
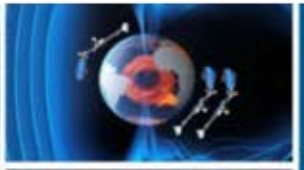


GOOD HOPE
FOR EARTH SCIENCES
27 August - 1 September 2017
Cape Town, South Africa
IAPSO-IAMAS-IAGA

Using Digital Object Identifiers in INTERMAGNET

Simon Flower
British Geological Survey (NERC), Edinburgh, UK

What is a Digital Object Identifier (DOI)?

“The DOI system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers, called DOIs, for use on digital networks.”

“What are we identifying with this identifier? …whether two things being identified are "the same thing“… This deceptively easy question is one of the most difficult encountered in all discussions about identifiers (but the one most commonly overlooked)”

https://www.doi.org/index.html

ISO 26324

What's important in our community?

In our community DOIs should:

- Allow scientists to cite data in their publications
- Give recognition & acknowledgement for data providers
- Allow discovery and easy online access to research data
- Include metadata to uniquely identify the dataset and provide relevant information to the user
- Allow re-use of the data in a reproducible way

DOIs in INTERMAGNET: Goals

Create a system that:

- Allows institutes to receive acknowledgement for their work in creating data
- Is realistic for scientists and publishers to use
- Conforms to DOI principles [referencable, reproducible, …]

Data citation principles: https://www.force11.org/group/joint-declaration-data-citation-principles-final

DOIs in INTERMAGNET: Guiding principles

- INTERMAGNET does not own any data. Data and acknowledgement belongs to INTERMAGNET's members, the data providers
- Concentrate first on definitive data
- INTERMAGNET is a network, not an institute. It needs to proceed with the agreement of it's members
- We can't think about DOIs without considering how data is licensed

A related topic: Evolution of INTERMAGNET data policy

The license for using data released through INTERMAGNET needs to be updated. It's important that the updated license is:

- Agreed by the members
- Recognised globally
- Machine readable

Options:

- Creative Commons (6 permutations)
- Open Data Commons (2 permutations)

DOIs: Evolution not revolution

Conditions of Use

The member institutes of INTERMAGNET invest considerable resources to operate their magnetic observatories to INTERMAGNET standards. It is important that the institutes producing the data have a measure of the scientific return on their investment. Accordingly, we have the following Conditions of Use.

Conditions of Use for data provided through INTERMAGNET

The data made available through INTERMAGNET are provided for your use and are not for commercial use or sale or distribution to third parties without the written permission of the institute operating the observatory. Publications making use of the data should include an acknowledgment statement of the form given below. A citation reference should be sent to the INTERMAGNET Secretary (secretary@intermagnet.org) for inclusion in a publications list on the INTERMAGNET website.

Acknowledgement of data from observatories participating in INTERMAGNET

We offer two acknowledgement templates. The first is for cases where data from many observatories have been used and it is not practical to list them all, or each of their operating institutes. The second is for cases where research results have been produced using a smaller set of observatories.

Suggested Acknowledgement Text (template 1)

The results presented in this paper rely on data collected at magnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice (www.intermagnet.org).

Suggested Acknowledgement Text (template 2)

The results presented in this paper rely on the data collected at ***observatory name***. We thank ***institute name***, for supporting its operation and INTERMAGNET for promoting high standards of magnetic observatory practice (www.intermagnet.org).

1st possibility: INTERMAGNET creates DOIs

- INTERMAGNET creates DOIs for INTERMAGNET accredited definitive data (e.g. 1 DOI for each existing CD/DVD).
- These INTERMAGNET DOIs reference DOIs that institutes may have created for their data.

PROs	CONs
Simple to manage	Primary DOI does not reference the data provider
	May require moving data to a new INTERMAGNET repository
Do institutes want INTERMAGNET to create DOIs?	

2nd possibility: INTERMAGNET harvests DOIs

- Institutes create their own DOIs
- INTERMAGNET offers to help create DOIs
- Institutes record DOIs with INTERMAGNET
- INTERMAGNET maintains a DOI database for INTERMAGNET observatories
- When users search for data on an INTERMAGNET portal they can also retrieve DOIs associated with the data

PROs	CONs
DOIs directly reference institutes	Complex to manage (but possible)
	Data referenced by institute's DOI may be different to data at INTERMAGNET

3rd possibility

?

Practicalities

Large data citation lists are inevitable. We need to work with publishers to ensure this is viable.

We need to store DOIs in a searchable database. The Geomagnetic Community Metadata provides for this, but the system needs to be implemented.

Research Data Alliance: Data Citation of Evolving Data. Should we conform with this?

https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf

Future work

This describes the ‘easy’ problem – definitive data. We also need to think about raw data and provisional data products.

RDA Recommendations (1)

A. Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

R1 – Data Versioning: Apply versioning to ensure earlier states of data sets can be retrieved.

R2 – Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.

R3 – Query Store Facilities: Provide means for storing queries and the associated metadata in order to re-execute them in the future.

B. Persistently Identify Specific Data Sets

R4 – Query Uniqueness: Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.

R5 – Stable Sorting: Ensure that the sorting of the records in the data set is unambiguous and reproducible

R6 – Result Set Verification: Compute fixity information (checksum) of the query result set to enable verification of the correctness of a result upon reexecution.

https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf

RDA Recommendations (2)

R7 – Query Timestamping: Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at the time a user issued a query.

R8 – Query PID: Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID.

R9 – Store Query: Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.

R10 – Automated Citation Texts: Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.

R11 – Landing Page: Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.

R12 – Machine Actionability: Provide an API / machine actionable landing page to access metadata and data via query re-execution.

https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf

Possible licenses

License	Variant	Details
Creative Commons version 4.0		
	CC BY	Does not protect against commercial use
	CC BY-SA	Mandates that the same license is applied to derived products Does not protect against commercial use
	CC BY-ND	Prevents new products being developed from the original data Does not protect against commercial use
	CC BY-NC	Allows new products to be developed from the original data Protects against commercial use
	CC BY-NC-SA	Mandates that the same license is applied to derived products Protects against commercial use
	CC BY-NC-ND	Prevents new products being developed from the original data Protects against commercial use
Open Data Commons		
	ODC-BY	Does not protect against commercial use
	ODC-ODbl	Mandates that the same license is applied to derived products Does not protect against commercial use