Introduction to (data) identification & citation

Maggie Hellström

ICOS Carbon Portal & Lund university

DATA FAIRNESS
LECCE 1-5 JUL 2019 INTERNATIONAL SUMMER SCHOOL FOR ENVIRONMENTAL & EARTH SCIENCE INFRASTRUCTURES

Couplings to FAIR principles

- **F1.** (meta)data are assigned a globally **unique and persistent identifier**
- **F3.** metadata clearly and explicitly **include the identifier** of the data it describes
- **A1.** (meta)data are **retrievable by their identifier** using a standardised communications protocol
- **A2.** **metadata are accessible**, even when the data are no longer available
- **I3.** (meta)data **include qualified references** to other (meta)data
- **R1.** meta(data) are richly described with a **plurality of accurate and relevant attributes**

Internet of Things (IoT)

“Everything” is getting connected

On-going trend to hook up devices to the Net

Two-way communication: the fridge can tell us what to buy, and we can tell it to lower its temperature a bit

Needs a “device registry” with basic metadata (kind of device, owner, URL, …)

Requires all devices to have a unique address

Same for data!

Image © https://corporatefirst.com.au

To share or not to share…

Using what others are sharing

Many researchers want to (and do) use others' data – but face issues:

- Difficult to find relevant data in the first place
- Detailed descriptions missing or unambiguous
- Available formats not directly applicable
- Quality and intended use often unclear
- Citing data sources is seen as complicated (lack of standards)

Y. Socha, ed. (2013). "Out of Cite, Out of Mind".

To share or not to share…

To share with others

Many researchers want to share their data – but they have many questions:

- Unclear where to store data to make them available
- Time-consuming to provide detailed descriptions
- Uncertainty who will pay for storage & effort
- Worries they will be “scoped” and/or not given credit
- No clear way to put “data publication” into their CV

Y. Socha, ed. (2013). “Out of Cite, Out of Mind”.

Identification

- In science we want clarity and reproducibility
- A name or (short) description isn't enough for unambiguous identification
- Persistent and unique "ID number" = PID
  - Pointer to object + some basic information
  - Can be resolved = looked up at any time by anyone
  - End user doesn't have to worry about the exact location
- Not only articles and data should be PID-ed
  - Places & measurement sites
  - Physical samples
  - Photos, videos, audio recordings
  - People
  - Instruments & sensors

Starting point: identity?!

Identifying a person

- “My friends John and Mary”
- Name (“John Smith”, “Mary Jones”)
- Address (“123 Enni Street, Samtown”
- Photograph, biometric data
- Passport or driving license number
- Social security number
- Fingerprint or DNA profile

Identifying a digital object

- “Meteo data from Norunda station, collected on December 11, 2017”
- Name (“SE-Nor_meteo_2017-12-11.dat”, “Matlab.exe”)
- Location or path (“C:\ICOSdata\”, “https:\\docs\icos-cp.eu\api\”
- File size, date, checksum
- Project-internal identifier, including versioning info
- Persistent and unique digital IDentifier (PID)

What can be ID-ed?

- Everything (almost)!
- Commonly used for entities with a digital representation
  - Articles, reports, data (doi or similar)
  - People (ORCID ID) & organizations (ISNI, ORCID)
  - Samples (IGSN, biobank IDs and similar)
  - Instruments, measurement sites, projects etc.
  - Services for computation, storage, …
  - Software
  - Model algorithms and model runs (parameter sets)
  - Variable definitions (types)
  - …

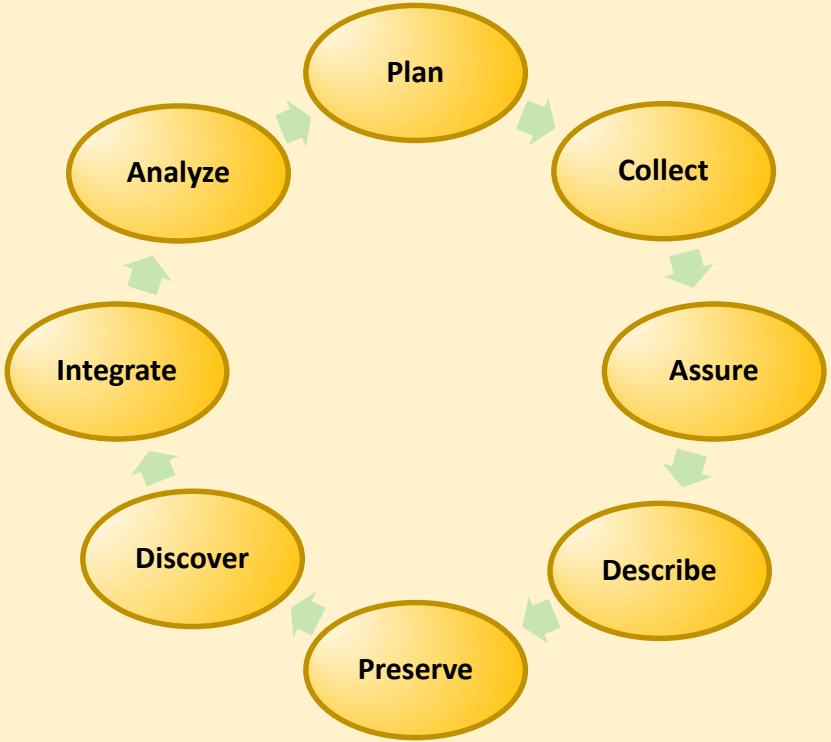
https://orcid.org/
http://www.isni.org/
http://www.geosamples.org/
https://datacite.org/

I&C in the data lifecycle!?

Split up into groups of 3-4 people and discuss (5+5 mins):

A. Where do we need to identify objects and resources?
B. How do we refer to them in the best way?

NOTE: half the groups start with question A, the other half with B!



Persistent?

- Persistent refers to the identifier! (Not necessarily the resource…)
- Persistent often interpreted as
  - Permanent (“for ever”)
  - Immutable (“cannot be changed”)
  - 24/7 accessibility
- Persistent function dependent on
  - Sustainability of registration agency
  - Correctness of metadata (esp. the resource locator)!
  - Existence & accessibility of the related resource and/or its metadata

R.E. Duerr et al., “On the utility of identification schemes for digital earth science data: an assessment and recommendations”, Earth Sci Inform (2011) 4: 139. https://doi.org/10.1007/s12145-011-0083-6 and other sources

Digital?

- Digital refers to the identifier (not necessarily to the resource)
- The resource should have a “digital representation”
  - Many resources & objects – like data – are already digital
  - People, instruments, samples etc. are physical – but can be catalogued
  - The identifier should resolve to a landing page with information about the physical object
- Digital often interpreted as
  - Only usable for digital resources (wrong!)
  - Allows direct access to or retrieval of object (wrong!)

R.E. Duerr et al., “On the utility of identification schemes for digital earth science data: an assessment and recommendations”, Earth Sci Inform (2011) 4: 139. https://doi.org/10.1007/s12145-011-0083-6 and other sources

Unique

- The identifier needs to be globally unique – or there will be confusion!
- Important to consider when “translating” from one PID system to another
  - For example when basing a Handle PID on a project-internal identifier that might not be unique
- A given PID should never be reused for “another object”
  - But it may be OK to use the same PID for an object that itself changes over time!

R.E. Duerr et al., “On the utility of identification schemes for digital earth science data: an assessment and recommendations”, Earth Sci Inform (2011) 4: 139. https://doi.org/10.1007/s12145-011-0083-6 and other sources

Handle System

- The Handle System was developed by the US Corporation for National Research Initiatives (CNRI) in the early 1990-ies
- It is a proprietary registry assigning persistent identifiers, or handles, to information resources, and for resolving "those handles into the information necessary to locate, access, and otherwise make use of the resources".
- Since 2014, the DONA Foundation administers the system's Global Handle Registry
- Handles are built from a prefix which identifies a "naming authority" and a suffix which gives the "local name" of a resource
- The complete string must be globally unique

Wikipedia article “Handle System” (https://en.wikipedia.org/wiki/Handle_System) and ENVRIplus deliverable D6.1 (Hellström et al. (2017))

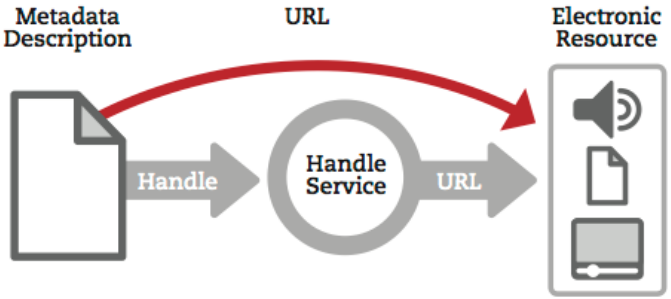
Different types of PID systems

- Archival Resource Keys (ARKs)
- **Digital Object Identifiers (DOIs)**
- Extensible Resource Identifiers (XRIs)
- **Handle System (Handles)**
- Life Science Unique Identifiers (LSIDs)
- Object Identifiers (OIDs)
- **Persistent Uniform Resource Locators (PURLs)**
- Uniform Resource Identifiers/Names/Locators (URIs/URNs/URLs)
- Universally Unique Identifiers (UUIDs)

R.E. Duerr et al., “On the utility of identification schemes for digital earth science data: an assessment and recommendations”, Earth Sci Inform (2011) 4: 139. https://doi.org/10.1007/s12145-011-0083-6

Indirection

- Indirection (also called "dereferencing") is the ability to reference something using a name, reference, or container instead of the value itself.
- This requires that the “something” must be registered in some kind of lookup system or database – a Registry - with enough metadata attached to allow it to be found
- The lookup database must be available to anyone (a human or a computer) wishing to use it



The DOI Handbook and ePIC web site (figure courtesy of CLARIN ERIC).

Resolving (Handle) PIDs

- Resolution is made using a Resolver, a server which may be different than the ones involved in exchanging the handle for the metadata.
- Unlike URLs, which may become invalid if the metadata embedded within them becomes invalid, handles do not become invalid and do not need to change when locations or other metadata attributes change.
- This helps to prevent link rot, as changes in the information resource (such as location) need only be reflected in changes to the metadata, rather than in changes in every reference to the resource.

Wikipedia article “Handle System”, https://en.wikipedia.org/wiki/Handle_System, accessed 2018-06-27

Landing pages

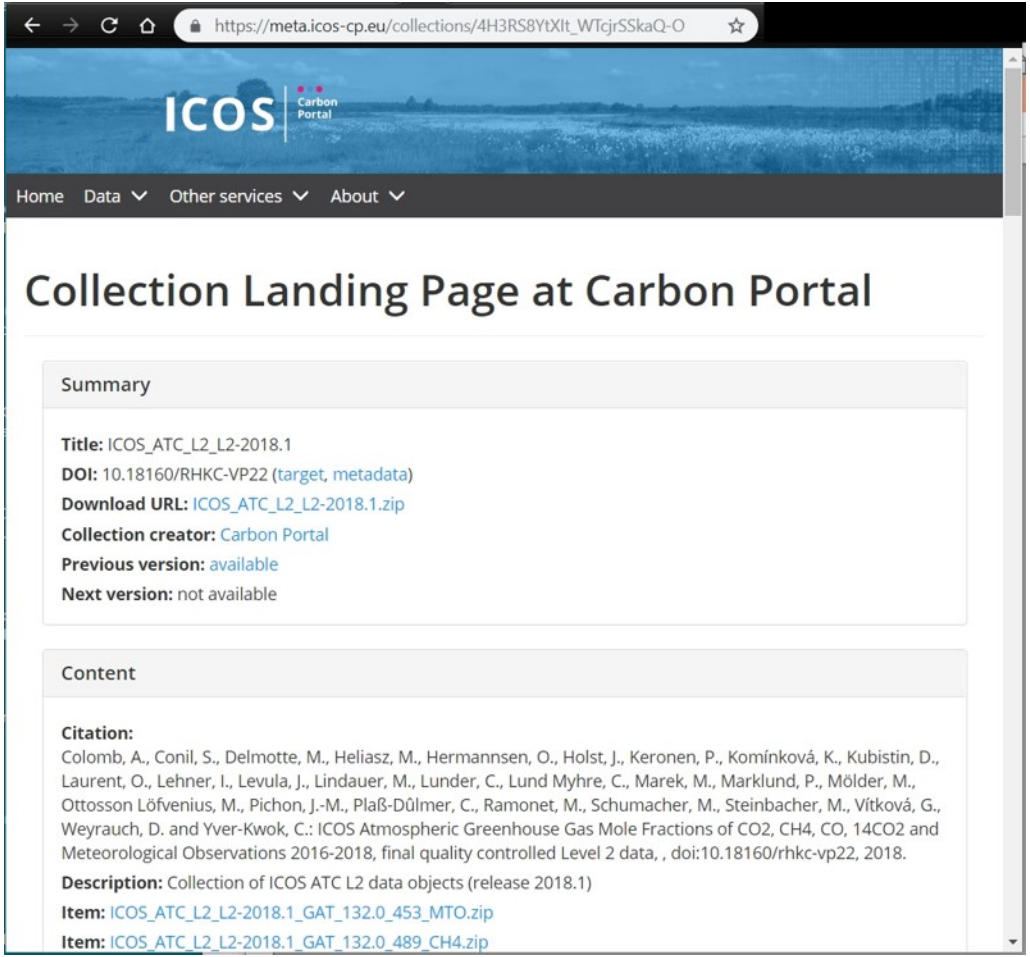
- By “landing page(s)” is meant a set of information about the data via both structured metadata and unstructured text and other information
- The identifier included in a citation should point to a landing page or set of pages rather than to the data itself
  - the metadata and the data may have different lifespans, the metadata potentially surviving the data
  - the cited data may not be legally available to all, even when initially accessioned, for reasons of licensing or confidentiality – so the landing page provides a method to host metadata and/or access credentials can be validated
  - resolution to a landing page allows for an access point that is independent from anymultiple encodings of the data that may be available

J. Starr et al. (2015), "Achieving human and machine accessibility of cited data in scholarly publications". PeerJ Computer Science 1:e1, https://doi.org/10.7717/peerj-cs.1

Best practices for dataset descriptions

- Minimally the following metadata elements should be present in dataset descriptions (in catalogue & on landing pages):
  - Dataset identifier (e.g. a PID, preferably an actionable link)
  - Title & Description
  - Creator - the person(s) and/or organizations who generated the dataset and are responsible for its integrity.
  - Publisher/Contact - the organization and/or contact who published the dataset and is responsible for its persistence.
  - PublicationDate/Year/ReleaseDate & Version
  - License for usage

J. Starr et al. (2015), "Achieving human and machine accessibility of cited data in scholarly publications". PeerJ Computer Science 1:e1, https://doi.org/10.7717/peerj-cs.1



Landing page example

Tombstones

- "Tombstones" are landing pages for objects or other resources that are no longer available
- They should continue to display – or link to – as much of the original metadata as possible
- They should also provide a reason ("cause of death")
- Examples:
  - Records of physical samples that have been destructively tested
  - A data file that has been purged from an archive after a 10-year storage period
  - Large-volume model output files that can be (easily) recreated
  - Keeping the link alive for an accidentally deleted data file

J. Starr et al. (2015), "Achieving human and machine accessibility of cited data in scholarly publications". PeerJ Computer Science 1:e1, https://doi.org/10.7717/peerj-cs.1

Preparing for landing...

Form groups of 4-5 people and do the following (10 mins):

- Check if your own RI or organization uses landing pages. Find an example you can share.
- If not, then check out some landing pages at e.g. the ICOS Carbon Portal (https://icos-cp.eu/)
- Do these examples seems to follow the recommendations?

Common questions

- How do I decide what PID to use?
- Where to get (assistance to get) a PID?
- Can “anyone” register a PID for (my) data?
- What information is required?
- Does it cost anything?
- Who is responsible?
- How do I decide what resolving the PID should “do”?

Selecting the right provider

- A suitable PID type for raw and intermediate data products that should be referrable also outside of a RI or project is the “basic” Handle – as provided by e.g. the European Persistent Identifier Consortium (ePIC)
- For “publication-level” data, the Digital Object Identifier – as provided by e.g. DataCite – is considered to be a very good choice
- Important aspects to consider include the trustworthiness, the sustainability of both the PID registry and the resolver(s) involved, and cost
- Because the cost models of the PID registration authorities and their national members can differ a lot between countries, it is important to investigate this carefully

Publishing data: store & PID in one go!

- Large research organizations may operate their own data repositories. Features often include metadata catalogues, PID minting and web-based discovery services
- Smaller RIs, research groups and individual scientists can use existing data repositories.
- Trustworthiness, reliability and sustainability are important aspects, not just storage capacity and costs
- An important “benchmark” for repository management is the OAIS Reference Model*
- Core Trust Seal (CST) is a certification system for repositories – see https://www.coretrustseal.org/ for more info!

CCSDS (2012), “Reference model for an open archival information system (OAIS)” (a.k.a. the Magenta Book), https://public.ccsds.org/pubs/650x0m2.pdf

ePIC

- ePIC stands for European Persistent Identifier Consortium
- Founded in 2009; current members include CSC, DKRZ, grnet, GWDG, KTH & SurfSara
- Provides PID services for the European Research Community – both minting & lookup
- ePIC PIDs are based on the Handle system
- If a repository wants to issue their own ePIC PIDs:
  - Register their own prefix at Handle.Net (https://www.handle.net/prefix.html)
  - Sign a agreement with one of the ePIC centers
  - Incorporate the ePIC API into the operational data management scripts

The ePIC web site (2018), http://www.pidconsortium.eu/

DataCite

- DataCite DOIs are alphanumeric strings assigned to uniquely identify an object. They are tied to both a metadata description of the object (in the DataCite metadata store) as well as to a digital location, such as a URL (external to DataCite)
- In order to create new DOIs and assign them to your content, it is necessary to become a DataCite member or work with one of the current members. (See list at https://datacite.org/members.html)
- To mint a new DOI, one can either use the DOI Fabrica web interface or the API of the DataCite Metadata Store
- Required information includes a name, a metadata description following the DataCite Metadata Schema and at least one URL of the object.
- Once created, information about a DOI is available through our different services search, event data, OAI-PMH and others).

The DataCite web site (2018), http://www.datacite.org/

The DataCite metadata schema

- DataCite offers users the possibility to upload a rich metadata set for each of the registered objects
- Reasons include enhancing search capabilities and supporting good citation practices
- The metadata schema is described at https://schema.datacite.org/ and https://schema.datacite.org/meta/kernel-4.1/
- Only a small subset of metadata fields are obligatory – but the more information that is submitted, the better
- However, the DataCite metadata store should never be considered the only or the authoritative source of information about the object! (That responsibility should rest with the repository where the data object is stoired.)

Who gives a PID?

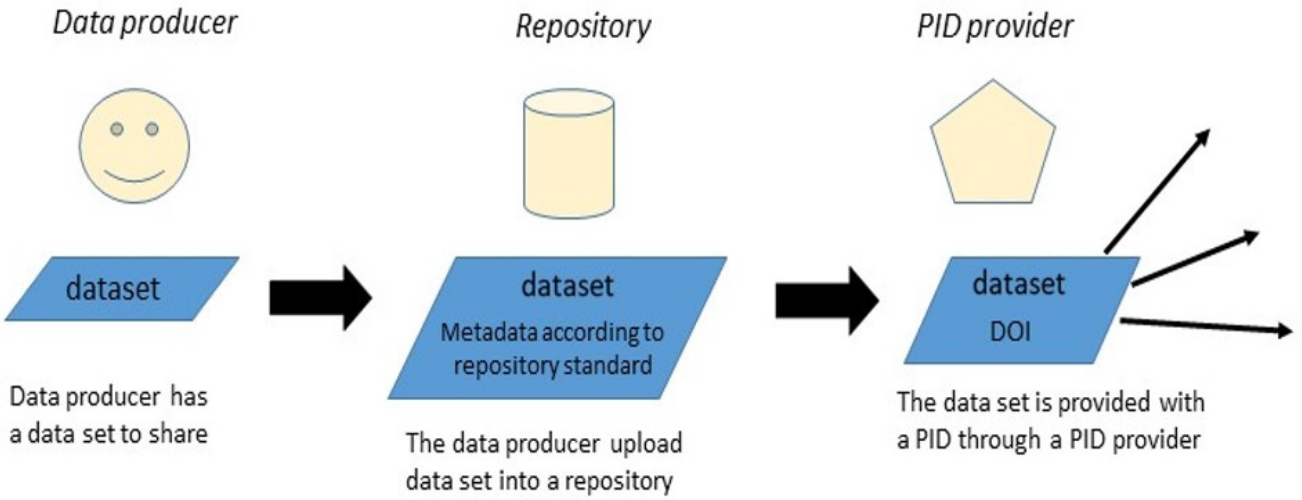
Split up into groups of 3-4 people and discuss (7 mins):

- How does your RI or organization get PIDs for your data sets? Do you create these yourselves, or do you rely on a repository to do this for you?
- Bonus question: Is there a national coordinator for (Handle) PIDs in your country?

Citation (or referencing)

- Give an unambiguous pointer to resources that were used
- Allows studies to be reproduced and validated
- Acknowledge other peoples' work!
- References should be understood by humans and computer processes
- Using PIDs + authors, resource name & relevant dates
- Data citation information should be just as "harvestable" for bibliometrics and usage statistics as references to articles etc.

How to make data citable



Citing a dataset

- Traditionally, datasets that have been used in research are referred to in the text of resulting scientific papers
- If the article authors were the data collectors, often no detailed information about the data sets (names, location, identifiers, etc.) were given
- When data from other sources were used, it might be referred to as “data from Hansson et al., used with permission” and/or by having invited Hansson to be co-author
- In some cases, the data might be mentioned in the bibliography alongside all other references – but not following any standardized formats
- With the increasing use of persistent identifiers for data (and other resources), it should be easy – and natural – to cite these properly

Some examples

Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC)
Gu J.J., E.A. Smith, and H.J. Cooper. 2006. LBA-ECO CD-07 GOES-8 L4 Gridded Surface Ration and Rain Rate for Amazonia: 1999. Data Set. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/831

Federation of Earth Science Information Partners (ESIP)
Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://dx.doi.org/10.5060/D4MW2F23z.

Some examples

GBIF

Chavan, V. S. (1996). Amphibians of the west coast of India. 1223records, published online, http://www.vishwaschavan.in/indfauna/amphibians_west_coast/,released on 12 June 1998, doi:10.5284/1000164.

-or-

http://www.ncbi.org.in/indfauna/ (2012), Hornbills and India, 989 records, accessed on 12 January 2012:22:10:10 hrs, user doi: 99.6672/100.324.2012, publisher doi: 10.3897/ncbi.ncl.2001.

Dryad

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012) Data from: More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters doi:10.5061/dryad.75nv22qj

Potential problems

- Correct citation is dependent on the information about the people (and institutions) involved being correct
- Currently, only those individuals identified as creators (and in certain cases publishers) will be included automatically
- There is no adequate support for an editor role for collections
- People may be mis-identified by automatic citation indexer processes – unless they are clearly identified by their ORCID numbers
- When data portals harvest metadata from other sources and insert into their own catalogues, the information about the original creators of the dataset may be lost, and replaced with the name of the harvesting source

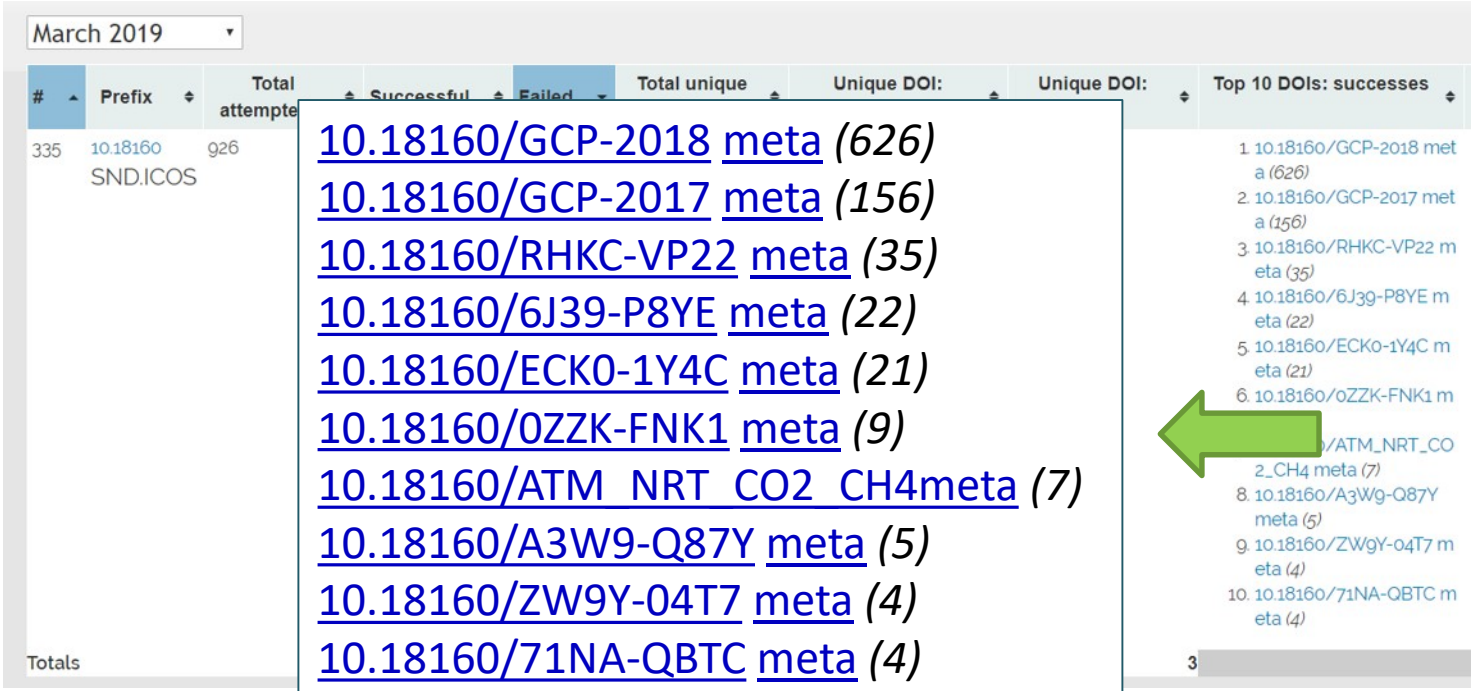
DataCite statistics

March 2019

#	Prefix	Total attempted	Successful	Failed	Total unique DOIs	Unique DOI: successes	Unique DOI: failures	Top 10 DOIs: successes
335	10.18160 SND.ICOS	926	918	8	34	31	3	1. 10.18160/GCP-2018 meta *(626)* 2. 10.18160/GCP-2017 meta *(156)* 3. 10.18160/RHKC-VP22 meta *(35)* 4. 10.18160/6J39-P8YE meta *(22)* 5. 10.18160/ECK0-1Y4C meta *(21)* 6. 10.18160/0ZZK-FNK1 meta *(9)* 7. 10.18160/ATM_NRT_CO2_CH4 meta *(7)* 8. 10.18160/A3W9-Q87Y meta *(5)* 9. 10.18160/ZW9Y-04T7 meta *(4)* 10. 10.18160/71NA-QBTC meta *(4)*
Totals		926	918	8	34	31	3	

Citation statistics for ICOS-minted DOIs (prefix 10.18160), March 2019

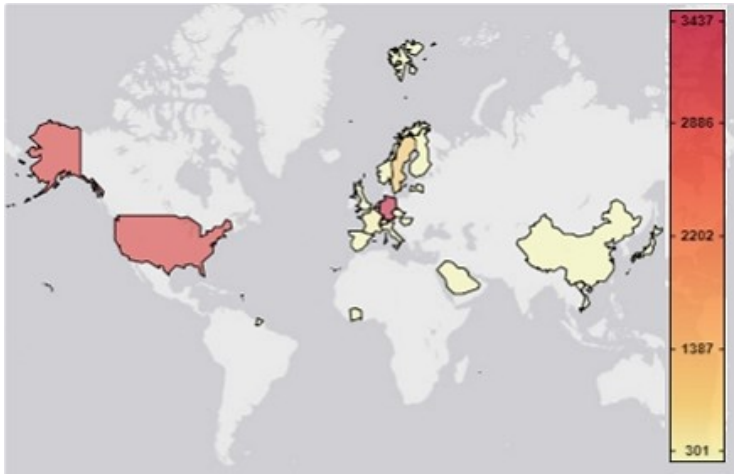
DataCite statistics

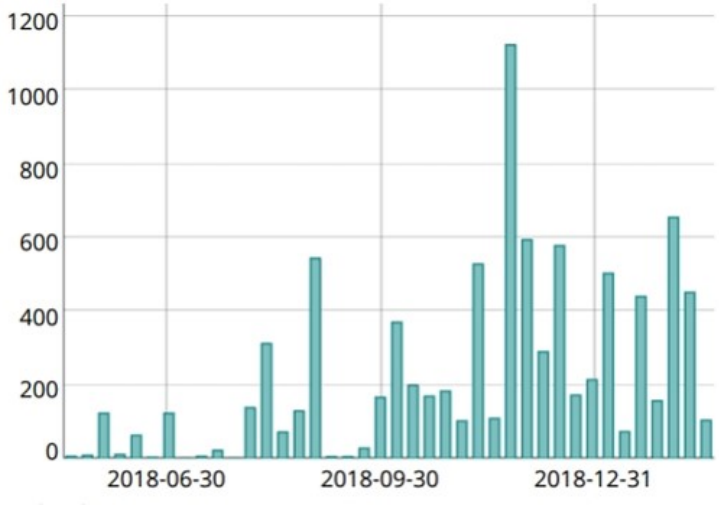


10.18160/GCP-2018 meta *(626)*
10.18160/GCP-2017 meta *(156)*
10.18160/RHKC-VP22 meta *(35)*
10.18160/6J39-P8YE meta *(22)*
10.18160/ECK0-1Y4C meta *(21)*
10.18160/0ZZK-FNK1 meta *(9)*
10.18160/ATM_NRT_CO2_CH4meta *(7)*
10.18160/A3W9-Q87Y meta *(5)*
10.18160/ZW9Y-04T7 meta *(4)*
10.18160/71NA-QBTC meta *(4)*

Citation statistics for ICOS-minted DOIs (prefix 10.18160), March 2019

Download statistics





Geographical (left) and temporal (right) distribution of downloads of all Level 2 data sets (CO2, CH4, CO, 14C and meteorological variables) from the ICOS Atmospheric Thematic Centre March 2017 – February 2018.

Give me credit!

Split up into groups of 3-4 people and discuss (5 mins):

- What data management activities do you want credit for?
- Who should provide the statistics – repositories, your RI, or both?

Wrap-up

We've looked at many things this afternoon!

- Identification and citation in context
- What can be PID:ed
- PIDs: what are they and how do they work?
- Landing pages and tombstones
- Providers of PIDs
- Data citation & download statistics