minid: A BD2K Minimal Viable Identifier Pilot

Draft of November 7, 2015

For more information: http://minid.bd2k.org and info@minid.bdk.org

Overview

Biomedical researchers generate many digital objects of different types, sizes, and quality. Current practice provides for publishing selected data products into repositories, obtaining a unique and persistent identifier (e.g., a DOI) for each such product, and recording these identifiers in registries such as BioCADDIE[1]. But what about the many intermediate, often transient, data products that are created during a research project? They are surely not all worthy of preservation and indexing. But imagine if every such data item had a unique identifier that, furthermore, could be resolved to access some basic metadata. Researchers would then be able to refer unambiguously to any data item, embed references to any data item in different records (e.g., audit logs, provenance records), and know how to access data where access is possible. Moving to a culture where we pass around unique and persistent identifiers to data and know what these identifiers mean seems likely to have a transformative effect on science.

Achieving this goal will require an environment in which digital objects are uniformly and robustly identified, and in which identifiers are used consistently throughout the research lifecycle. We hypothesize that the key to achieving this goal is to provide tools that make the creation and use of identifiers trivial. These tools will focus on just the simplest activities: creating identifiers, resolving identifier to access metadata, and updating metadata. They will reduce each of these actions to a simple API call that requires no forethought to use. We can then embed these API calls into our tools to achieve uniform, universal naming.

We coin the term **Minimal Viable Identifier** (**minid**) to denote an identifier that is sufficiently simple to make creation and use trivial, while still having enough substance to support our goals of more easily findable, accessible, interoperable, and reusable (FAIR) data[2]. We present some proposed **minid** properties and a potential approach to providing those properties. We have launched a lightweight pilot project to gain practical experience with the concepts and tools that we present. We welcome discussion on these ideas and participation in the pilot project.

A Minimal Viable Identifier Service

We conceive of **minid**s as being supported by a **minid** service that:

- Allows an authorized individual to request a new **minid**, supplying minimal metadata, and to obtain an identifier in response;
- Creates a simple landing page for each **minid**, containing user-supplied metadata (who, what, where, when) in human- and machine-readable form;
- Supports resolution of an **minid** to the associated landing page URL;
- Allows **minid**s to be upgraded to DOIs or other unique and persistent identifiers, if/when desired; and

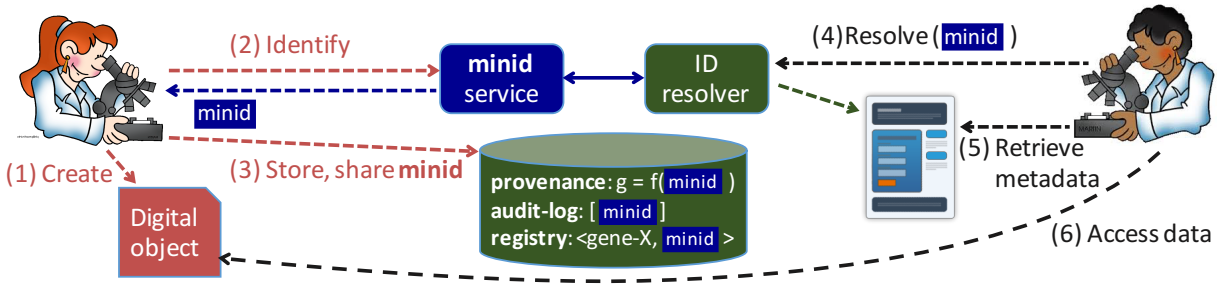
[1] https://biocaddie.org

[2] https://www.force11.org/group/fairgroup/fairprinciples

- Works for both individual files and for multiple files aggregated into a research object

This service might be used as shown in the figure below:

- A researcher creates a new digital object that she may want to reference and/or share later. Her tools register it automatically with the service to obtain a **minid**
- Her tools further embed the **minid** within audit logs and provenance records that she may later share with a colleague.
- A colleague coming across a **minid**, e.g., in a registry, audit log, or provenance record, resolves it to access the landing page, and thus accesses the associated metadata and data.



Note that we only need to build the blue parts in this figure: the **minid** service and the APIs for accessing it. For everything else, we can leverage existing tools, standards, and infrastructure. Thus, we can easily have something operational in a 3-4 month time frame.

We believe that this approach will advance FAIR data by making data more:

- Findable: **minid**s allow researchers to pass around names for data with well-defined meaning, rather than, as at present, passing volatile URLs, sharing Dropbox folders, or (heaven forbid) sending email attachments
- Accessible: **minid**s provide a mechanism for accessing addresses and checksums that can be used to determine whether supplied bits are those originally identified
- Interoperable: **minid**s provide an interoperable mechanism for resolving identifiers to basic metadata
- Reusable: **minid**s provide a substrate on which can be built methods for quality control, indexing, etc., to enable data reuse

Note that we are not talking about an alternative to conventional DOIs and registries, but a stepping stone towards them. By increasing the number of intermediate data products that have identifiers, we simplify the subsequent tasks of identifying and defining provenance of published digital objects.

Proposed approach

There are three things to be done:

1. Define a streamlined method for creating identifiers for data.
2. Define simple policy around the use of such identifiers. For example:
    - A user MUST authenticate before creating a **minid**
    - A user SHOULD provide basic metadata when registering a **minid**: who (identity), what (basic description, checksum), when (time), and where (URL)
    - A user MUST follow specified conventions for any metadata provided
    - A user SHOULD update WHERE metadata if digital object location changes
3. Implement and deploy in an operational ecosystem consisting of infrastructure, tools, and data

Note that we are not proposing a new standard or even a process towards a standard. This is an experiment; we aim to have working code in a few weeks and substantial experience within six months.

Further notes

In defining an identifier service, we follow policies advocated by McMurry et al.[3]

Naming infrastructure: In this pilot we propose to use the existing EZID naming infrastructure for managing identifiers. EZID supports both Archive Resource Keys (ARKs) for rapid identifier creation and Digital Object Identifiers (DOIs) for long-term citations. ARK and DOI identifiers can be linked, and DOIs connected to DataCite. These identifiers are actionable in that they can be resolved using Web protocols to retrieve basic provenance data and additional metadata, such as what storage providers have the actual bits. We have an EZID account and can easily create millions of identifiers in both name spaces.

Aggregations: We believe that it is also important to be able to capture data sets (aggregations). Here, we are exploring the Bagit format[4], a draft Internet standard that we are working with other groups (Goble group in the UK, BagIt standard authors) to adapt to big data. Bagit defines how to (a) list a manifest of files, (b) specify simple metadata, and (c) organize files in a directory structure; it can include pointers to data that are not directly included in the bag (a "holey bag"). We see Bagit as integrating naturally with **minid**s: each element in a bag can have a **minid** and a bag itself can be given a **minid**. Thus, we can provide all of the benefits of identifiers to collections of files (data sets).

Policies: We will define an initial and minimal set of policies for name space management, versioning notation, and data landing pages. Policy will address minimal metadata models for ARKs and DOIs. At early stages metadata might be thin, encompassing perhaps just author, creation time, contact information, and description. Over time, we may want to allow for increasing degrees of annotation, incorporating tools and vocabularies such as those being developed by CEDAR and within initiatives such as DataCite, which has extended the DOI metadata schema for data[5] and bioCADDIE working group 3, which is developing a metadata schema for indexed data objects[6]. We will also want to interface with identifier activities such as bioCADDIE identifier working group 2 and the FORCE11 DataCitation Implementation Group.

Tools: We are developing a minimal tool set for minting identifiers, binding to minimal provenance data (description, author ORCID, dates, etc.), computing hash codes, etc. See below for an initial description.

Versioning: We are not tackling versioning: if two digital objects give different checksums, they will have different **minid**s. (But one could build a versioning system on top of **minid**s.)

[3] https://zenodo.org/record/31765
[4] https://tools.ietf.org/html/draft-kunze-bagit-06
[5] https://schema.datacite.org
[6] https://biocaddie.org/group/working-group/working-group-3-descriptive-metadata-datasets

Prototype

In order to validate that current infrastructure can be used to rapidly create and deploy **minid**s, we created a small **minid** creation and resolution prototype. This prototype leverages the EZID identifier and resolution service operated by the California Digital Library, which supports both ARKs and DOIs. Our prototype provides a command line client for creating **minid**s. Once created, **minid**s can be resolved using a web browser or a programmatic interface. For example, given a data file "`Foster-PublicationSupplement_Final.pdf`," a **minid** can be created with the following command:

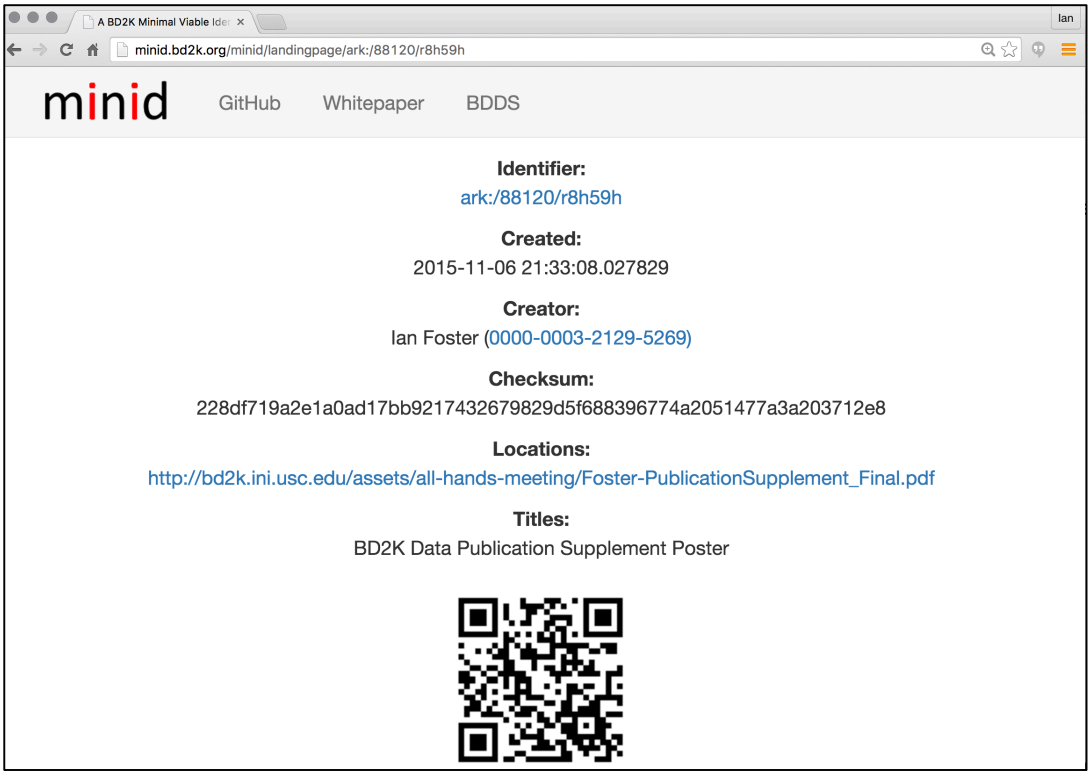
```
% minid Foster-PublicationSupplement_Final.pdf --register --title "BD2K Data Publication Supplement Poster"
```

The prototype combines the title with other minimal metadata (creator name, ORCID, creation date, checksum, data location) and uses EZID to create and return the new identifier, in this case an ARK:

```
Created new minid: ark:/88120/r8h59h
```

A programmatic interface (i.e., a simple REST API) can also be used to achieve the same result.

Once registered, a **minid** can then be resolved, for example by the EZID Name-to-Thing (N2T) resolution service. For example, going to `https://n2t.net/ark:/88120/r8h59h` will redirect the browser to the landing page for the object with our new **minid**, as shown in the screenshot.



A **minid** can also be resolved programmatically in a machine-readable JSON representation, for example:

```
% curl -i -L -H "Accept: application/json" https://n2t.net/ark:/88120/r8h59h
```

which returns the following JSON document:

```
{
  "checksum": "228df719a2e1a0ad17bb9217432679829d5f688396774a2051477a3a203712e8",
  "created": "Fri, 06 Nov 2015 21:33:08 GMT",
  "creator": "Ian Foster",
  "identifier": "ark:/88120/r8h59h",
  "locations": [
    {
      "created": "Fri, 06 Nov 2015 21:33:08 GMT",
      "creator": "Ian Foster",
      "link": "http://bd2k.ini.usc.edu/assets/all-hands-meeting/Foster-
PublicationSupplement_Final.pdf",
      "uri": "http://bd2k.ini.usc.edu/assets/all-hands-meeting/Foster-
PublicationSupplement_Final.pdf"
    }
  ],
  "orcid": "0000-0003-2129-5269",
  "titles": [
    {
      "created": "Fri, 06 Nov 2015 21:33:08 GMT",
      "creator": "Ian Foster",
      "title": "BD2K Data Publication Supplement Poster"
    }
  ]
}
```


Summary

The identification of data is not a task to be performed only at the end of the research data lifecycle, but throughout. By providing unique names for every version of code and data as it is produced, we can facilitate the linking of data to the resources from which it was derived, and thus make research more reliable and reproducible.

To this end, we propose to prototype a shared set of tools, methods, and infrastructure for providing digital objects with permanent identifiers so that (a) they can be reliably referenced and (b) researchers can ensure that a dataset's content has not been modified since its registration. The focus of this work is on the creation and curation of names and minimal metadata. Data itself may not be permanent and indeed may be transient; nevertheless, these methods will still ensure that data consumers get what they expected when they access data.

Acknowledgments

These ideas grew out of discussions at the BD2K California Workshop held at Palm Springs, CA, on October 9-10, 2015. We gratefully acknowledge the support of the NIH Big Data to Knowledge program.