ID Assignment Checklist Data Type Example 2

Data Type: "Dynamic" dataset (actively updated, but available for citing in order to facilitate analysis -- the results of which may then be published), e.g., time-series data.

Use Case: A colleague on a collaborative research project wants to publish findings from the project and needs to get an ID for a snapshot (or version) of the entire, continually updated dataset that has been analyzed to create the findings. As a project collaborator, the colleague does have the right to register identifiers for this data.

Ask the research support staff at your institution what are the standard ID schemes used by your community or data domain.	The most commonly used ID schemes for data that are intended for publication are: ➢ Digital Object Identifiers (DOI) ➢ Archival Resource Keys (ARK)
Check to make sure that the ID schemes used by the community are flexible enough to use for the types of datasets you and your collaborators produce.	The ID schemes for these kinds of data often use the DataCite Metadata Schema which requires both a DOI identifier and a "resourceTypeGeneral" which would be "Dataset" for the time series data. See *DataCite-Metadata Kernel*, v4.4, pg. 48 for definitions of this property. Other schemas commonly used include Dublin Core and Dublin Kernel.
Check data similar to yours to see whether the ID schemes have been successfully applied to the use cases that you anticipate for your data. For example, can the IDs be used ➢ At varying levels of granularity?	For this use case, the recommendation is usually made to assign an identifier at the dataset level but include enough information about the snapshot (or version) to differentiate it from another snapshot. Different identifier schemas and different identifier registration organizations have different solutions for acknowledging the version of the dataset that represents either the particular time frame covered by the snapshot needed for the published findings, or for the version of the dataset to which the

	findings refer. Solutions may include assigning a minor version number to the dataset each time it is republished, so long as ongoing additions don't change pre-existing records in a scientifically significant way as decided by the publisher.[1] Other solutions may include a time range designation for the snapshot of the dataset.[2] In either case, the time designation should be included in the data citation, and described clearly on a landing page for each version of the dataset. Examples of landing pages can be found in the endnotes below.
➢ Or extended to reflect the relationships among disparate data entities?	The ID schemes for these kinds of datasets usually have ways of describing relationships among versions that include describing a minor version to be the "same as" a previous version if the data is scientifically the same, or if there is more than one copy of the dataset at different locations, for instance, or "version of" if the dataset contains new information. Another method used by the DataCite schema when assigning DOIs involves a "RelatedItem" property which would allow for the DOI of the earlier or later version of the dataset [*DataCite-Metadata Kernel*, v4.4, pg. 9]. The ARK ID scheme allows for both subsetting of a data object and versioning. See ARK Identifiers FAQ.
Ask if the infrastructure and services of the allocation agent that provides and registers the IDs has sustained organizational & community support. For ex., see below:	DOIs are probably the most commonly used ID scheme for this kind of data type and use case, and are accepted by most data repositories and archives. ARKs are also used quite widely for datasets that are versioned or dynamic.

---

[1] See GBIF IPT User Manual at: https://ipt.gbif.org/manual/en/ipt/2.5/versioning. Also note on this page, more complete instructions for the kind of information that should be included on time series datasets landing pages.

[2] An example for a time series range might be: "Time subset: 1991-09-01 to 2004-01-01." For an example of a landing page that points to various versions of a time series dataset, see: https://catalogue.ceda.ac.uk/uuid/e0659b01259145c8bfb0de6eb12c2690

➢ Do the data repositories or archives where you plan to store your data for the longer term support these IDs?	You can find both general and domain specific data repositories and archives to find out which identifiers are accepted by going to:
	   ➢ Scientific data's Data Repository Guidance (from nature.com)
	   ➢ DataCite's Repository Finder
	   ➢ FAIRsharing registry

➢ Do the publishers that you intend to use accept these IDs?	Most publishers accept DOIs; from the DOI Factsheet, DOIs are "currently used by well over 5,000 assigners, e.g., publishers, science data centres, movie studios, etc." See https://www.doi.org/factsheets/DOIKeyFacts.html
	
	As of September 2020, an estimated 8.2 billion ARKs have been created by more than 725 registration agents. They are used as permalinks (and identification) for (among other publishing entities):
	   ➢ the Data Citation index (linked to the Web of Science)
	   ➢ Wikipedia articles
	   ➢ Wikidata records
	   ➢ Internet Archive collections
	   ➢ ORCID researcher profiles
	   ➢ See an interactive map of registration organizations for ARKs at ARK Identifiers FAQ

➢ Does the registration agent provide training or help with the description (metadata) requirements or other technical questions?	Training / help in meeting the metadata requirements for DOIs and ARKs can often be found by checking with the research support staff at your institution, from DataCite and CrossRef (commonly used registration agents for DOIs in North America) documentation, ARK Alliance Resources, and tutorials and other learning resources from training catalogues such as ESIP's Data Management Training Clearinghouse.

Pull together the important information required by the ID scheme to describe & make your data more FAIR (Findable… Accessible… Interoperable…Reusable) ➢ What's important for finding your data in a catalog or archive and making it accessible over time?	When you apply to a registration agency to create a DOI or an ARK, you'll need to provide the following information (See more complete definitions from the DataCite Metadata Scheme, pgs. 12 – 17): ➢ **Creator**: Your name or name of the primary creator of the data object ➢ **Title**: What it's called ➢ **Publisher**: Organization making the data publicly available ➢ **Publication Year** ➢ **Resource Type** (general description of the data type using a controlled vocabulary) Although expressed in XML, you can see examples of what kind of information should be included, in general for a dataset.
➢ Include the contextual information most important for others to reuse your data & more easily interoperable	Other information that may make your data more reusable and interoperable would be: ➢ **Language** in which the text/documentation is expressed ➢ **Size and format** of the file(s) (if they are intended to be downloaded) ➢ **GeoLocation** information is important to understand the context of visualization or for reuse ➢ **Rights/license for reuse** (This information may be dictated by your institution, repository or publisher.)