PIDs for Data and Metadata

Metadata Workshop, 22 Jan 2014

Tobias Weigel
Deutsches Klimarechenzentrum (DKRZ)

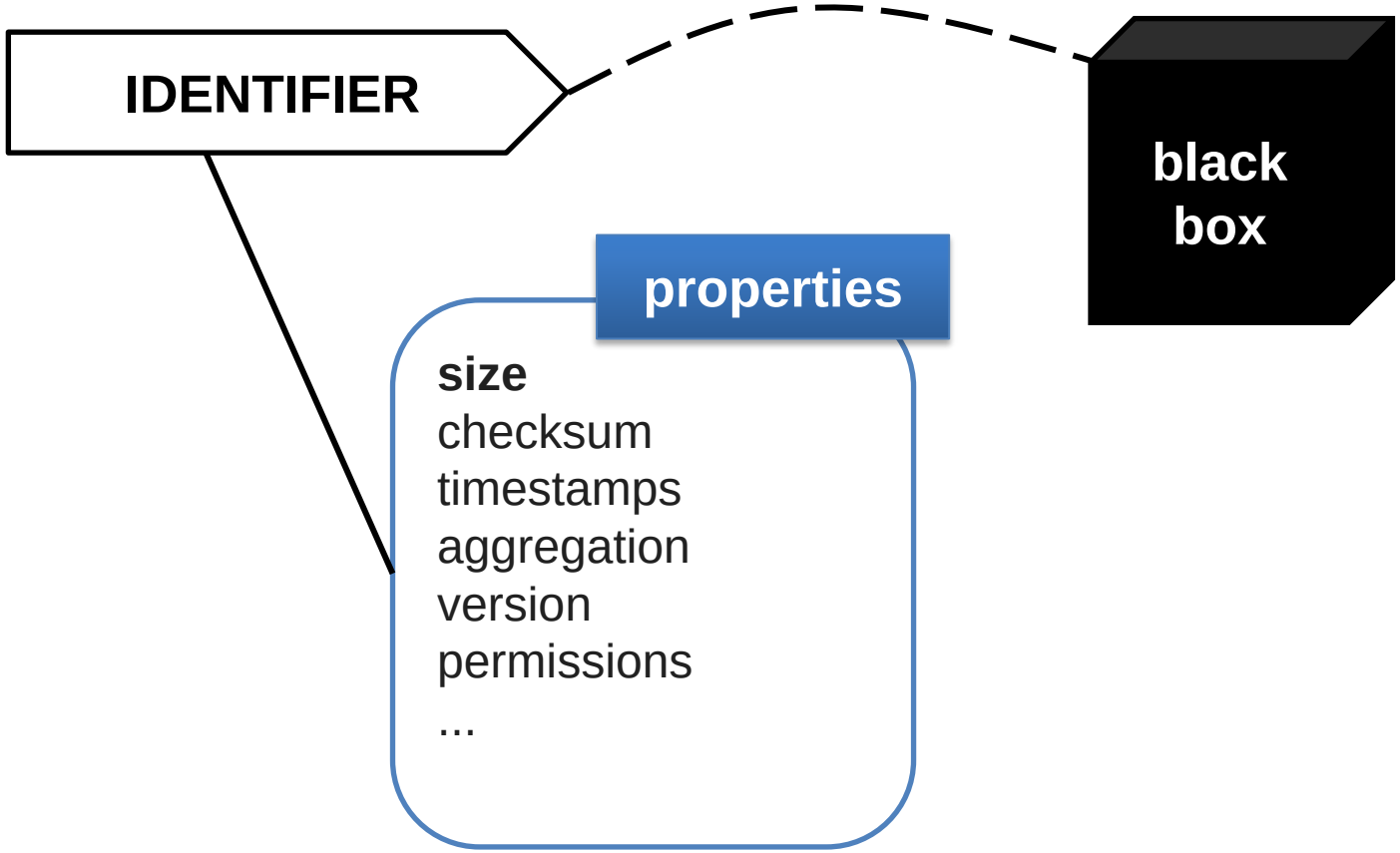
Persistent Identifiers come in various formats

- 10876/abc123
- 10.1594/WDCC/CMIP5.NCCNMpc
- ark:/13030/tf5p30086k
- http://purl.org/dc/elements/1.1/
- urn:lsid:ubio.org:namebank:11815

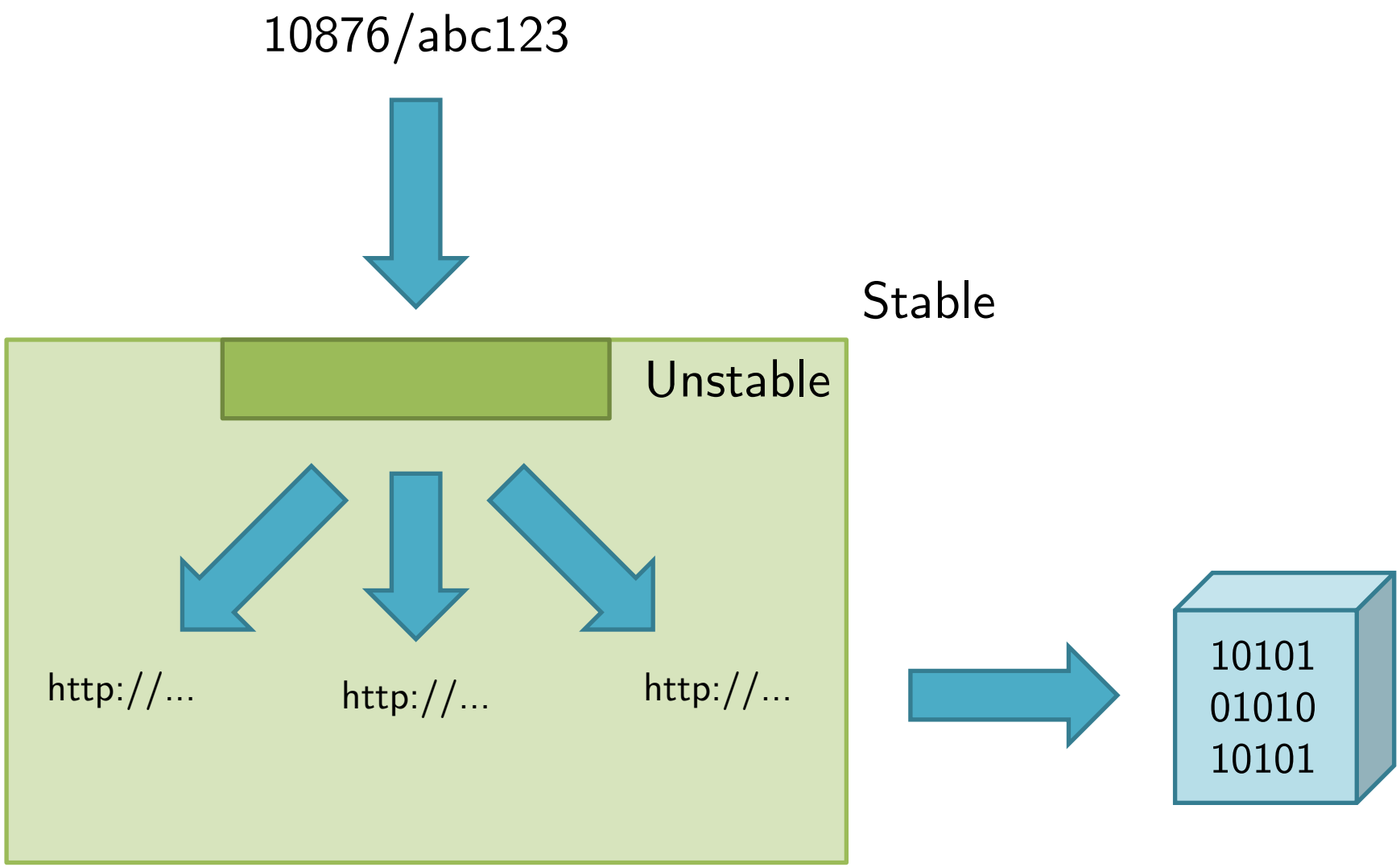
There are many PID systems / infrastructures

- Handle System
- Digital Object Identifier (DOI)
- Archival Resource Key (ARK)
- Persistent URL (PURL)
- Life Science Identifier (LSID)
- Uniform Resource Name (URN)
- ...

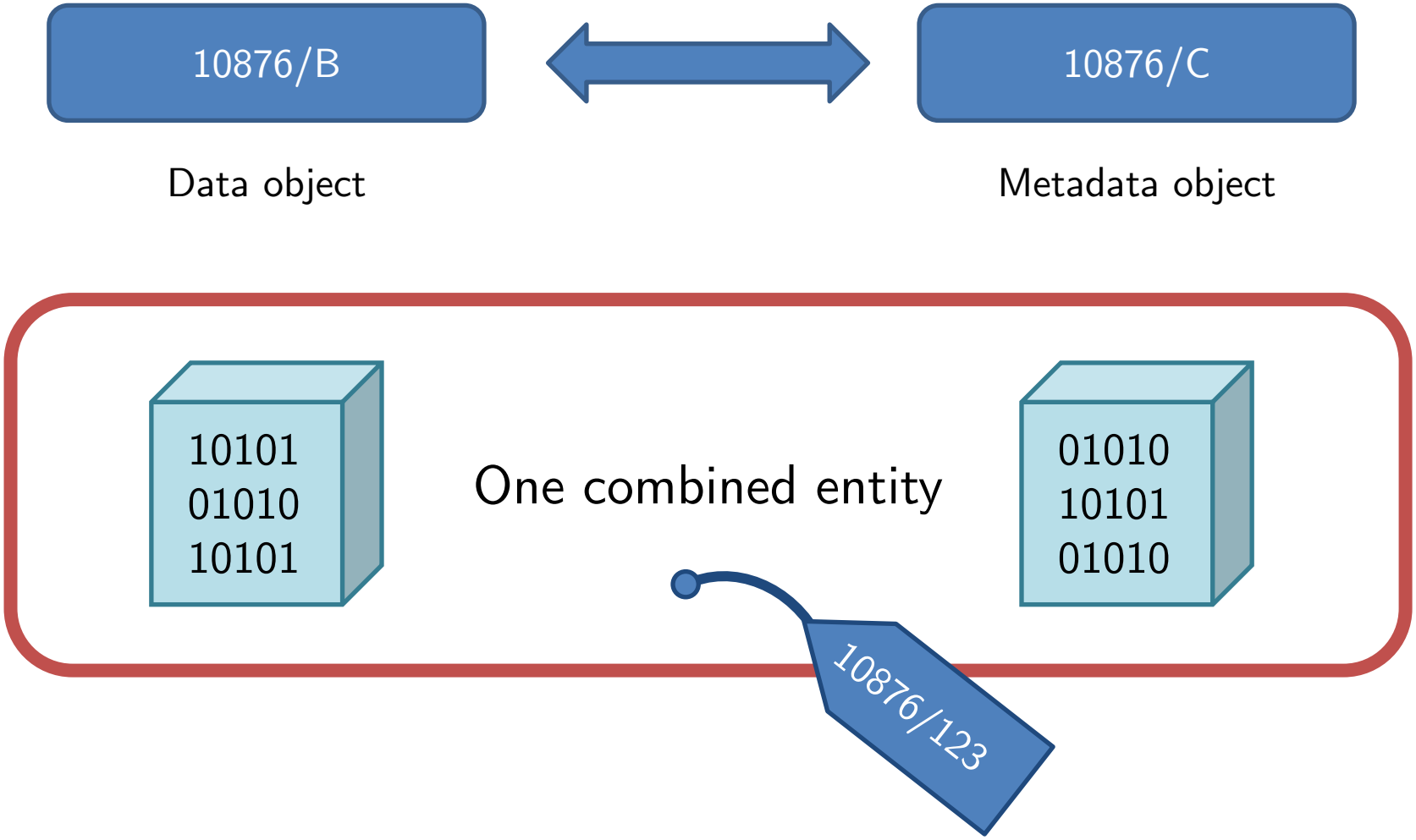
What are PIDs?



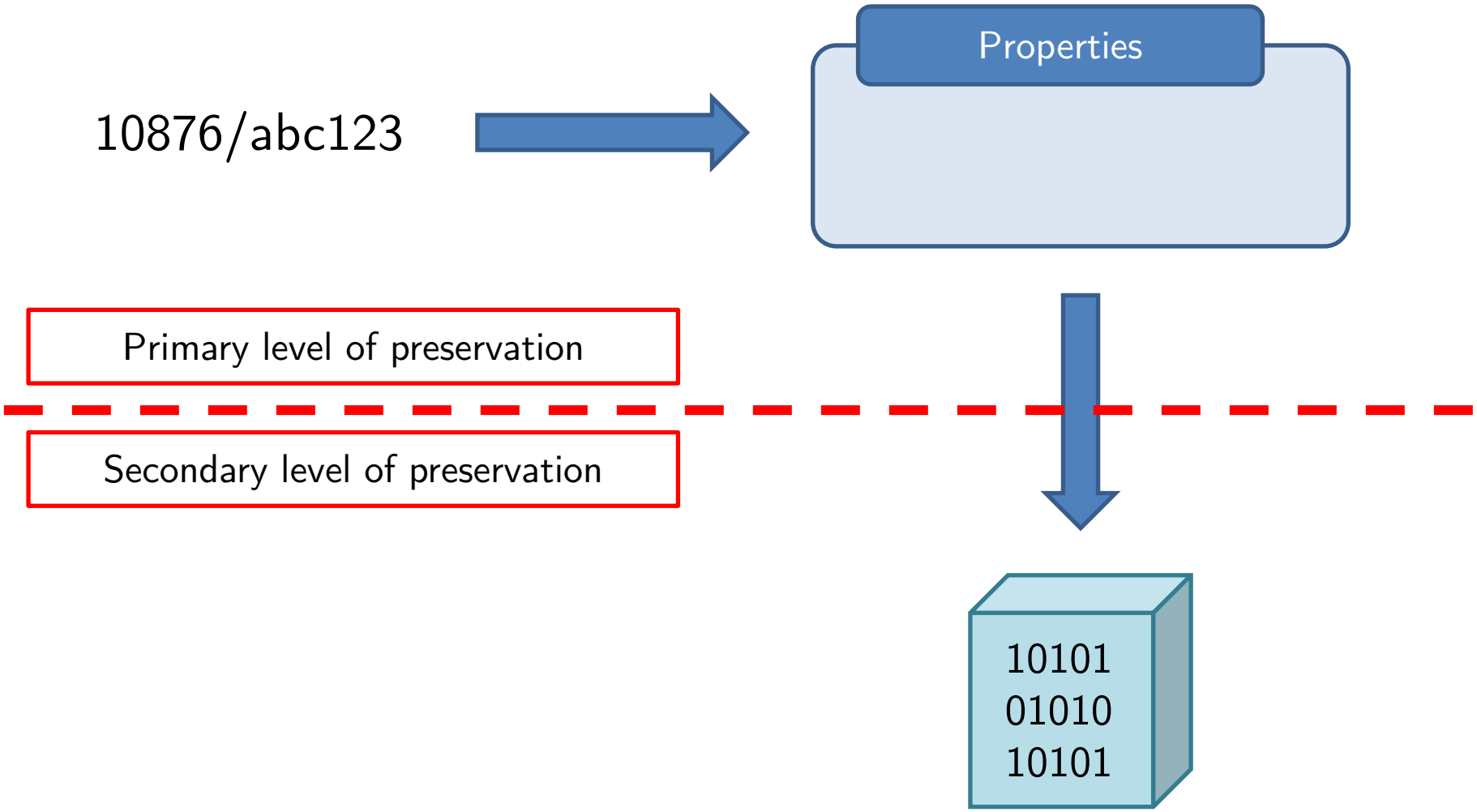
PIDs establish a redirection layer



Data and metadata can be linked together

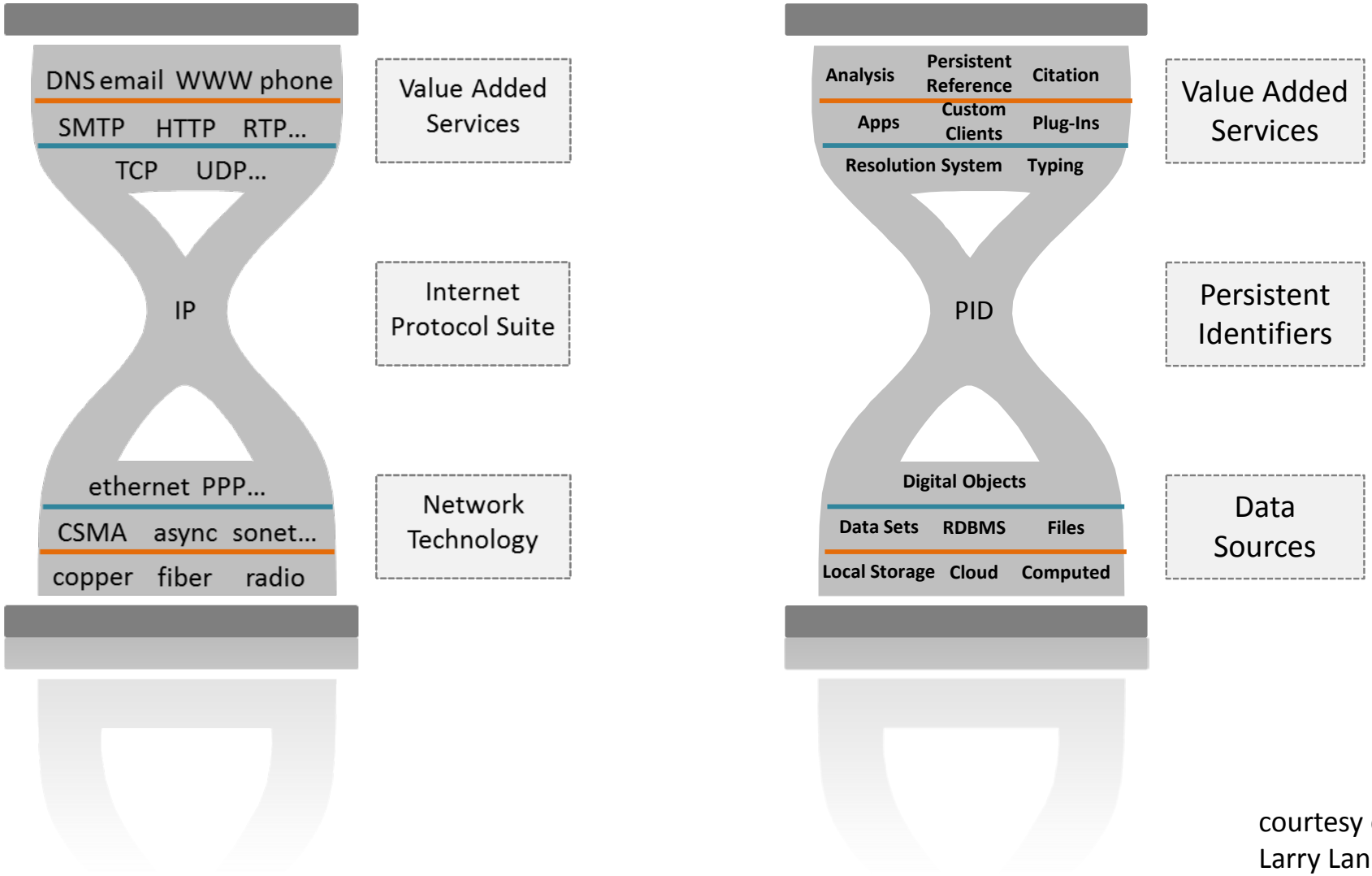


Levels of preservation



Weigel et al. 2013

PID layer stack



courtesy of Larry Lannom

Further PID usage scenarios

RDA WG on PID Information Types

Research Data Sharing without barriers

http://rd-alliance.org

- WG co-chairs: Tobias Weigel (DKRZ), Timothy DiLauro (JHU)

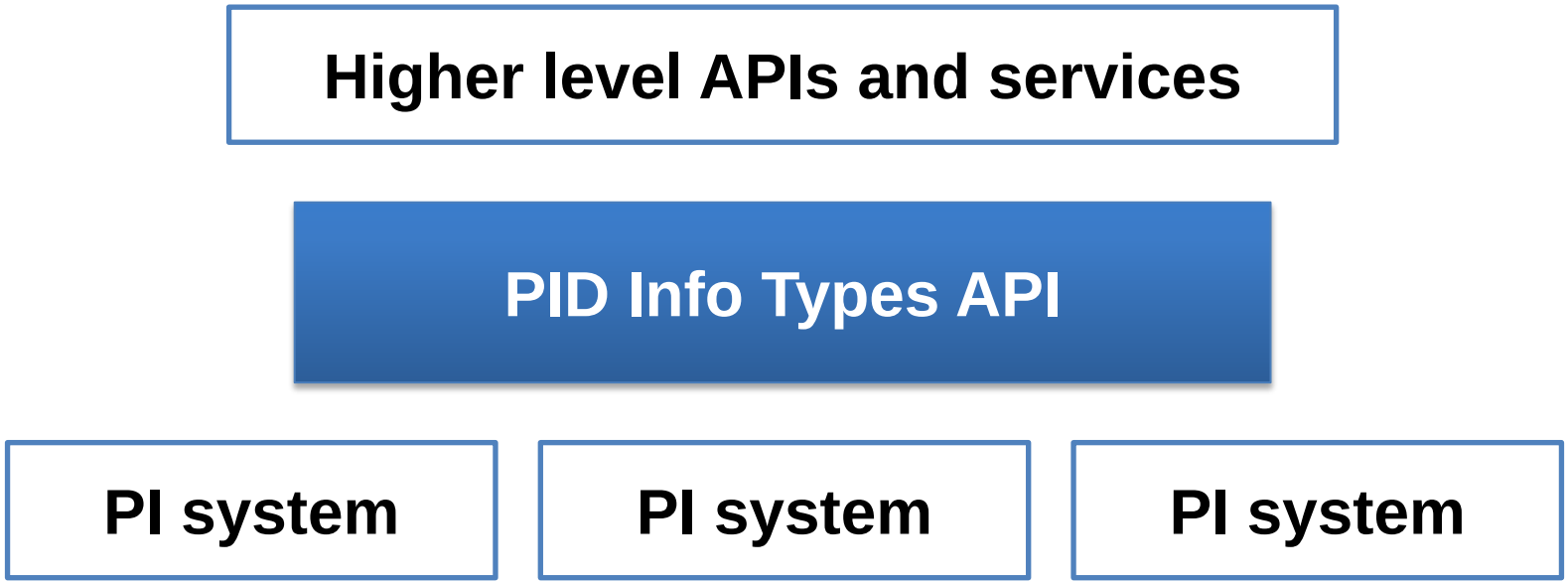
- Founded at 1st RDA Plenary, Göteborg, 03/2013
- Will be terminated at 4th plenary, Amsterdam, 09/2014

PIT WG tasks

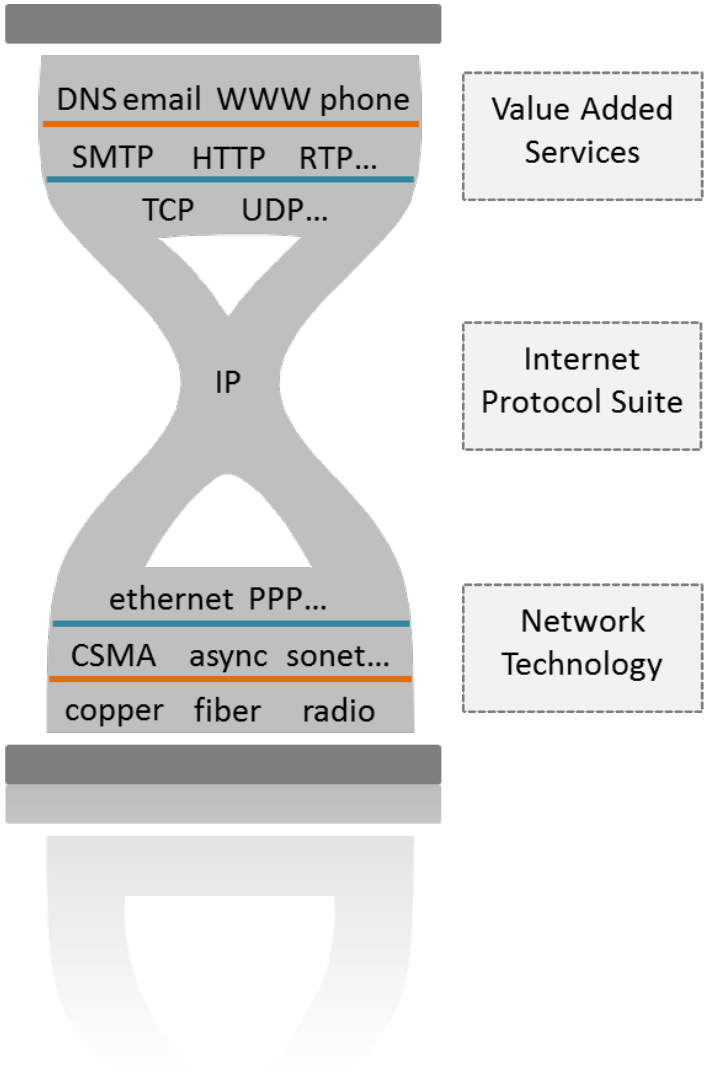
- Develop use cases (done)
- Develop API specification (active)
  - https://docs.google.com/a/org.dkrz.de/document/d/1P-_BOzQ_kZz6UUiqqNEeEJmHnmaXiTSxWRveCheow60/edit
- Sort out the greater architecture (active)
- Build demonstrator/prototype

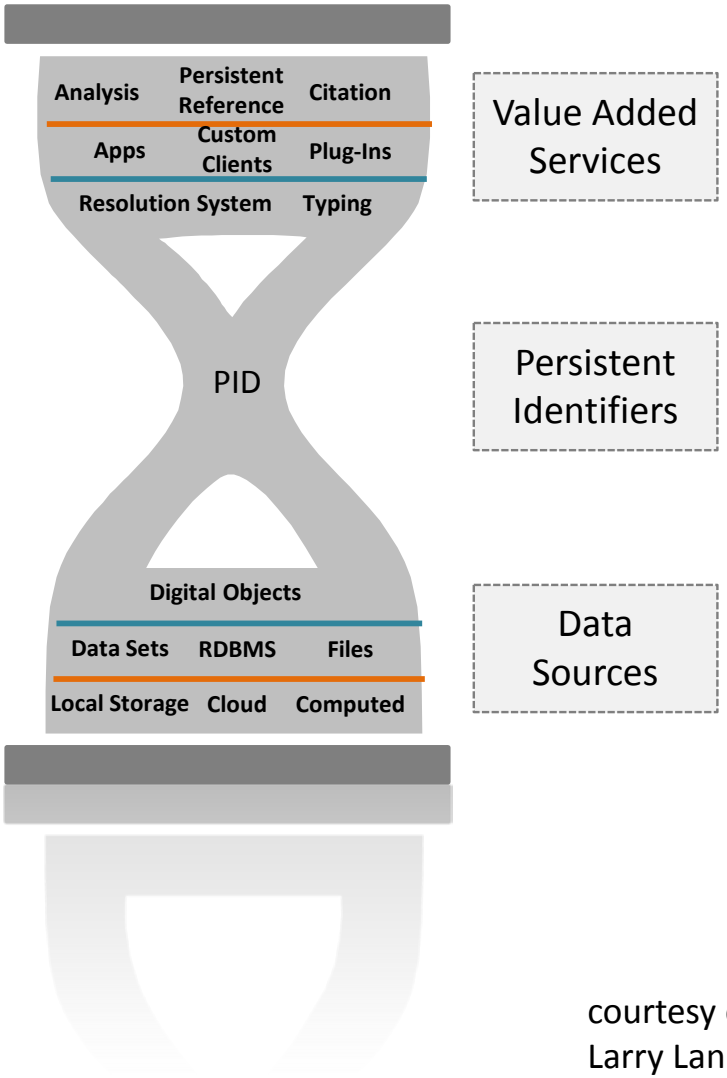
Interoperability through technical interfaces

- The PIT API – one approach for interoperability



PID layer stack

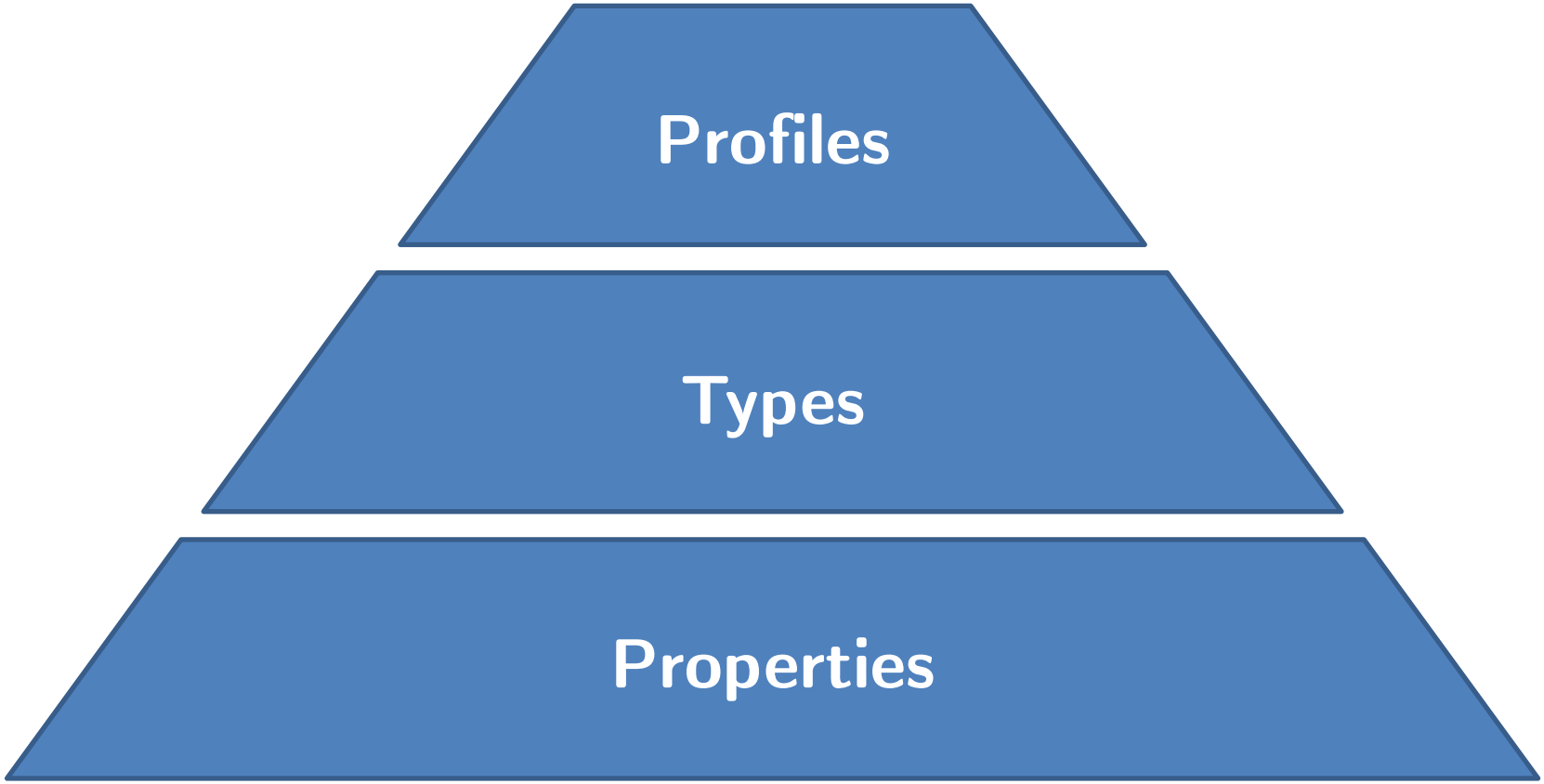




courtesy of Larry Lannom

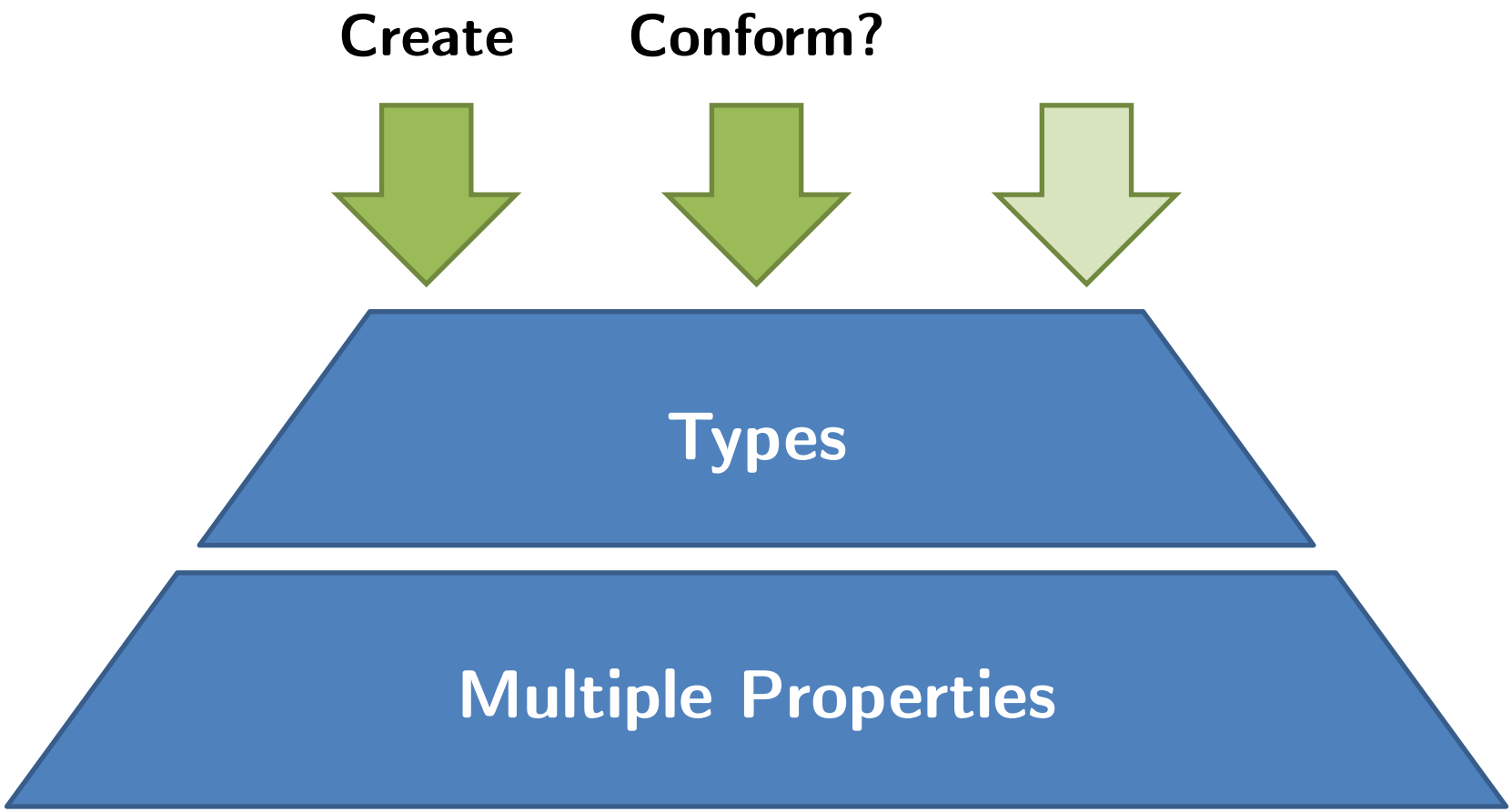
Organize properties to harmonize their use.

Tim, John Erickson, Tobias
Discussion will continue!



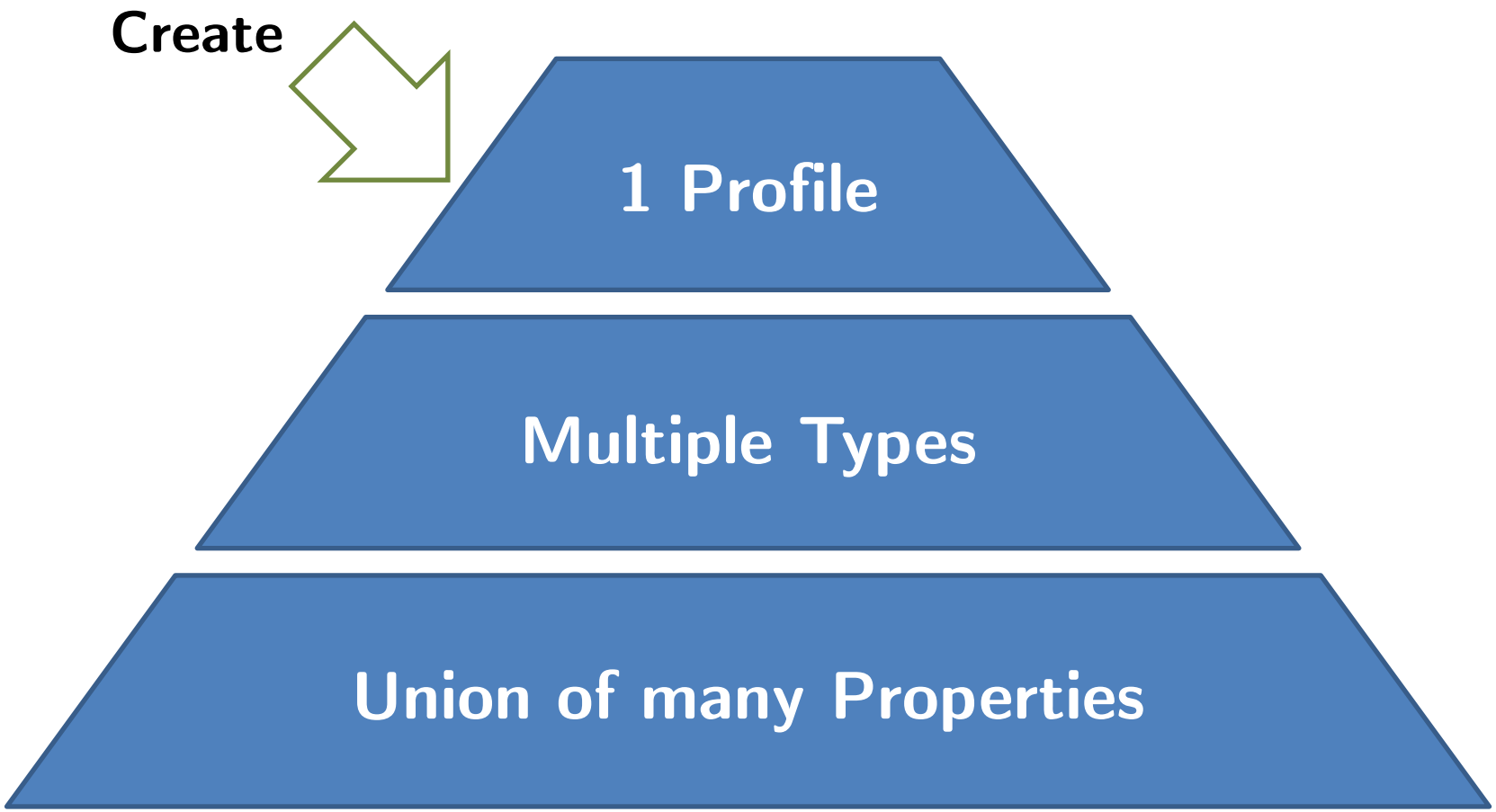
Core services defined on types.

Tim, John, Tobias ...



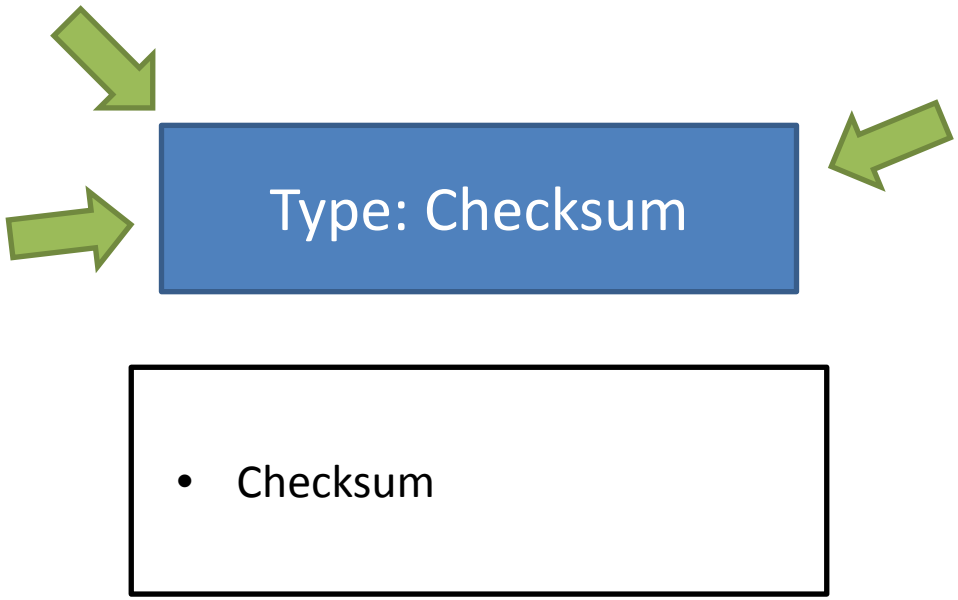
Create locally according to a profile.

Tim, John, Tobias ...



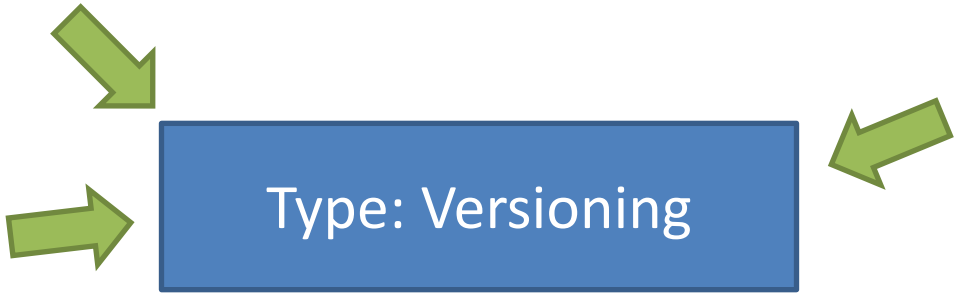
Checksums are a simple example.

- Type „checksum“
  - (or: verifiable object)



Versioning is an example for Collections.

- Type „versioning“ (for some particular use case!)
  - (or: versioning-enabled object)



- Version number
- PID of previous object
- PID of following object

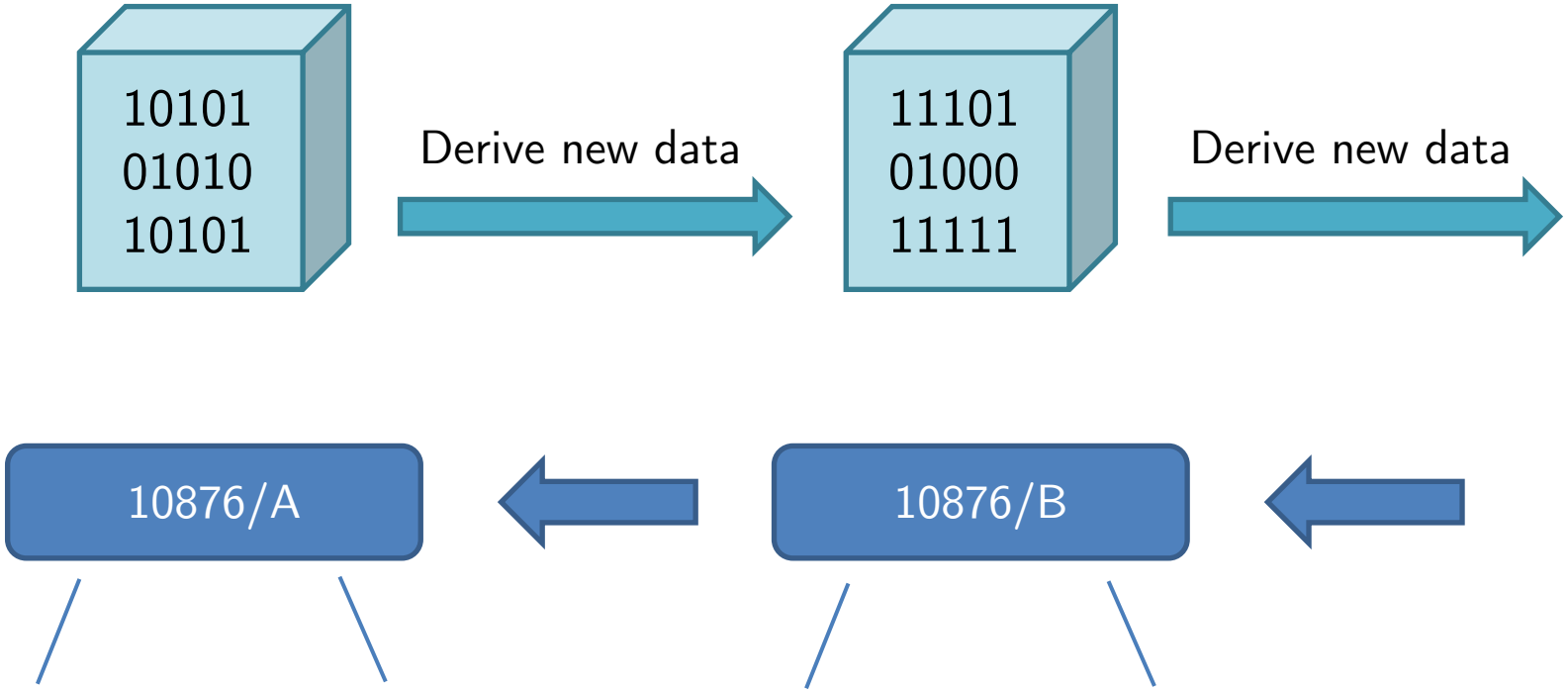
Profiles prevent proliferation of types.

Profile: Checksum+Versioning

Conform?

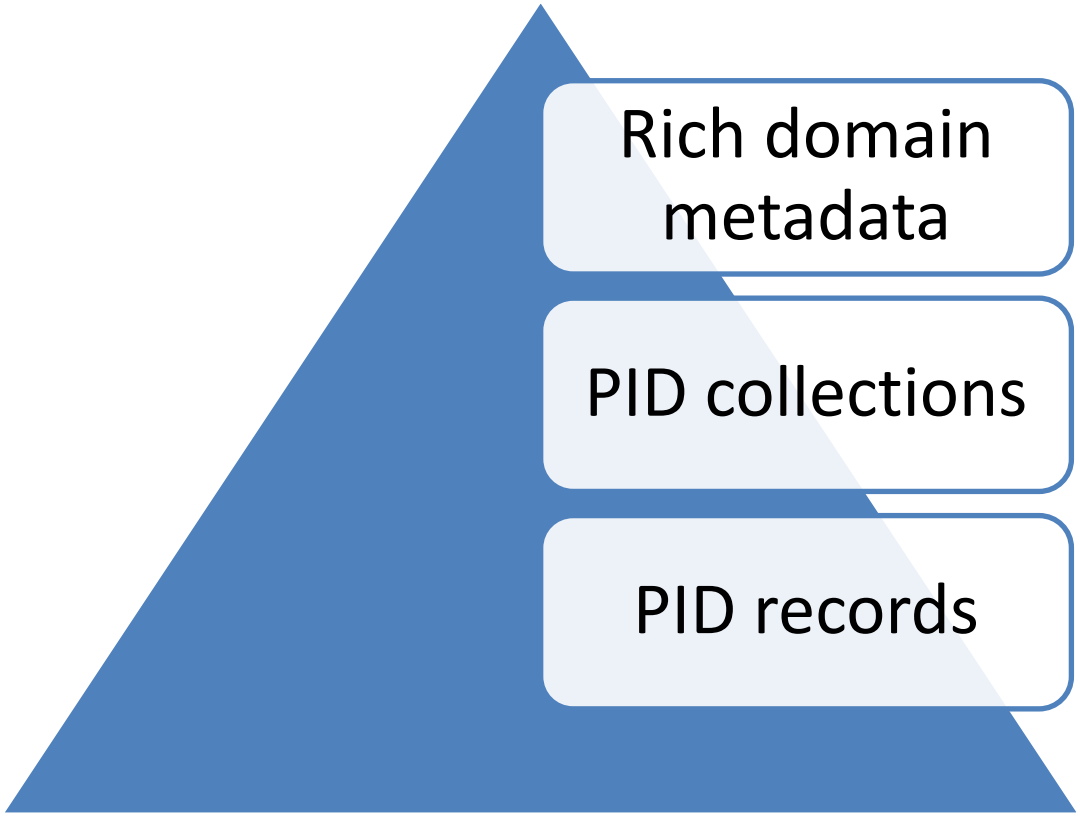
- Checksum
- Version number
- PID of previous object
- PID of following object

PIDs may be used to trace provenance



... plus more detailed metadata records/objects in a higher layer.

PID layer stack – with collections



- But: Avoid duplication of information!

What are PID collections?

- Abstract Data Types based on PID Records
  - Actionable!

- Prototypic implementation: Linked Lists, Arrays, Hash maps
- Keep structural information within PID record
- Provide interoperability between PID systems

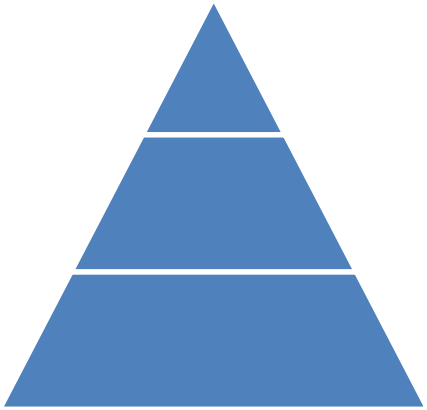
- https://github.com/TobiasWeigel/lapis
- More details: Weigel et al. 2014

More PID use cases

- Data-Metadata binding
- Versioning
- Provenance tracing
- Generic composites
- Hierarchies
- Custom data citation
- ...

Early workflow usage?

- Trade-off:
  - We do not know which objects are persistent
  - Assign PIDs as early as possible to track their provenance
  - But: must also limit the number of PIDs for objects that are eventually deleted



PIDs in ESGF / CMIP6

- Unify the various identifiers in use
- Assign identifiers early – when they enter the federation
- Provide versioning, tracking
- Easy migration/uplifting to DOIs

- Need to have good high-quality processes to ensure the **P**ersistency of PIDs

End-user perspective – a long-term vision?

- Provide PID graph discovery features on DOI landing pages
- Visibility into the context aggregated through a dataset‘s lifetime
- May not hold rich metadata for every step, but may be decodable for a human (forensics)

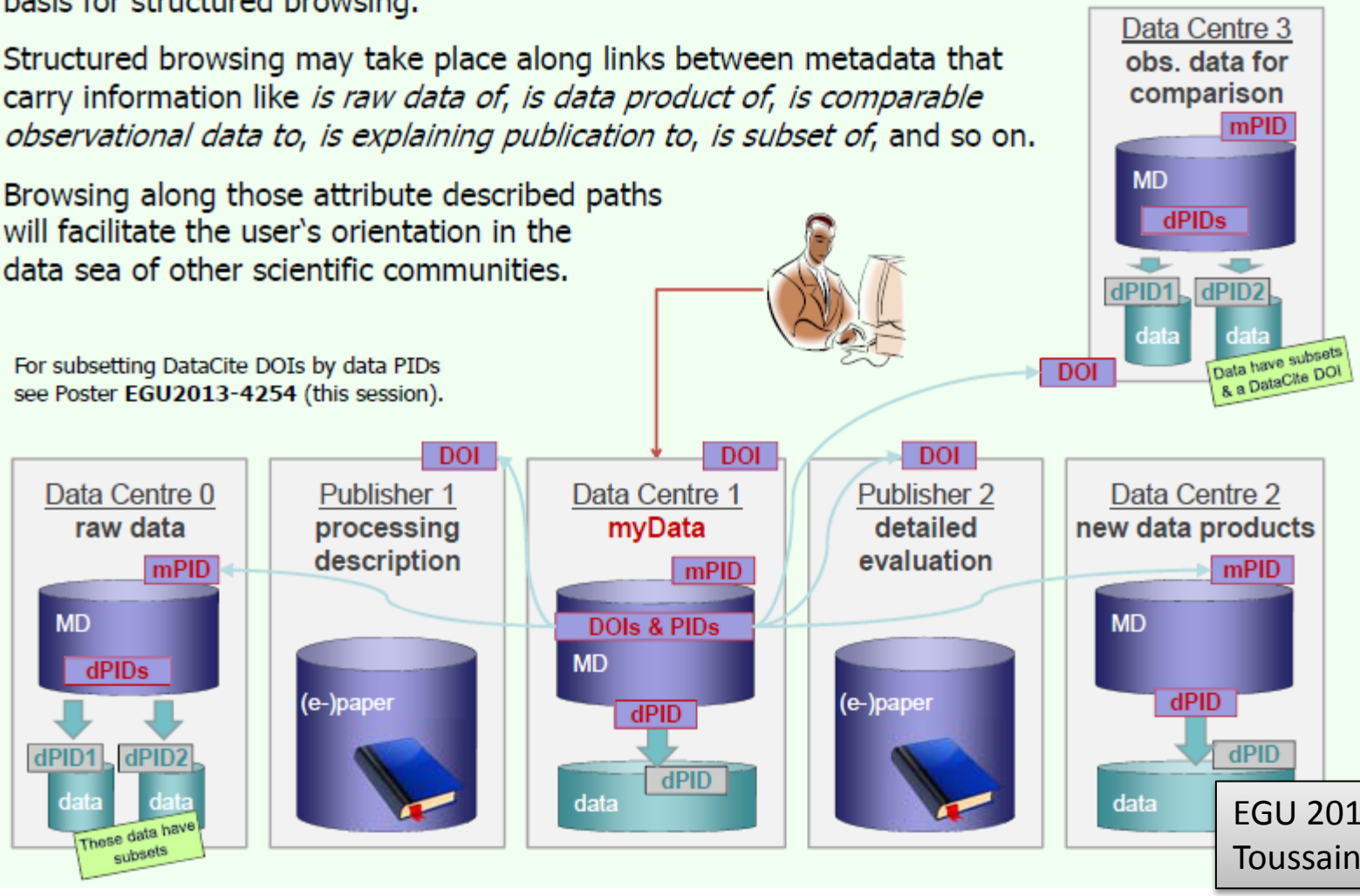
Surfing metadata via Persistent Identifiers

For cross community data usage mutual search functionality is needed. However, to know what to search, the first step is browsing metadata, to get an idea what data are availabe. As browsing in an unordered list is inefficient, data links by PIDs can be basis for structured browsing.

Structured browsing may take place along links between metadata that carry information like *is raw data of, is data product of, is comparable observational data to, is explaining publication to, is subset of*, and so on.

Browsing along those attribute described paths will facilitate the user's orientation in the data sea of other scientific communities.

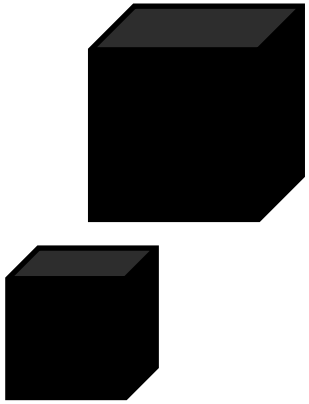
For subsetting DataCite DOIs by data PIDs see Poster **EGU2013-4254** (this session).



EGU 2013, Toussaint et al.

The end.

- Thank you for your attention.



- https://www.rd-alliance.org/working-groups/pid-information-types-wg.html

- Weigel, Lautenschlager, Toussaint, Kindermann (2013): A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, Vol. 12, pp 10-22. http://dx.doi.org/10.2481/dsj.12-036
- Weigel, Kindermann, Lautenschlager (2014): Actionable Persistent Identifier Collections. Data Science Journal, Vol. 12, pp. 191-206. http://dx.doi.org/10.2481/dsj.12-058
- Toussaint, Stockhause, Weigel, Höck, Lautenschlager (2013): Application of Handles in the European Data Project EUDAT. EGU General Assembly, EGU 2013-5475

Persistent Entity / PID

- Surrogate for an all-encompassing PID definition
- Persistent Entity ADT
- Two operations:
  - (key-)metadata resolution
  - resource resolution
- MD resolution remains even if data object is gone
  - primary level of preservation