# Madrigal Distributed Data System Architecture and Features: Full Implementation of FAIR Guiding Principles

Bill Rideout and Phil Erickson
MIT Haystack Observatory
July 1, 2020

The document "The FAIR Guiding Principles for scientific data management and stewardship" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/) describes goals for scientific databases. The CEDAR Madrigal database architecture and implementation meets and exceeds FAIR guiding principles in all aspects. For easy community reference, we describe the FAIR relevant and specific Madrigal implementation features and design principles below.

***FAIR Guiding Principle summary***

The FAIR Guiding Principles are summarized as follows:

**To be Findable:**

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

**To be Accessible:**

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

**To be Interoperable:**

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

**To be Reusable:**

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

***Madrigal implementation of each FAIR principle***

**Findable**

CEDAR Madrigal has automatically assigned each data file a globally unique identifier based on w3id.org redirects. CEDAR Madrigal uses a web interface to allow users to see both the unique identifier and a full reference to the data file. Because CEDAR Madrigal uses a w3id.org redirect, this unique identifier can be maintained even should another future organization or system take over the curation duties of the CEDAR Madrigal database.

The CEDAR Madrigal system does not allow old data to be replaced. This ensures the file pointed to by an identifier will always exist. Instead, the data can be updated with a new version, and the old version marked as "history" status, allowing data to be updated without any loss of reproducibility, since both updated and history versions of the file are available.

The entire CEDAR Madrigal database was designed around rich, shareable metadata. This was done so that a distributed database can appear to a user as a single database. The metadata from each site is shared to every other site, allowing each site to be searched as though it contains all data. The metadata itself is defined publicly in the documentation. The metadata is thus open, free, and universally implementable. Furthermore, the software to install the entire Madrigal database is free, open, and easily available. (Note that the CEDAR Madrigal database archives all public data at the remote data sites to increase data security and provide long term archiving.)

Standardization of CEDAR Madrigal database metadata goes beyond user interaction with the web interface, since this standardization allows API's to access all CEDAR Madrigal metadata and data. These API's are available in python, Matlab, and IDL, and have been stable since 2003. Indeed, a script written in 2003 will still work today.

**Accessible**

As described above, all CEDAR Madrigal data can be accessed with persistent, globally-unique identifiers, either through the web site or through APIs. The protocol is standard secure https transport. All data on the CEDAR Madrigal database is fully available to the world, so no authentication is necessary. However, all users are required to submit their names, emails, and affiliations when accessing data files so that the data Principal Investigators (PIs) know who has accessed their data. This information is not verified, a fact which has never been raised as an issue by any instrument PI. However, if needed, remote Madrigal sites do have a way of making their data private so that only local users can access it. That private data is not imported into the CEDAR Madrigal database until its status changes to public.

Since the data is all backed up at the CEDAR Madrigal database and cannot be overwritten, the FAIR principle that metadata should remain even if the data should disappear is not applicable, as the CEDAR Madrigal database does not allow data to ever disappear.

**Interoperable**

For the CEDAR Madrigal database, interoperability has been implemented as a core property. Community use cases have been successful in demonstrating the utility and flexibility of this operating principle which is also coded in FAIR principles. In particular, the first virtual observatory that directly accessed CEDAR Madrigal was the European effort STAP, implemented in 2007 entirely using the fully public API. This particular interface used the python API, but either the Matlab or IDL interfaces could have been used if desired.

**Reusable**

The richness of the publically available Madrigal metadata is demonstrated both by the richness of the public APIs that use it, and its use in interfaces to virtual observatories. Because the metadata are well-defined, additional APIs can be written by users in other languages, as was done for e.g. Mathematica. The API also offers access to the provenance of the data, indicating whether it is a historical file or the latest release.

The CEDAR Madrigal data format is stored in the standard HDF5 format. All parameters used in the files are standardized, such that the file can be readily analyzed by software agents. This allows users of the CEDAR Madrigal database to request data files with additional derived parameters, such as magnetic models. It also allows users to search the entire database according to specific data criteria.

The CEDAR Madrigal system also implements a core principle which always makes its data files fully self-describing, which is a considerable superset of FAIR Specifically, using this approach, a user with no knowledge of the CEDAR Madrigal data format could simply examine the Hdf5

file itself without any API infrastructure and fully understand the scientific information in the file. The file defines all parameters and units, and it provides documentation about the data PI and other metadata about the file.