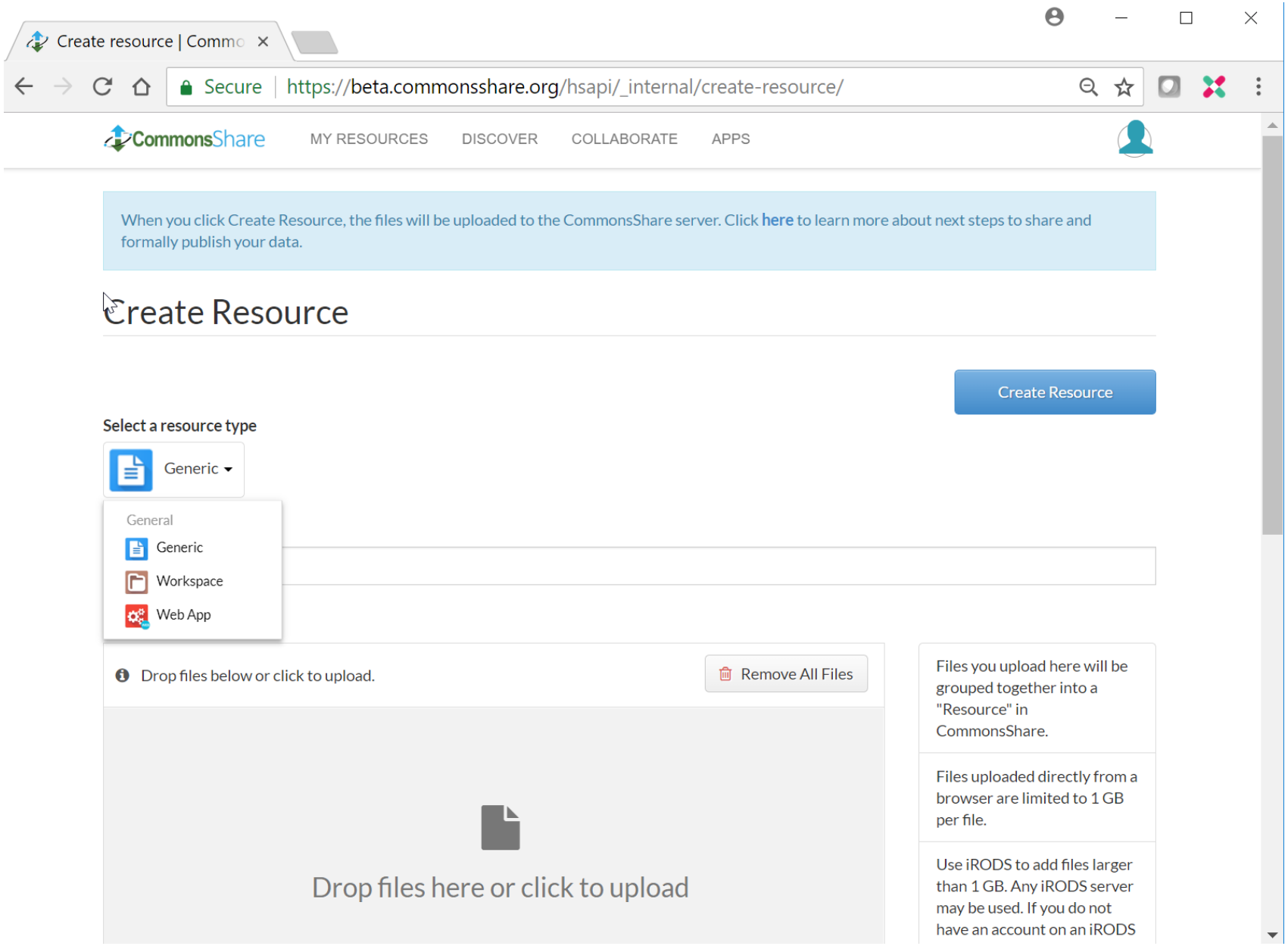
Helium Team Presentation to KC7

CommonsShare

- CommonsShare is Helium Team's KC5 web user interface to the NIH Data Commons
- https://beta.commonsshare.org/
  - Supports "Sign In With Globus"

NOTE: All minids referenced herein and created in beta.commonsshare.org are with test flag set to true meaning they are for testing purposes only.

CommonsShare Resources



- CommonsShare (CS)
  - In CommonsShare, ingested data becomes **resources**
  - One or more of any type of files can be added to Generic resources
  - CS Workspace resources can hold multiple other resources
  - Web App resources hold apps from the App Store
    - E.g. Jupyter notebooks

CommonsShare BDBag

- https://beta.commonsshare.org/resource/bf6932a87efc49aa94d02ab239957a87/

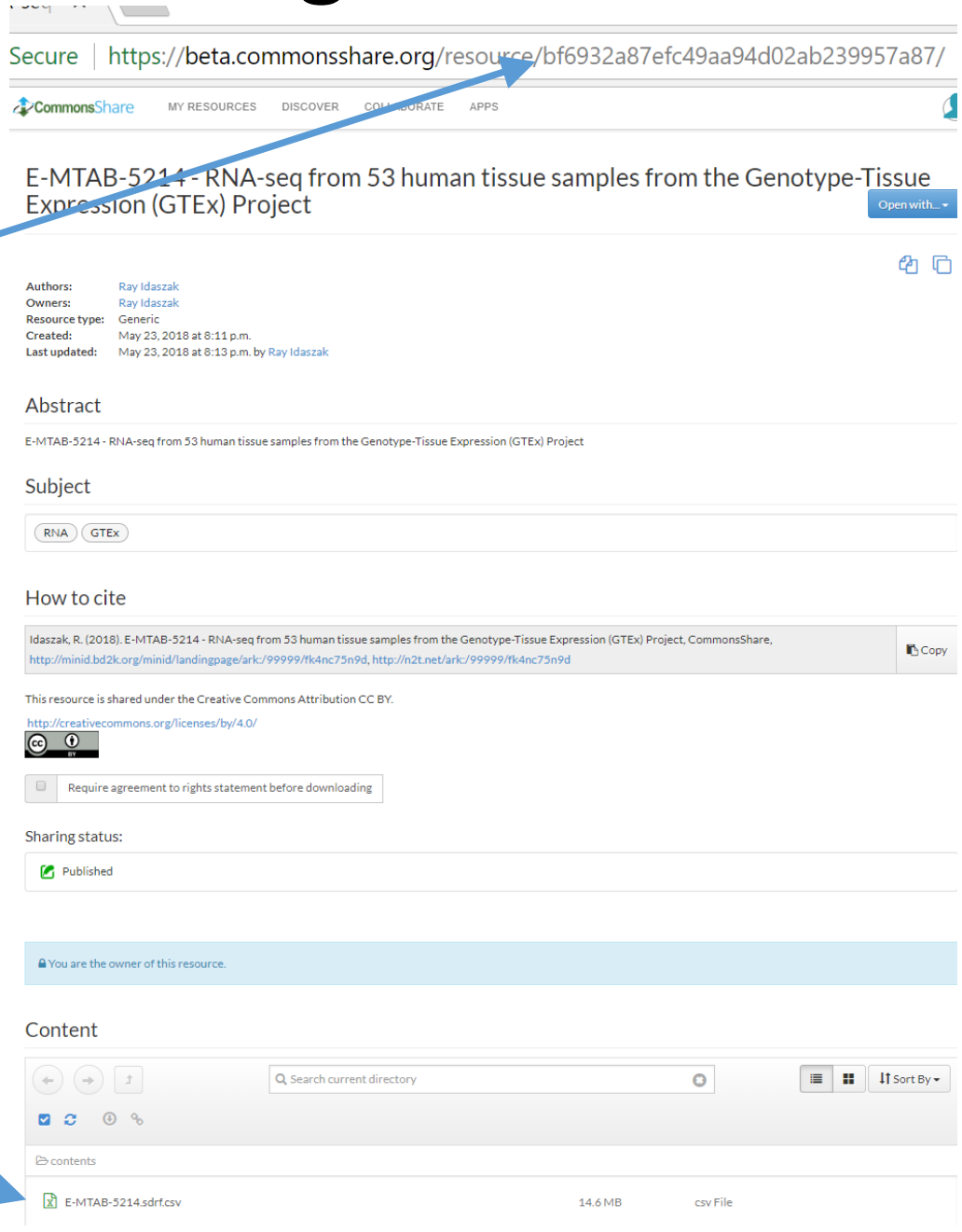
GUID bf6932a87efc49aa94d02ab239957a87/

/data

- bag-info.txt
- bagit.txt
- fetch.txt
- manifest-md5.txt
- manifest-sha256.txt
- tagmifest-md5.txt
- tagmanifest-sha256.txt

fetch.txt

irods://test.commonsshare.org:1247/commonssharetestZone/home/bagsdata/bf6932a87efc49aa94d02ab239957a87/data/contents/E-MTAB-5214.sdrf.csv 15348705 data/E-MTAB-5214.sdrf.csv



CommonsShare and fetch.txt

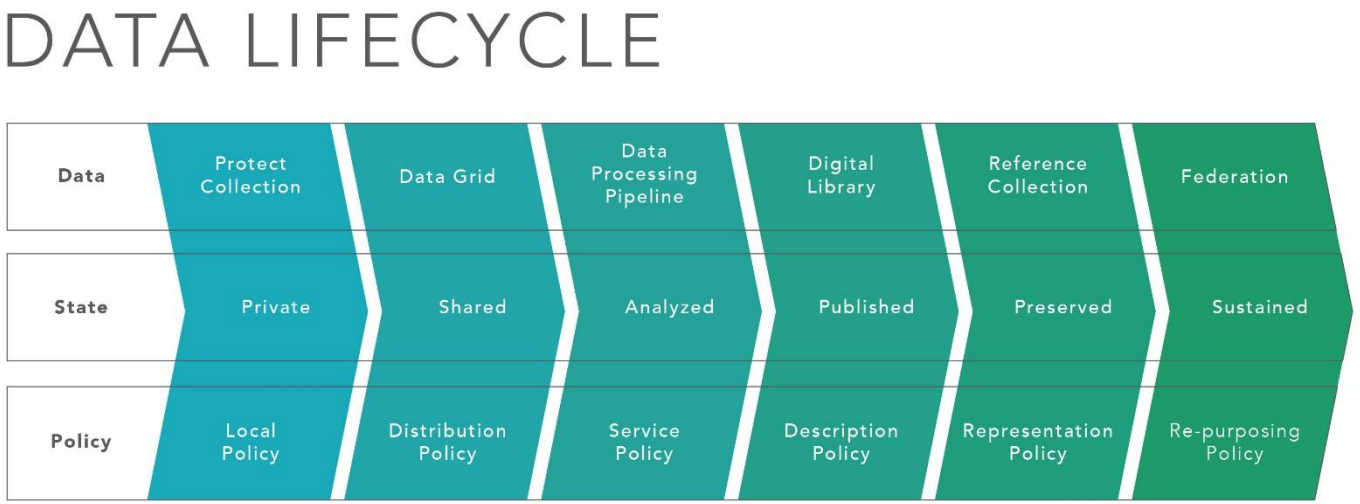
- CommonsShare
  - In CommonsShare BDBags, all payload is always referenced in fetch.txt
  - "…a fully complete BDBag with all data present in its data directory is semantically equivalent to an empty BDBag with its remote dependencies specified in the fetch.txt file…"*
  - Practically speaking, our Helium+ NHLBI DC use cases involve multi-TB BDBag payloads, so the fetch.txt-only approach reasonably ensures that the BDBag will reasonably be servable by http requests
    - i.e. a BDBag using only fetch.txt and no local payload will likely mostly stay under a GB thus servable by http

*K. Chard et al., "I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets," in Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016, pp. 319– 328; https://doi.org/10.1109/BigData.2016.7840618.

CommonsShare: Support for Data Lifecycle

iRODS
DATA LIFECYCLE

From Ingest to Institutional Repository



iRODS virtualizes the stages of the data lifecycle through policy evolution.

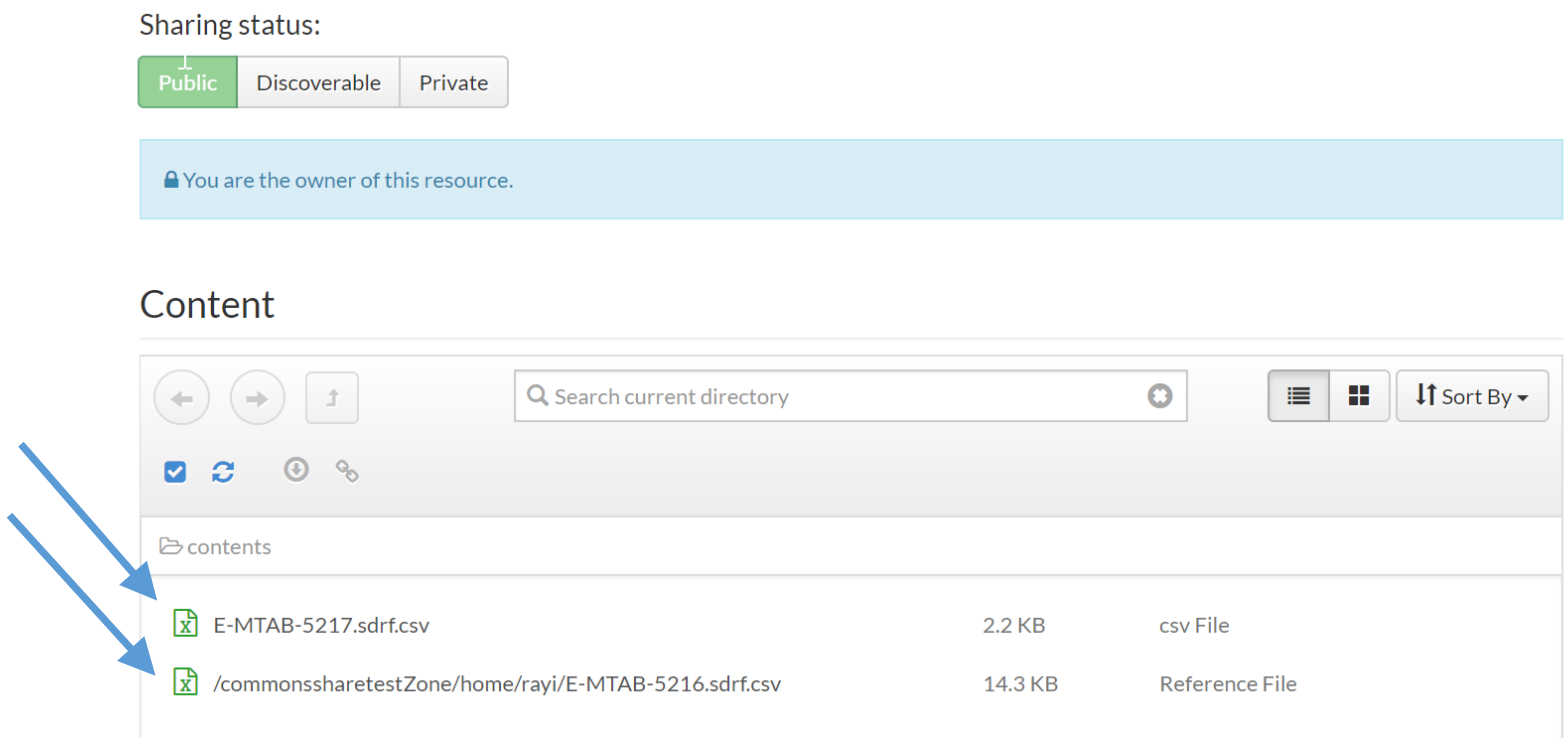
As data matures and reaches a broader community, data management policy must also evolve to meet these additional requirements.

Source: https://irods.org/images/data_lifecycle_diagram.jpg

irods.org

Pre-minid Helium Team CommonsShare resources are mutable

- CommonsShare
  - In CommonsShare, resources are assigned a GUID at time of creation
  - Resources are mutable until they are published with a minid
  - The idea is in supporting the full data lifecycle, artifacts and metadata are added intermittently to resources until ready to be published
  - The owner determines when to make a resource immutable so as to only create the ID provenance chain when ready and relevant
    - There will still be an audit trail, though primarily for handling sensitive data



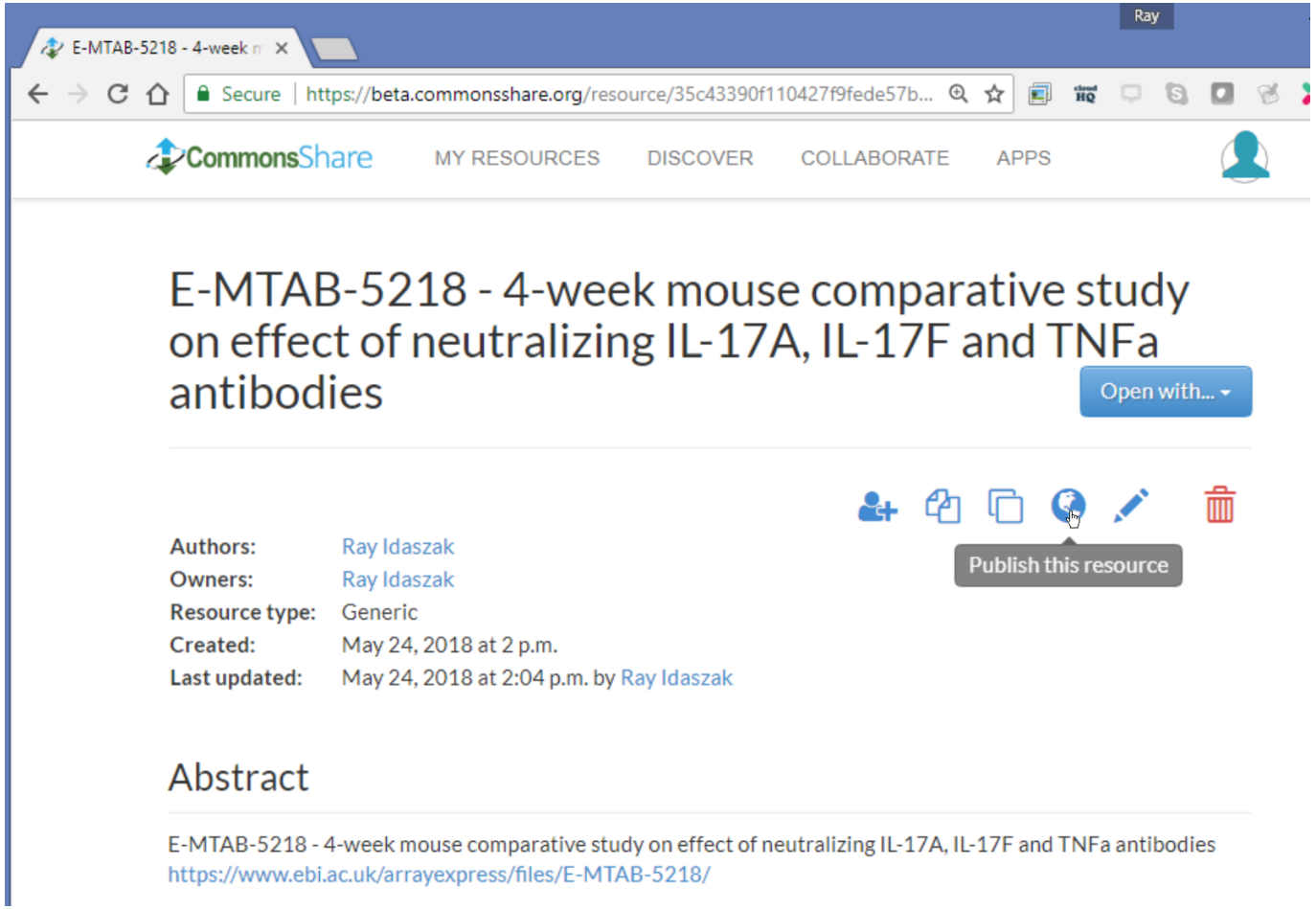
CommonsShare Identifiers

- Helium Framework Services and Interop: KC2-- Global Identifiers and metadata
  - "We have extended our identifier management layer to comply with the BD2K Minimum Viable Identifier (Minid) proposed by the Argon team. Team Helium also provides PrefixCommons, which supports robust curation of identifier prefixes for standardizing identifier resolution. The content within PrefixCommons is synchronized with N2T.net and Identifiers.org. Since Minids build on ARK, ***we will use N2T.net as the primary resolver to resolve identifiers globally***, however we will also utilize identifiers.org as the secondary resolver as we believe that it is important to have more than one resolver in case of outage or other issues. We have incorporated the code and services developed by the Argon team to mint and register minids, and are curating all prefixes from the three data stewards within PrefixCommons to support resolution of all identifiers within any stack."

  - https://docs.google.com/document/d/1R0Ucl6R6h0AjqwqOPr1f8_IW_6ZP8SBlTB_25jGb41U/edit?usp=sharing

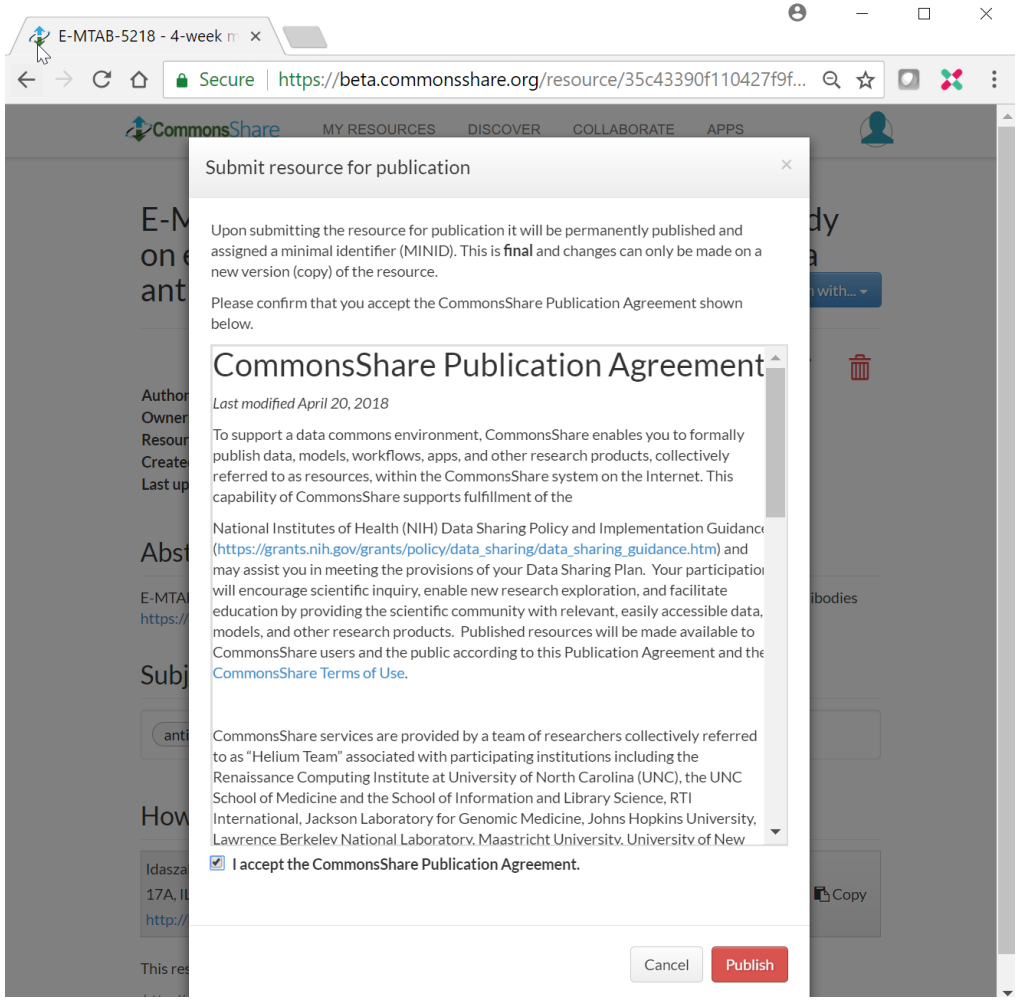
CommonsShare minting a minid

- Once a CommonsShare mutable resource with GUID is ready to be published as immutable with a minid, the user selects the “Publish this resource” icon on the resource landing page



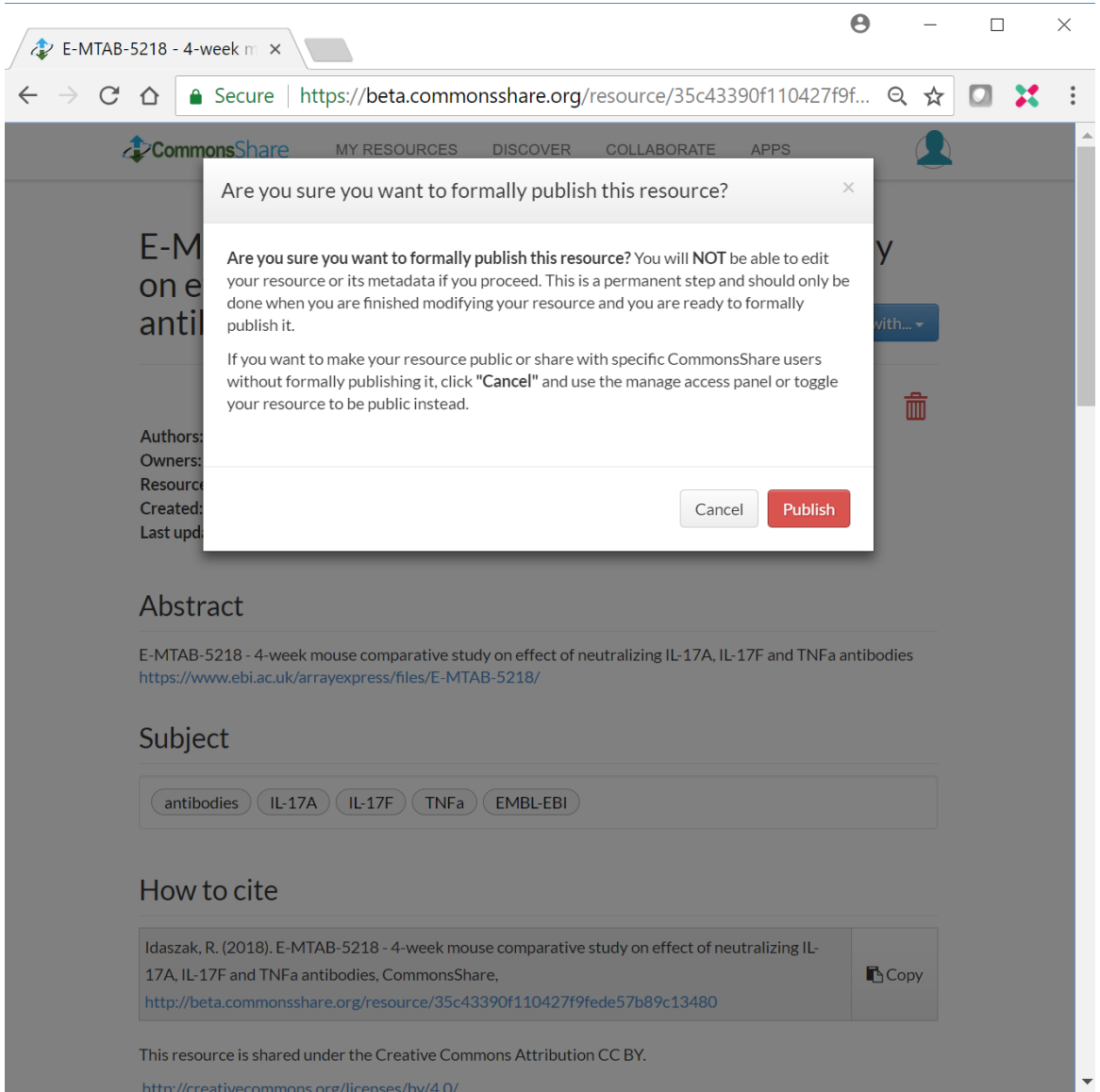
CommonsShare minting a minid

- Upon selecting “Publish this resource” a CommonsShare Publication Agreement is presented for the user to accept as they Publish.
  - Text is here: https://www.commonsshare.org/publication-agreement/



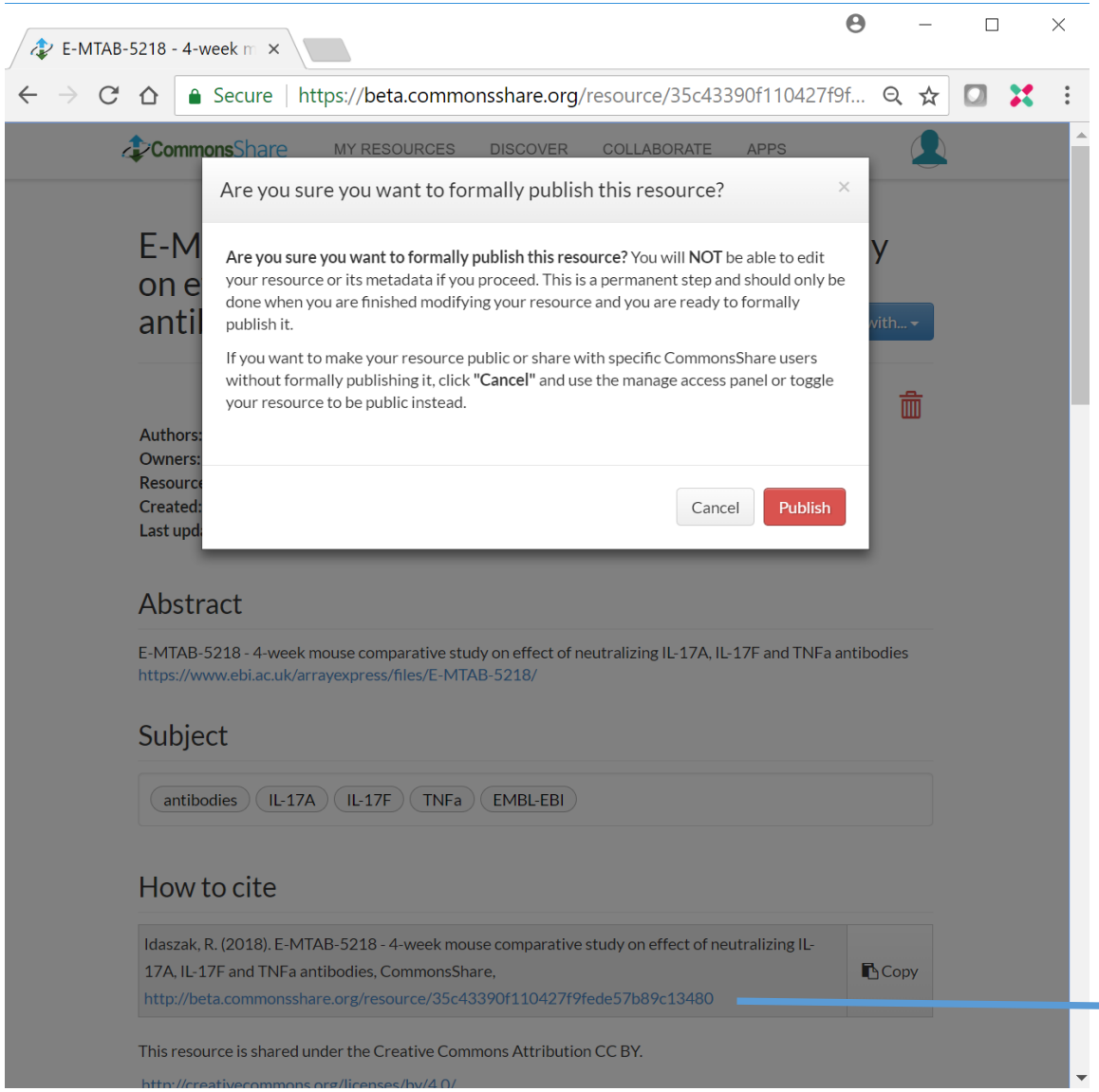
CommonsShare minting a minid

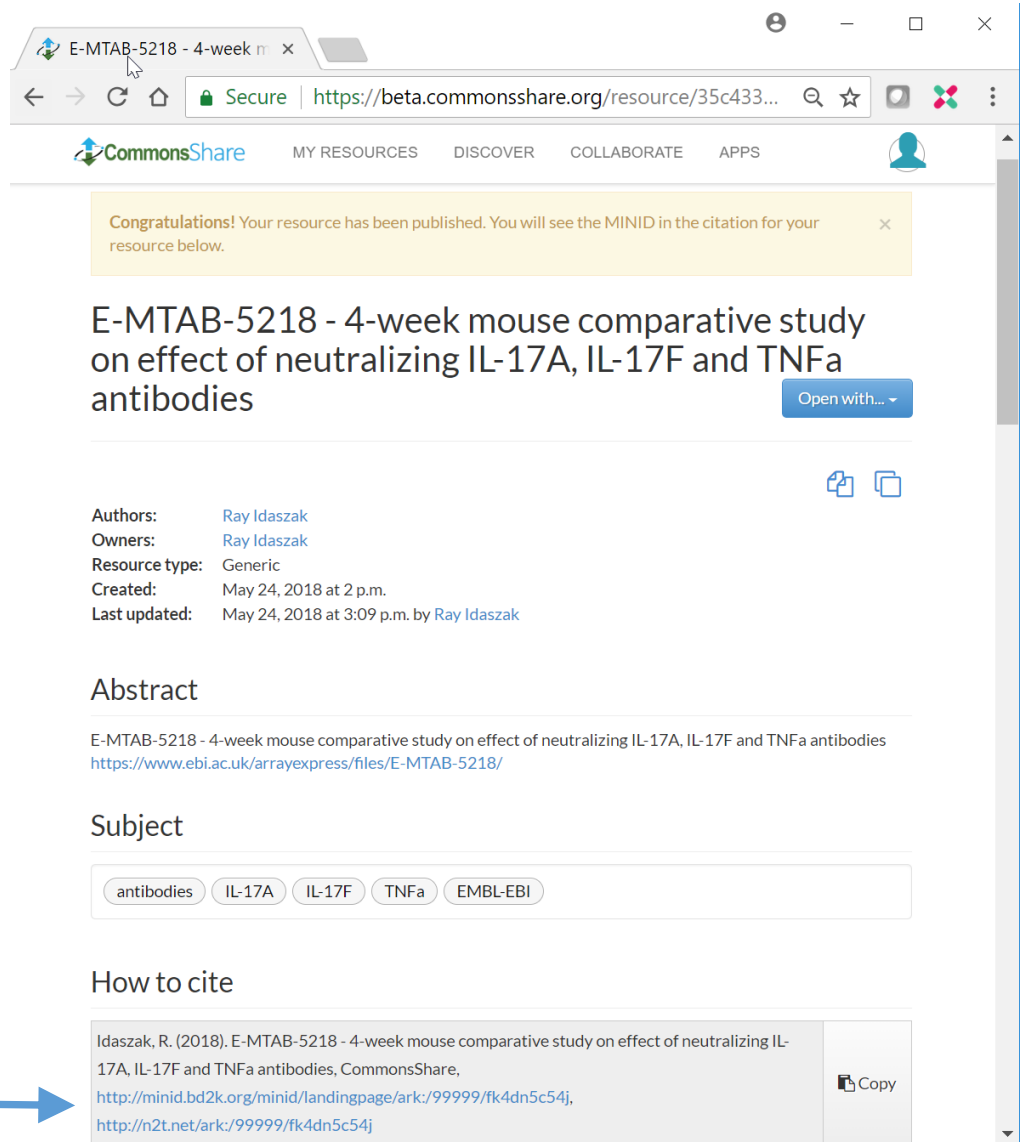
- …along with one more “Are you sure” message.



CommonsShare minting a minid

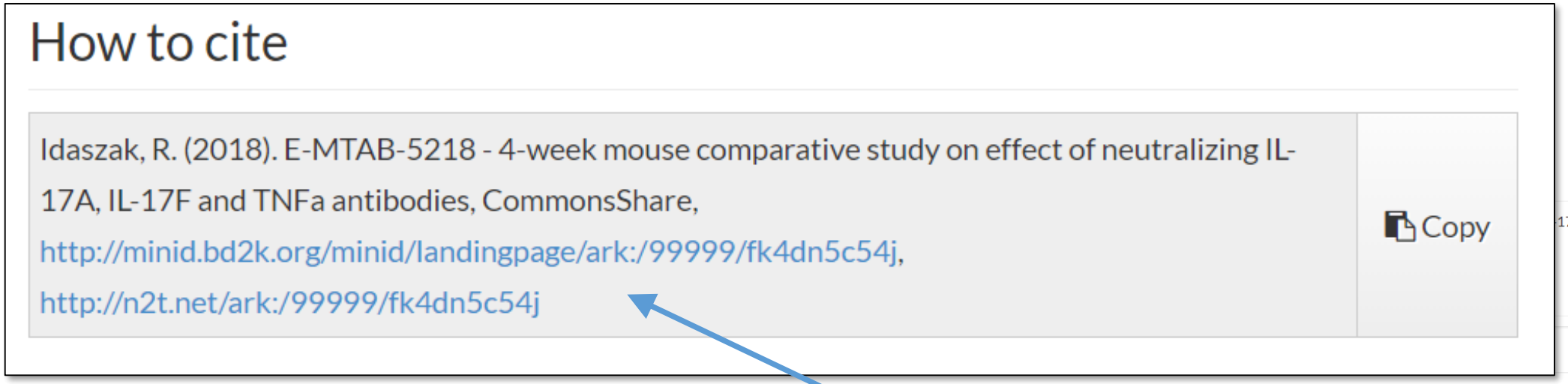
- Upon Publication, the “How to cite” on the resource landing page changes from the GUID to the minid and N2T URIs

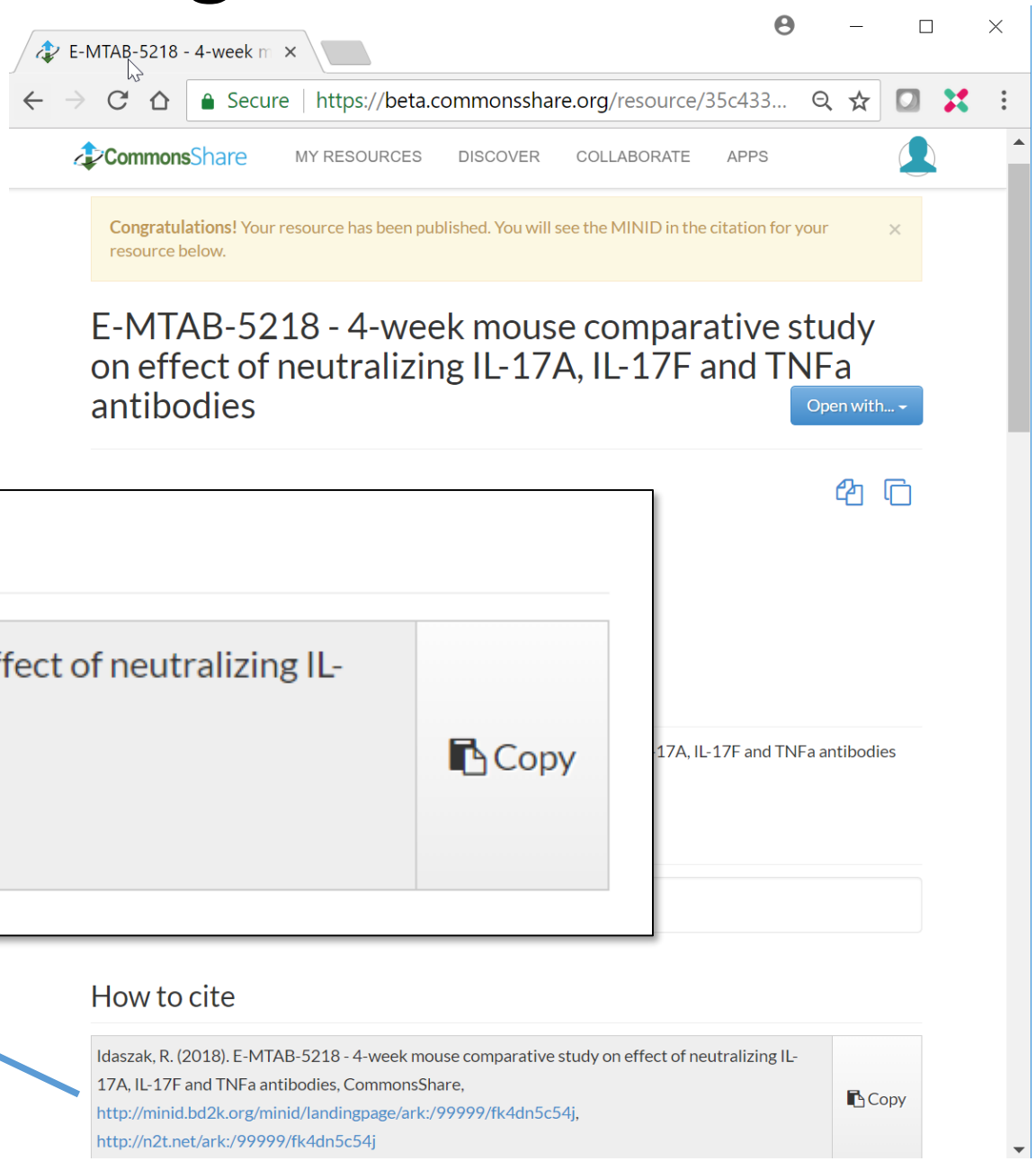




CommonsShare minting a minid

- The resource then becomes Published with a minid and N2T identifier
  - All editing functions and icons are removed from the published resource's landing page ensuring its immutability
- As stated earlier, Helium team uses N2T.net as the primary resolver to resolve identifiers globally

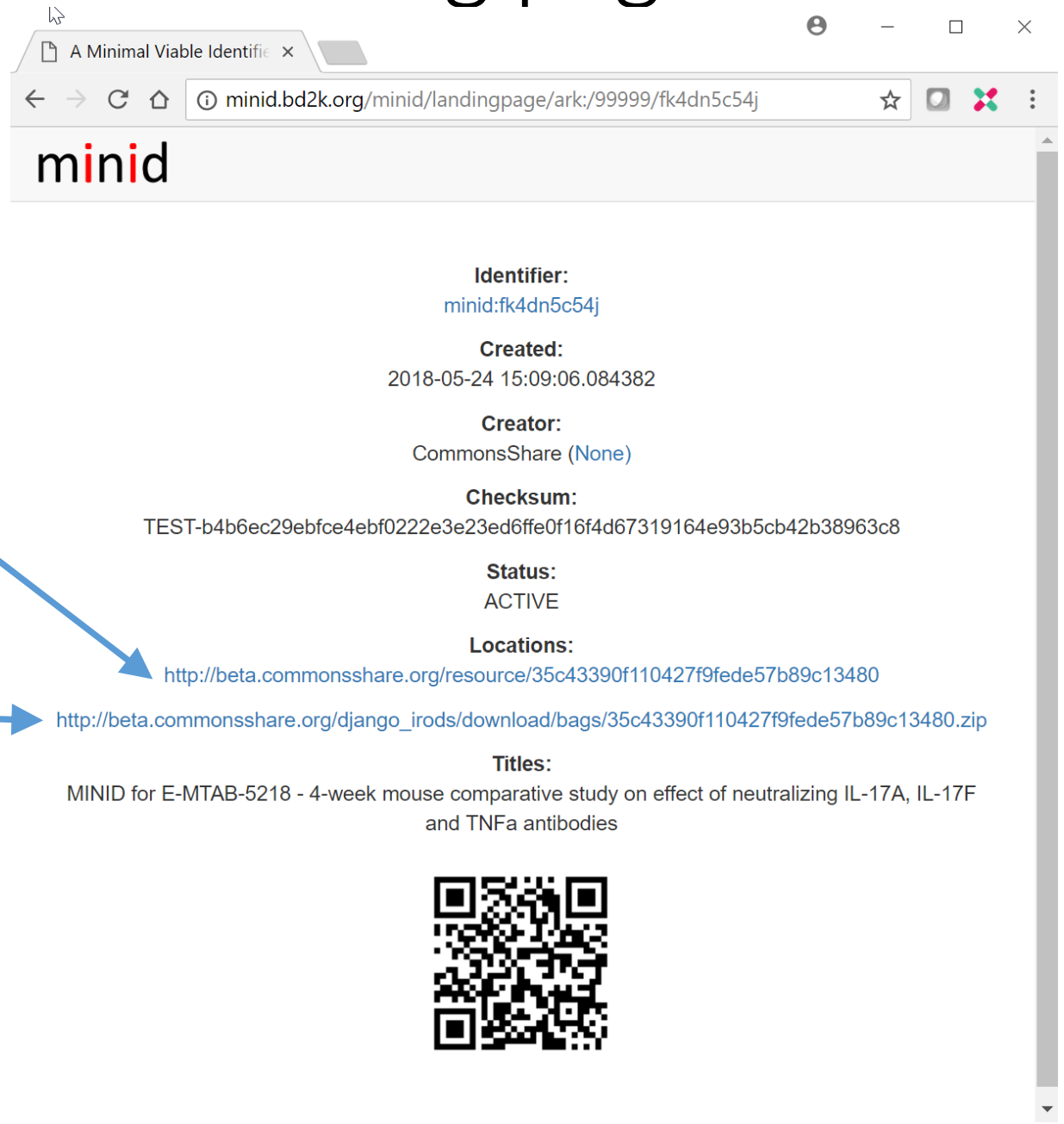




CommonsShare minid landing page

Link back to CommonsShare resource landing page
{Note: in discussion after this presentation, it was pointed out a new field needs to be created on the minid landing page for this type of location reference and will consequently be pursued}

BDBag .zip'ed contents



CommonsShare Minid & BDBag Future Work

- Helium Team Bring Your Own Data (BYOD)
  - Support for Open Science Framework, Dropbox, and additional support for Globus
  - Support for ingest from other iRODS zones
- Extended metadata considerations
  - CommonsShare supports entering extended metadata directly on the resource landing page
  - Aligning metadata models
- Emerging CommonsBag considerations