Credit where credit is due

Ensuring that data producers' efforts are properly accounted for when creating and using data collections

Maggie Hellström, ICOS & Lund University, Sweden
Markus Fiebig, ACTRIS & NILU, Norway
ENVRIplus Theme 2 "Data for Science"

ICOS INTEGRATED CARBON OBSERVATION SYSTEM

ACTRiS

ENVRI plus

ENVRIplus overview

- A Horizon 2020 project (2015-2019) funded by European Commission
- Partners: Environmental and Earth System RIs, projects and networks together with technical specialist partners
- Goal: a more coherent, interdisciplinary and interoperable cluster of Environmental Ris across Europe

"Environmental Research Infrastructures Providing Shared Solutions for Science and Society"

http://www.envriplus.eu

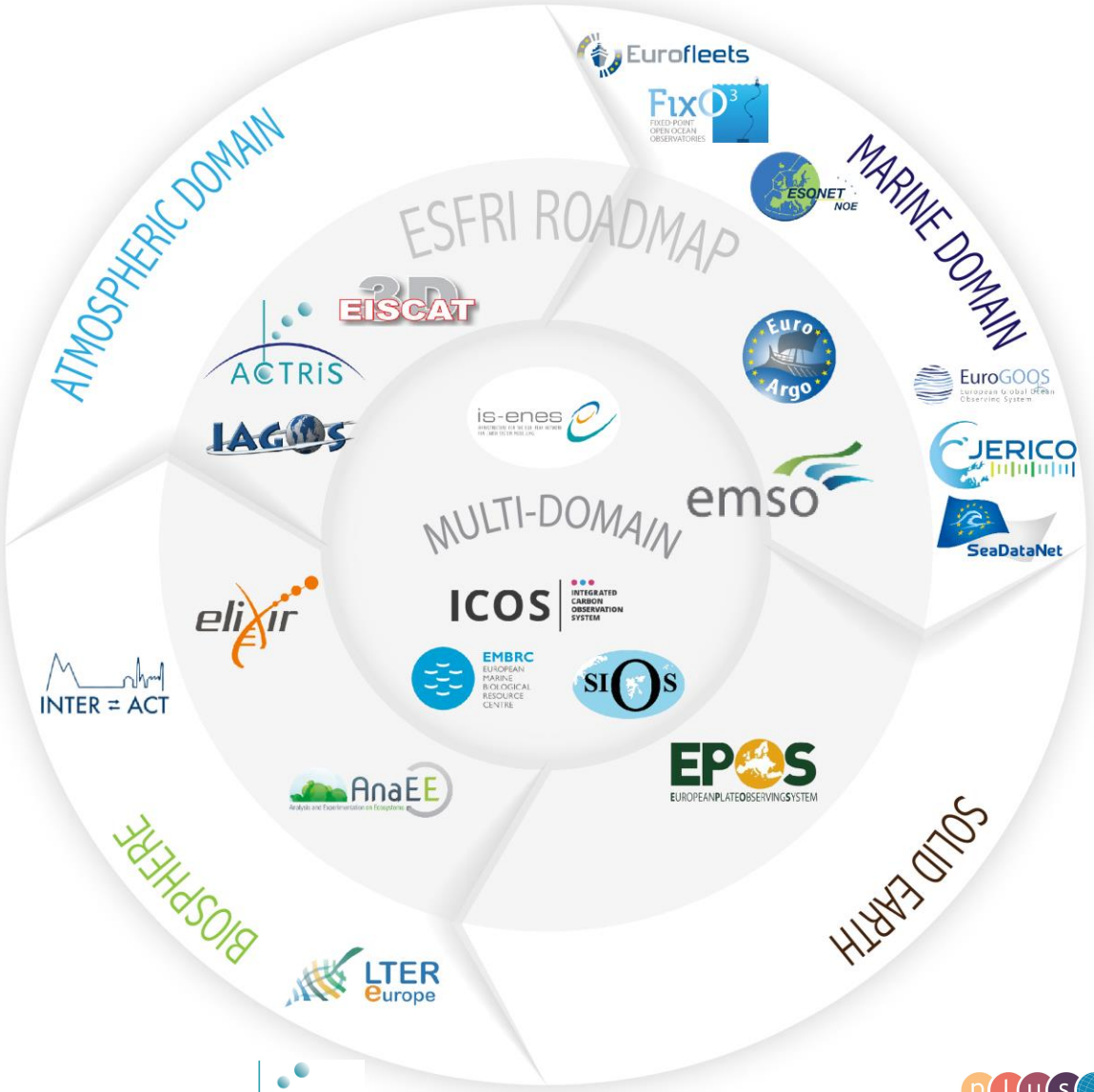
ICOS INTEGRATED CARBON OBSERVATION SYSTEM

ACTRiS

plus ENVRI

ENVRIplus members

- 20 RIs from 4 domains
- Many commonalities, but also many differences!
- Looking for common strategies & ”best practices” for data management – including PID & citation!



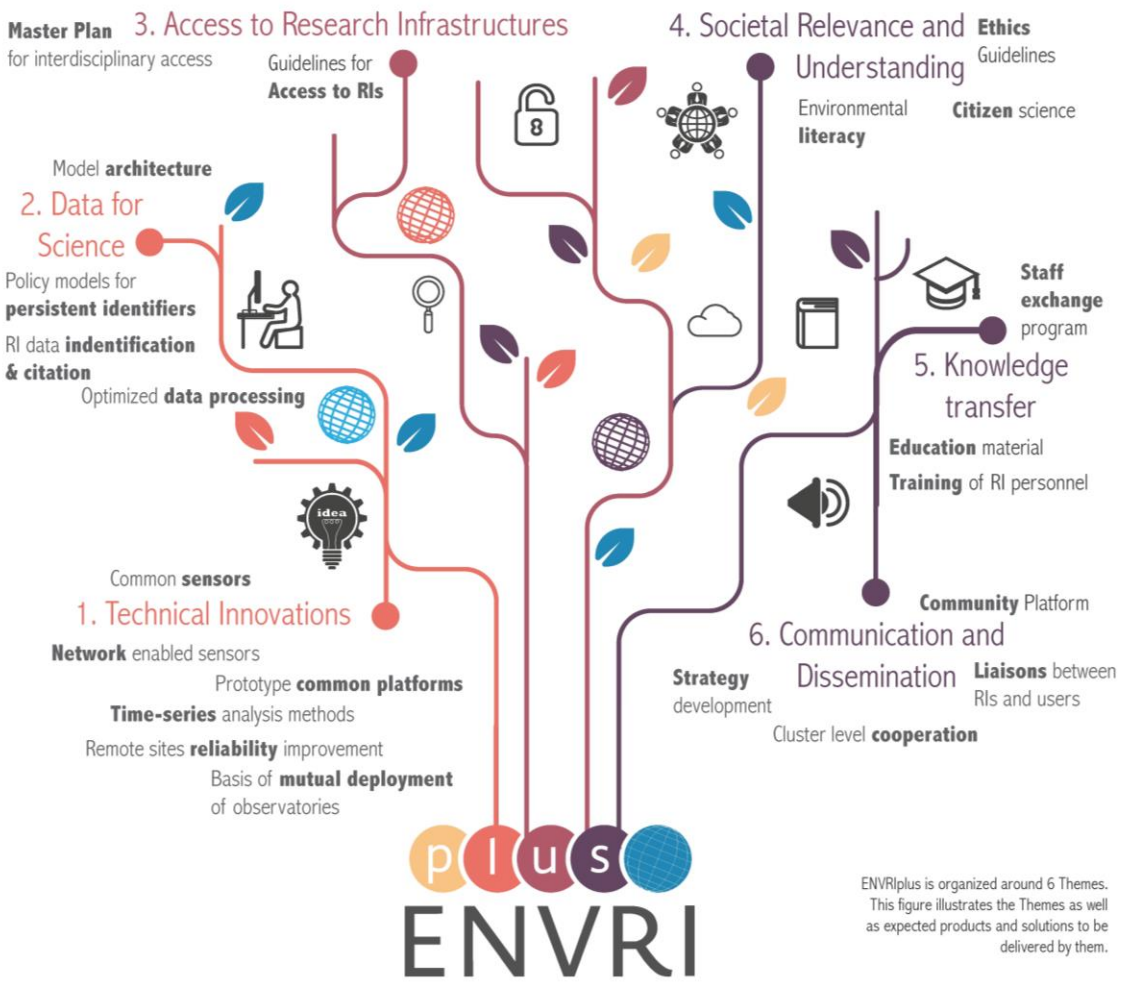
ICOS INTEGRATED CARBON OBSERVATION SYSTEM

ACTRIS

plus ENVRI

ENVRIplus themes

- 6 themes, 19 work packages
- Theme 2: “Data for science”
- WP5: Reference model guided RI design
- **WP6: “Data identification & citation”**
- WP7: Data processing and analysis
- WP8: Data curation and cataloguing
- WP9: Service validation and deployment



ICOS INTEGRATED CARBON OBSERVATION SYSTEM

ACTRiS

plus ENVRI

WP6: Inter RI data identification and citation services

- **RI requirements & needs survey**
- **Technology review**

Deliverable 5.1 (May 2016)

- **Tasks...**

http://hdl.handle.net/11346/839B

  - design common policy models for persistent identifier use
  - define services to be developed...
  - discuss with publishers & other actors...
  - use cases: develop “PID-based” data management for selected RIs...
- **Implementation case studies**
  - **Dynamic data citation (RDA recommendation implementation)**
  - Identification as support for provenance
  - **Data collections: usage accounting & credit assignment**

Who creates collections, and why?

- In principle, anyone can create a collection!
- “In the dark all cats are gray” – but all RIs are not the same (day or night)!
- EPOS, ENES, ICOS... create, curate and manage their own data
  - Collections are a way to organize large volumes of data products in different ways
- Others, like ACTRIS, rely on data contributions from “external” partners, like (small) research groups or individual scientists
  - Annual datasets may need to be built up from 10-100 small contributions

Scientific credit...

- Funding agencies are pushing towards increasingly more open data policies
- Publically-funded projects required to re-distribute their data, also for commercial use
- Reliable statistics of citations or (re-)use – is paramount for accumulating “scientific credit” – needed for grants & career advances
- Correct quantification of data usage is key
- “Easy” (easier) for individual data sets, but more tricky for collections!

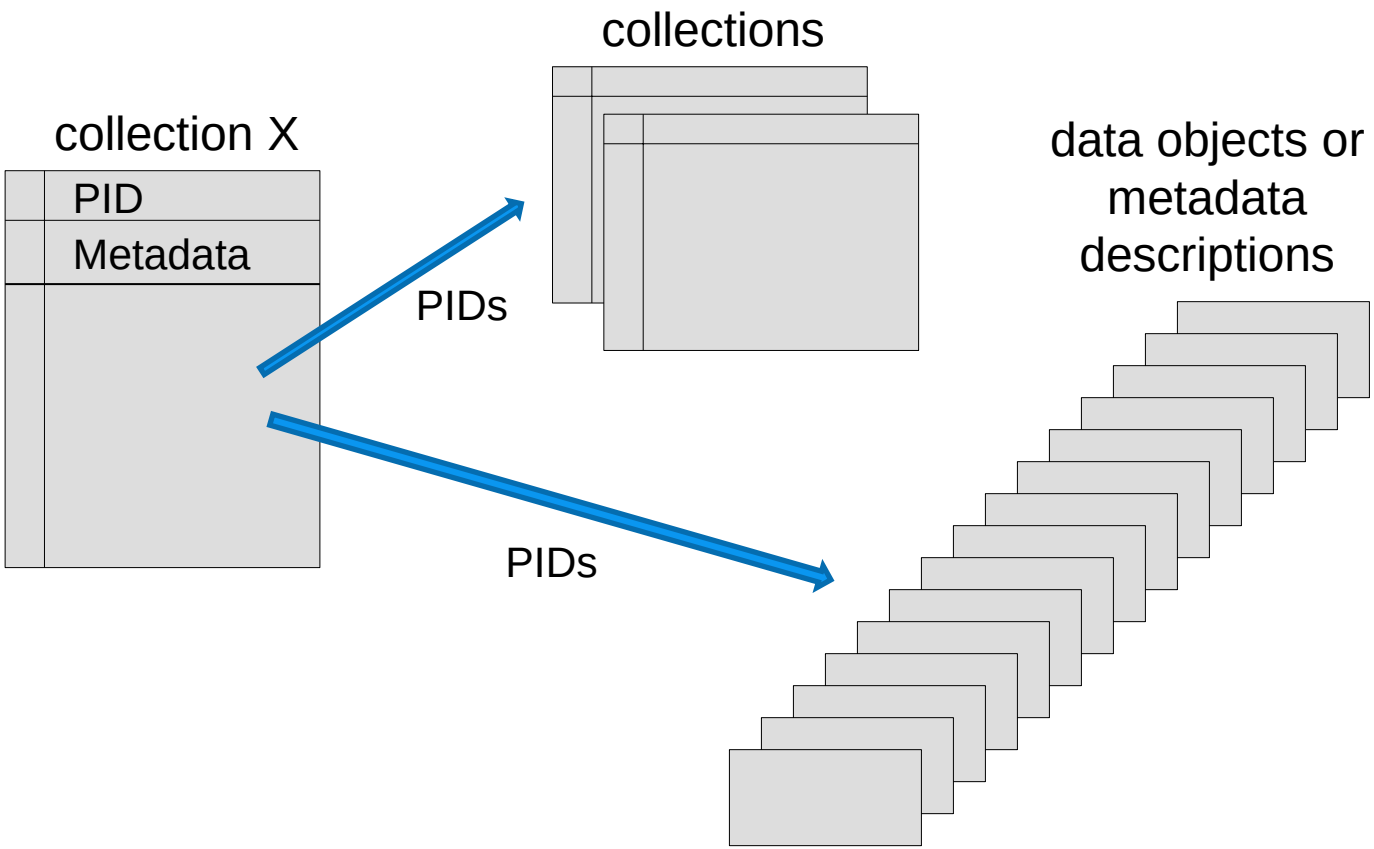
Many items, many ”authors”?!

- “Primary” identification of all data archived in data centers, with suitable & homogeneous granularity
- Collections of course need to include references to these primary identifiers of all data sets included
- But how easy is it to extract “authorship” of the member items when starting from a collection’s PID/DOI?
- Special case: collections resulting from queries to (dynamic) data sources (PID’d or even DOI’d) – could be a collection of subsets of many different items!

Extracting & assigning credit

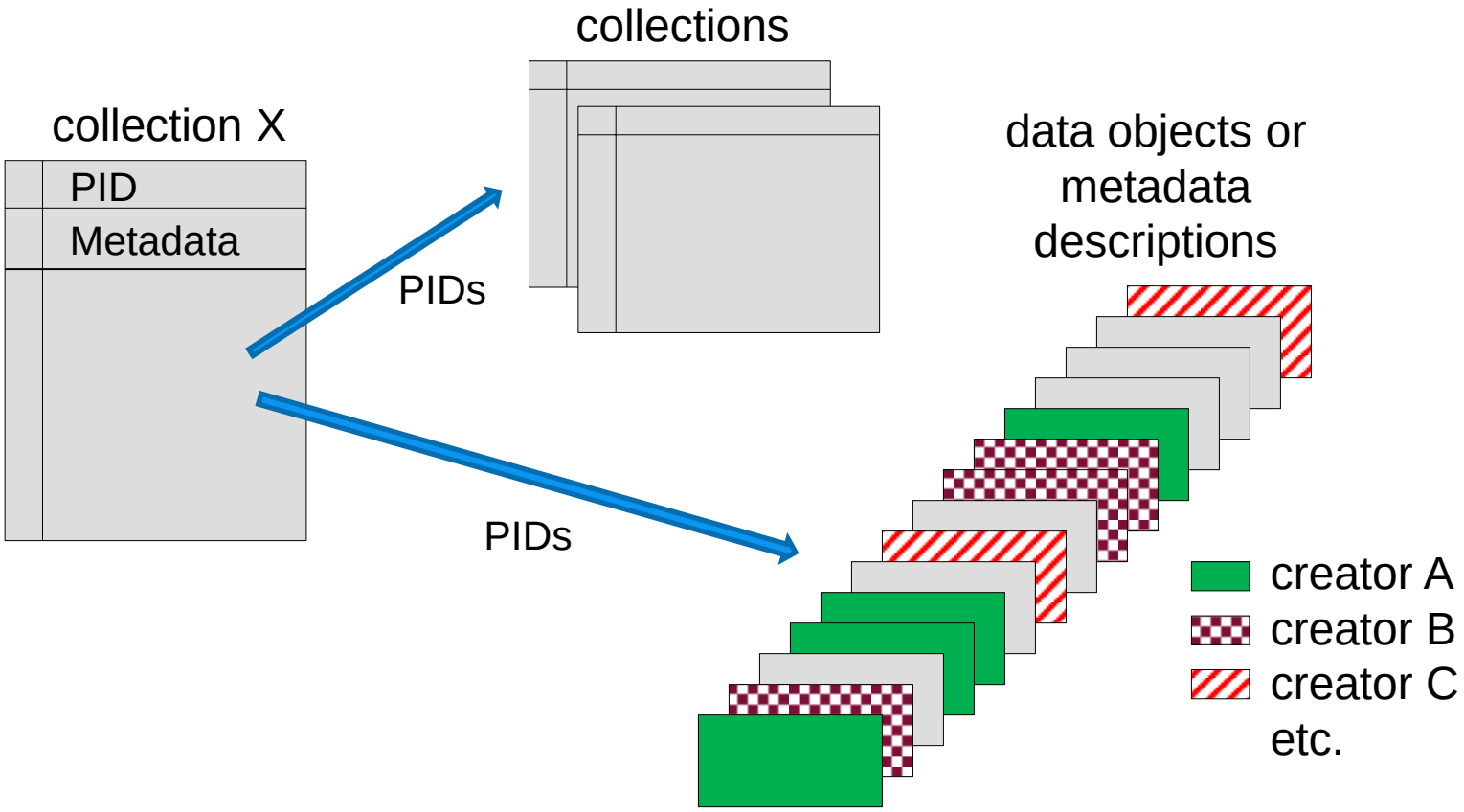
- Any services handling the collections (serving data, collecting citation statistics etc.) must be able to create “complete” citation information, including the owners of each member, on request
- Usage metrics-compiling agents should use such feature to resolve a collection citation in literature into “credit” of all member item authors/owners!
- Tricky: “proportional” credit assignment (see next slides)

Collections & their members



Adapted from a presentation by Peter Wittenburg...

Collections & their members



Adapted from a presentation by Peter Wittenburg...

Questions

- Where should the “author” information on the collection members best be stored?
  - In the collection’s PID record?
    - in its own Description field?
    - in member item-specific Description fields?
  - In the members’ own individual PID records?
  - In the members’ own metadata (at landing page level)?
  - In a “collection catalog”
    - maintained by the PID registry of the collection
    - maintained by RIs, libraries, ...

More questions...

- How can membership information best be retrieved?
  - Which (other) collections include member M(i)?
  - What information about M(i) is stored there?
- Can membership information - M(i) is a member of C(i) - be added to M(i)'s metadata somehow?
  - E.g. via annotations to its PID registry record?
  - Impossible if collection creator isn't also creator of members?
- If the collection C(i) is cited
  - How can "bibliometric credit" be assigned to its members?
  - If author A contributed more than one member to C(i), then what?
  - How to point to subsets of C(i)? (By using members' own PIDs, or in some other way?)

Thank you for your attention!

Get in touch: margareta.hellstrom@nateko.lu.se
markus.fiebig@nilu.no
ENVRIplus-coordination@helsinki.fi

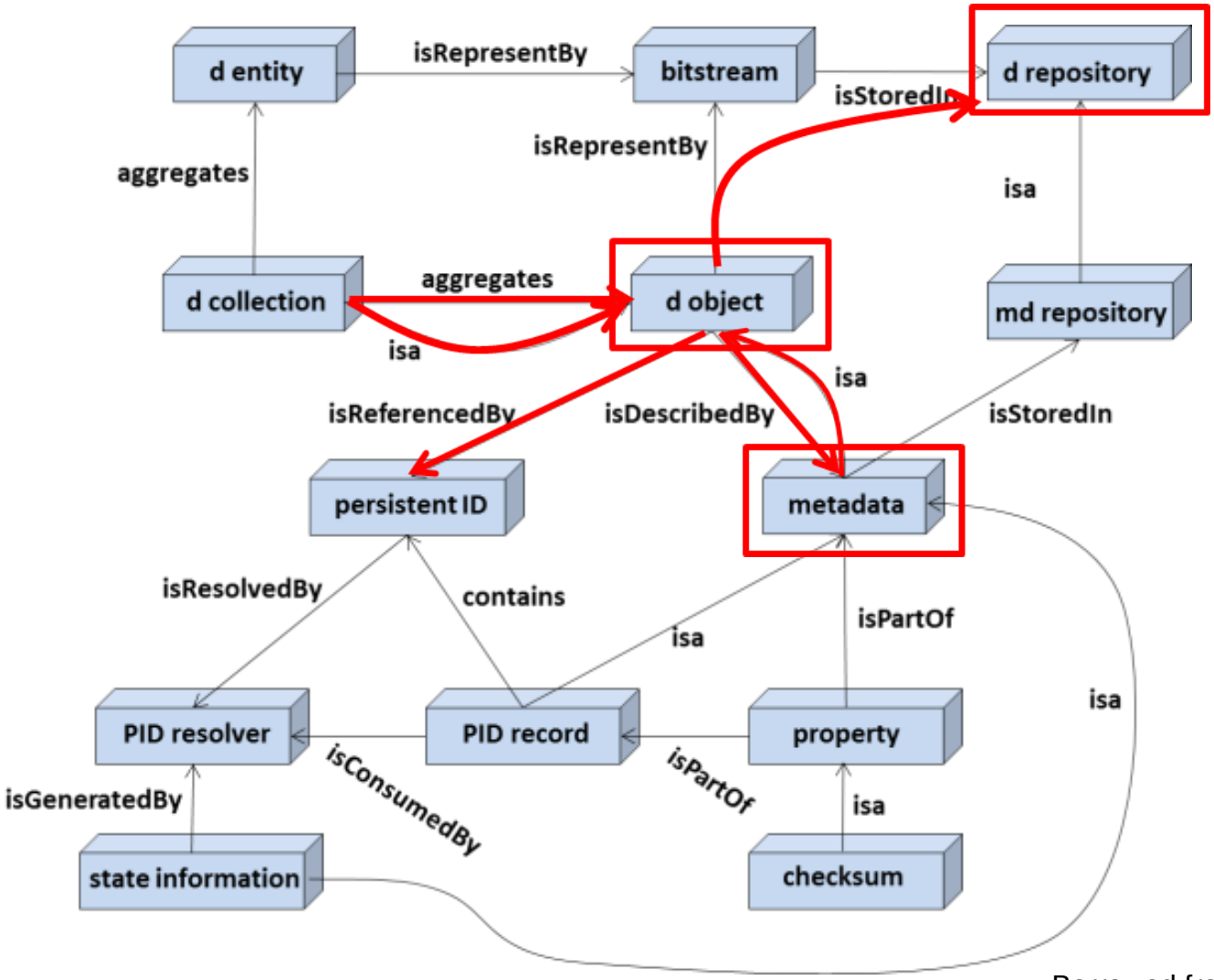
ICOS INTEGRATED CARBON OBSERVATION SYSTEM

ACTRiS

ENVRI plus

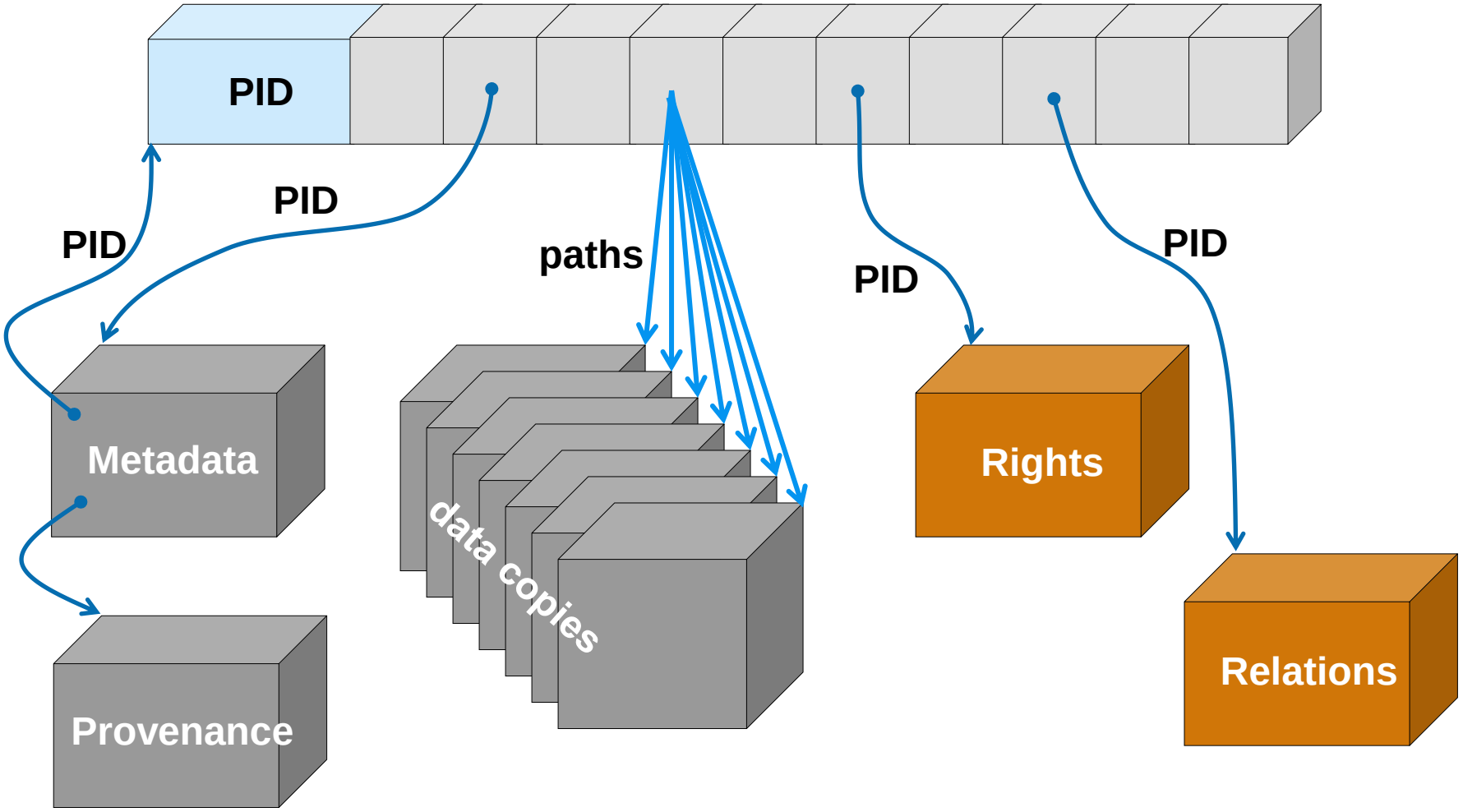
Some additional slides
(for the discussion)

The DFT Model of Data Organisation



Borrowed from a presentation by Peter Wittenburg...

PID-centric Data Object model...



Borrowed from a presentation by Peter Wittenburg...

Ideas from Research Data Collections WG

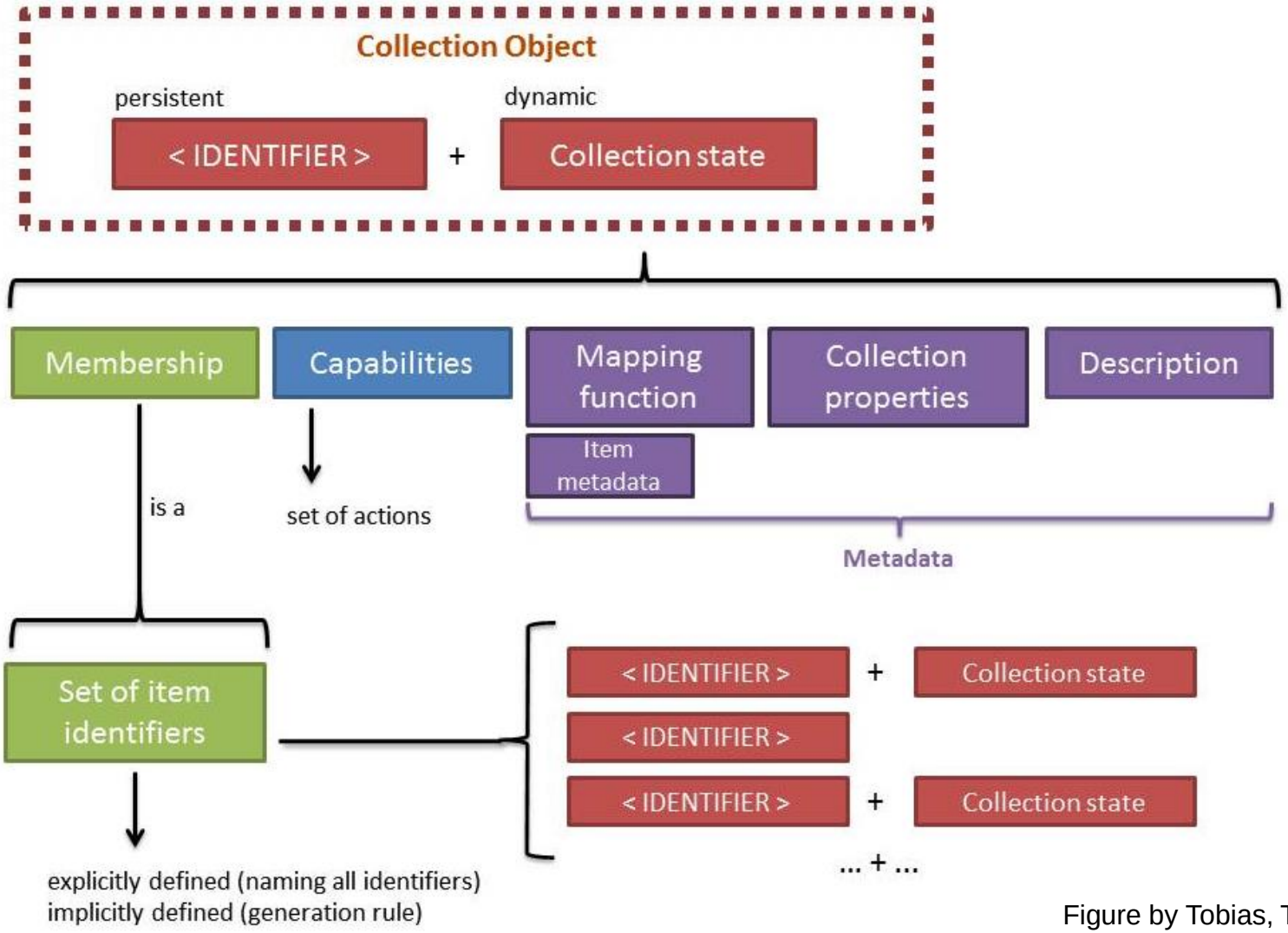


Figure by Tobias, Tom, Frederik, Ulrich, Bridget