Emerging domain agnostic functionalities on the handle-centered networks

Kei Kurakawa
National Institute of Informatics
Takayuki Sekiya
The University of Tokyo
Yasumasa Baba
The Institute of Statistical Mathematics

International Workshop on Sharing, Citation and Publication of Scientific Data across Disciplines
Joint Support-Center for Data science Research (DS), ROIS
NIPR / NINJAL, Tachikawa, Tokyo, Japan, 5-7 December 2017.

Overview

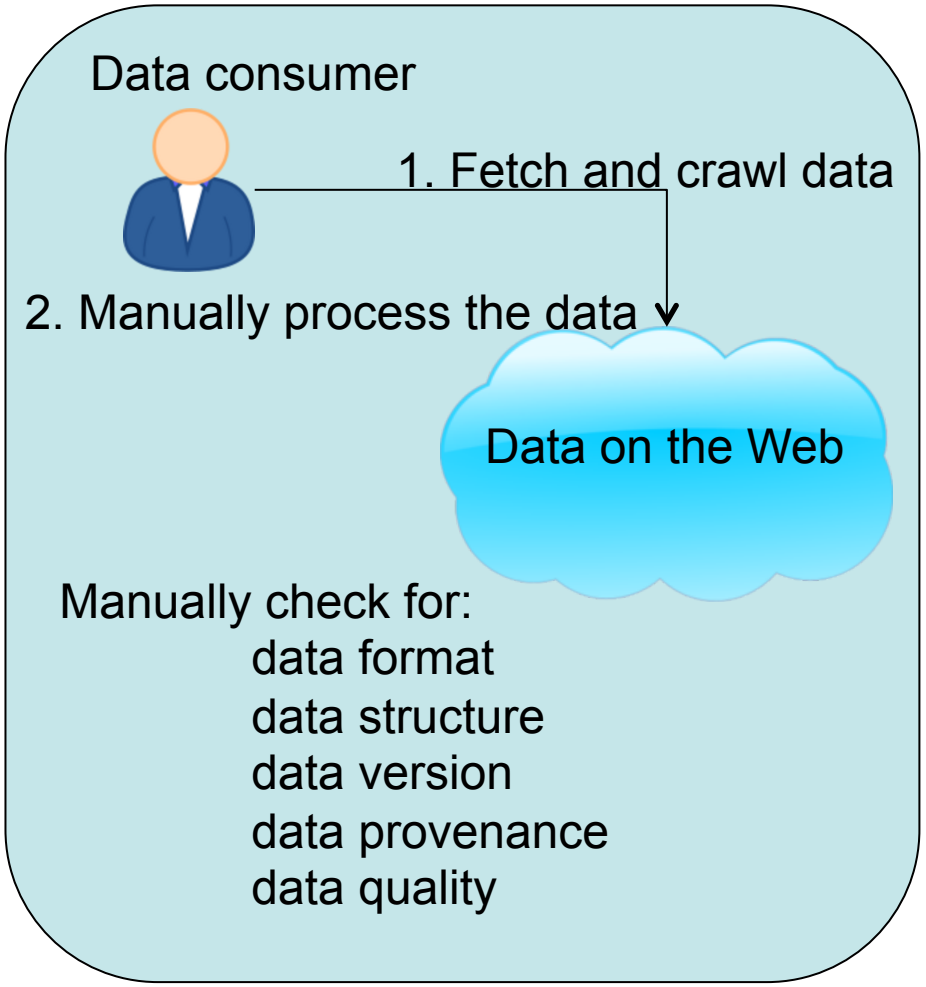
- Research data sharing
- Domain-independent automatic data processing environment on the PID centric information model for very large collections of distributed scientific data
- Kernel Information
- Handle-centered networks on Kernel Information metadata layer
- Future directions
- Summary

Research data sharing mind from Open Access, Open Data , Open Science

- Disciplinary historical events
  - Meteorology and geoscience
    - The first International Polar Year (IPY)(1882)
    - The first International Geophysics Year (IGY) (1957)
  - Biology
    - “Bermuda Rules” (1996)
- Interdisciplinary events
  - Budapest Open Access Initiative (2002)
  - Berlin Declaration (2003)
  - G8 Open Data Charter (2013)
- The movement reached at the slogan “research data sharing without barriers” of RDA (Research Data Alliance) among all disciplines, in order to innovate and develop societal and technological specifications for scientific data infrastructures.

Current procedure to aggregate and process the scientific data

- The procedure, which may be peculiar to each discipline, is a process of craftsmanship and too much time consuming task.
- The data consumer needs to understand the semantics of data structure in domain dependent schemes and choose ordinarily a community standard of tools on a specific computational environment to process the data.
- It seems to be difficult for outsiders of the expertise to do the same things.



A community objective

- Data on the web
  - Very large collections of scientific data, which is distributed on the web
  - PID centric information model
- Two major processes in the scientific data use
  - Data discovery
  - Automatic data processing
- To invest domain-independent automatic data processing environment on the PID centric information model for very large collections of distributed scientific data

PID centric information model and services

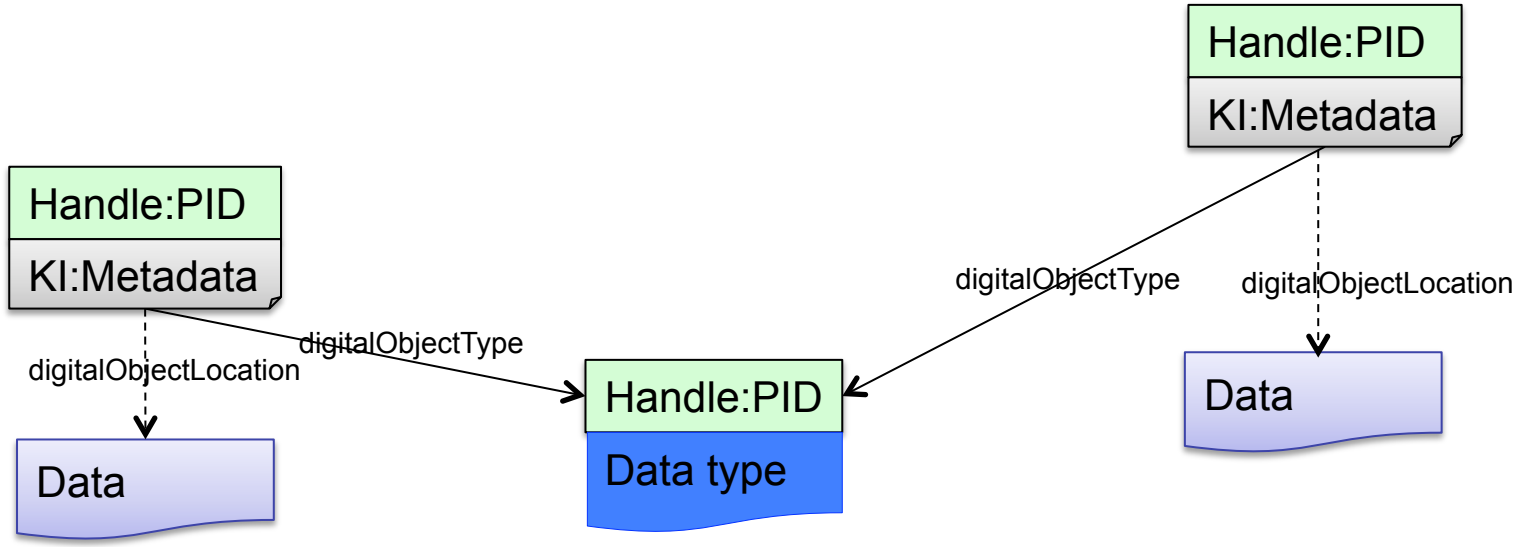
Working group outputs of the Data Fabric, Data Type Registries, PID Information Types, and PID Kernel Information

- Information elements
  - Handle : PID
  - Metadata
  - Data
  - Data type
- PID resolve service
  - Handle server
- Metadata service
  - Metadata repository
- Data services
  - Data repository
  - Data type registry

Kernel Information : Metadata

<Web Space>

Kernel Information represents a connection between Data and Data type.

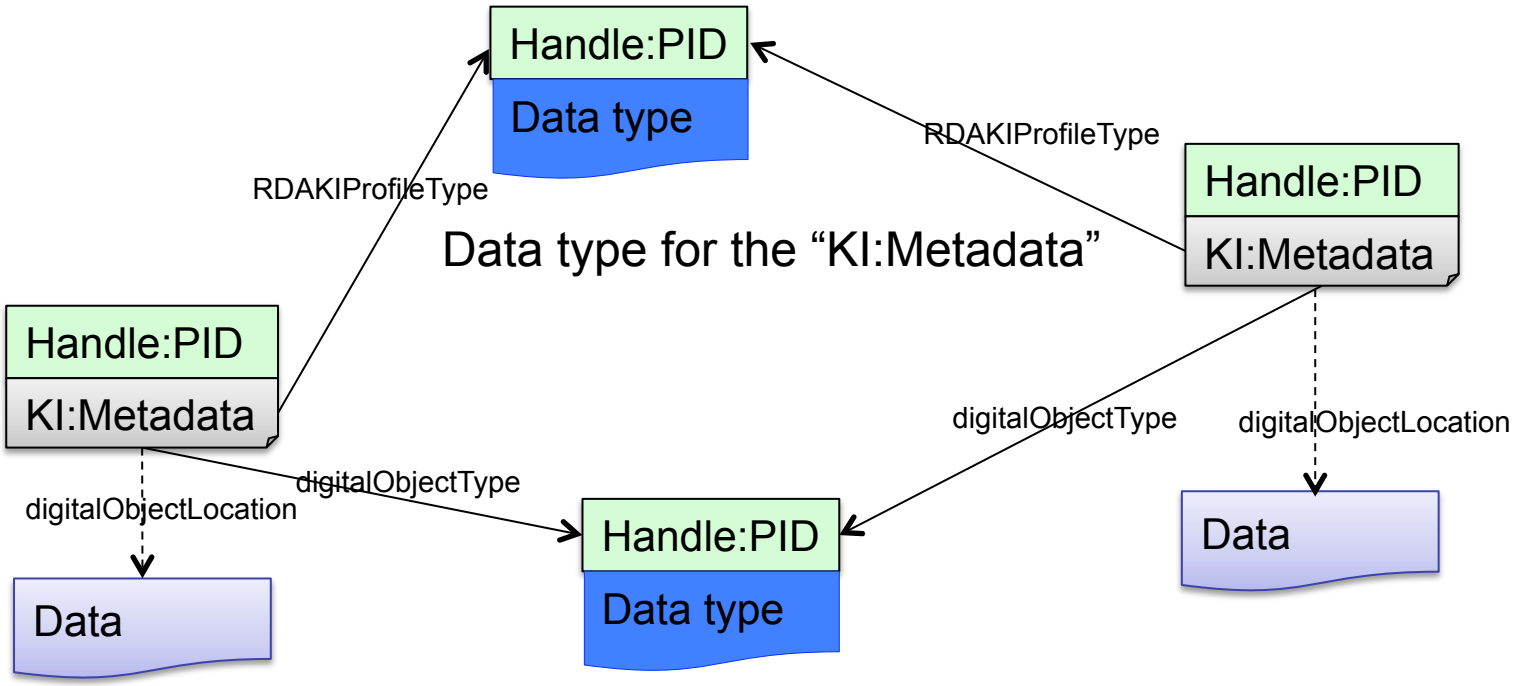


Data type for the “Data”

Kernel Information : Metadata

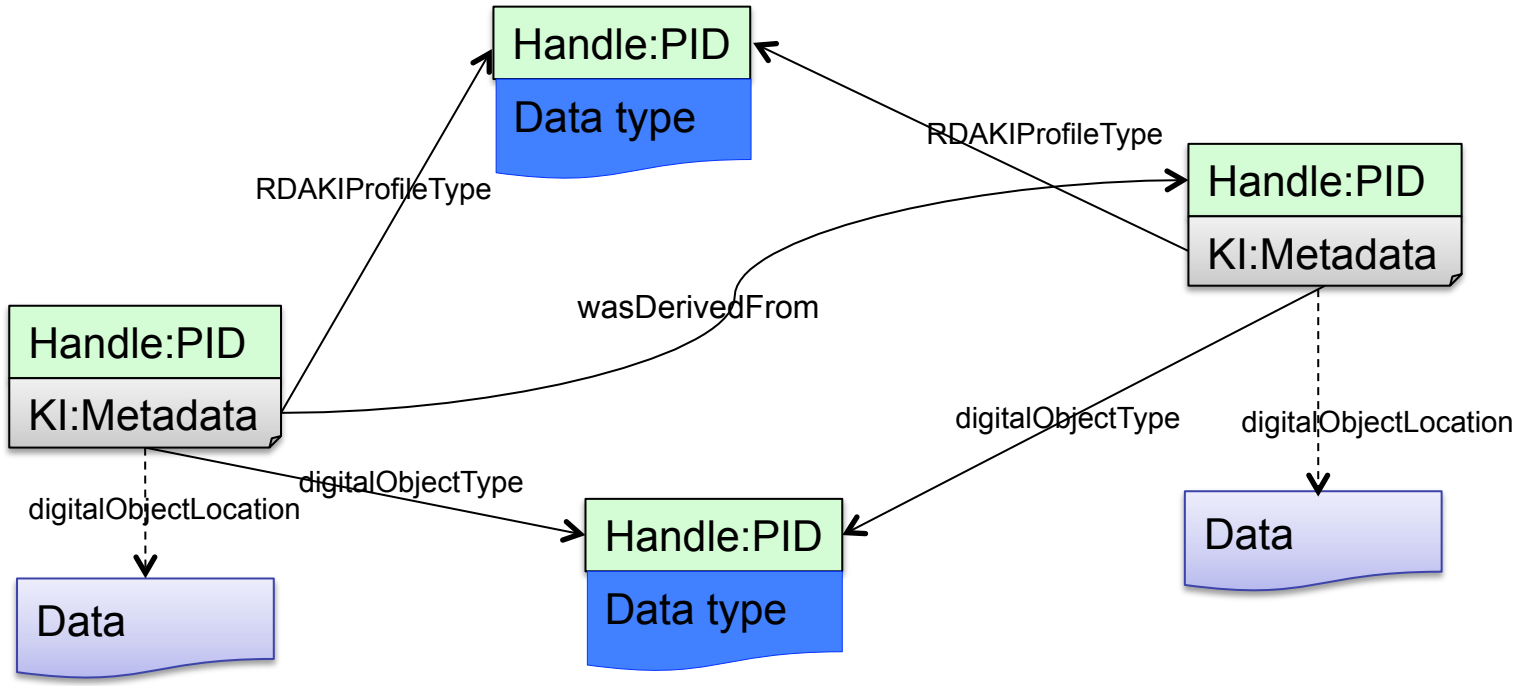
<Web Space>

KI:Metadata itself also should be data-typed.



Kernel Information : Metadata

<Web Space> KI represents structural relationships.



Kernel Information structural data relationships defined

- wasDerivedFrom
- specializationOf
- revisionOf
- primarySourceOf
- quotationOf
- alternateOf
- hadMember
- externalW3CPROVDoc

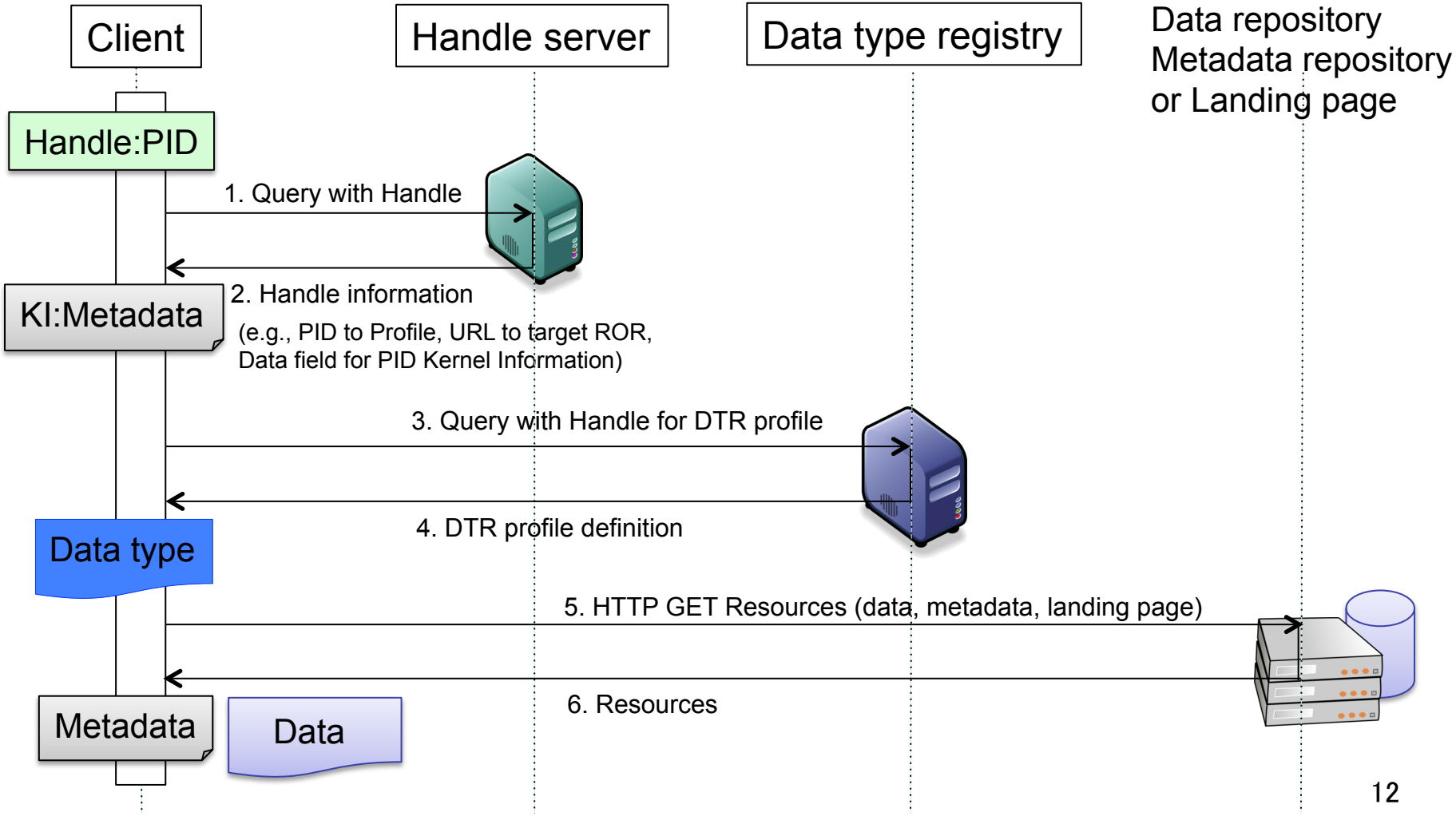
PID profile metadata 17.03.06 from the WG

Strawman PID Profile 17.03.06

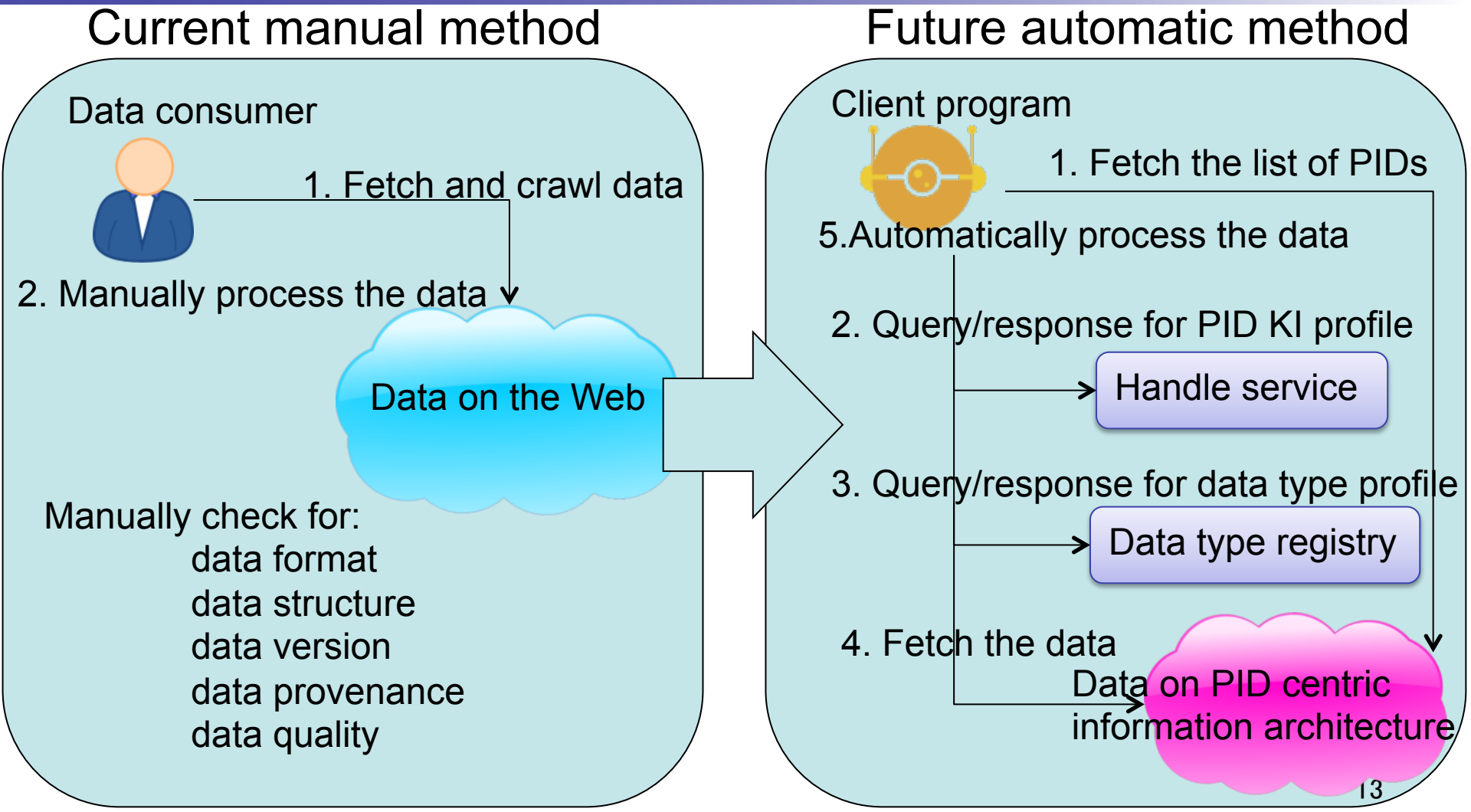
Assumption: PID is implied and needed to retrieve the profile. Has special role.

	Property identifier	Content format	Mandatory?	Explanation
1	PID	Handle	Yes	Global identifier for the object
2	RDAKIProfileType	Handle	Yes	Handle to the Kernel Information type profile; serves as pointer to profile in DTR. *DTR federation address is expected to be global (common) knowledge just like global handle server.*
3	digitalObjectType	Handle	Yes	Handle points to type defn in DTR. The type of the object (this should always be the same for this type of data, but would distinguish it from other data types). Distinguishing metadata from data objects is a client decision within a particular usage context, which may to some extent rely on the digitalObjectType value provided.
4	digitalObjectLocation	URL	yes	Pointer to the content object location (pointer to DO)
5	etag	Hex String	Yes	Checksum of object contents
6	lastModified	ISO Date	Yes	Last time of digital object modification
7	creationDate	ISO Date	Yes	Date of digital object
8	version	String	no	If tracked, a numerical version for the object
9	wasDerivedFrom	Handle	no	Transformation of an entity into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity.
10	specializationOf	Handle	no	Entity is of another shares all aspects of the latter, and additionally presents more specific aspects of the same thing as the latter.
11	revisionOf	Handle	no	A derivation for which the resulting entity is a revised version of some original.
12	primarySourceOf	Handle	no	Used for a topic refers to something produced by some agent with direct experience and knowledge about the topic, at the time of the topic's study, without benefit from hindsight.
13	quotationOf	Handle	no	Used for the repeat of (some or all of) an entity, such as text or image, by someone who may or may not be its original author.
14	alternateOf	Handle	no	Entities present aspects of the same thing. These aspects may be the same or different, and the alternate entities may or may not overlap in time.
15	hadMember	Handle	no	A membership relation is defined for stating the members of a Collection.
16	externalW3CPROVDoc	Handle	No	A URL referring to a fuller provenance document (in W3C PROV) from external repository

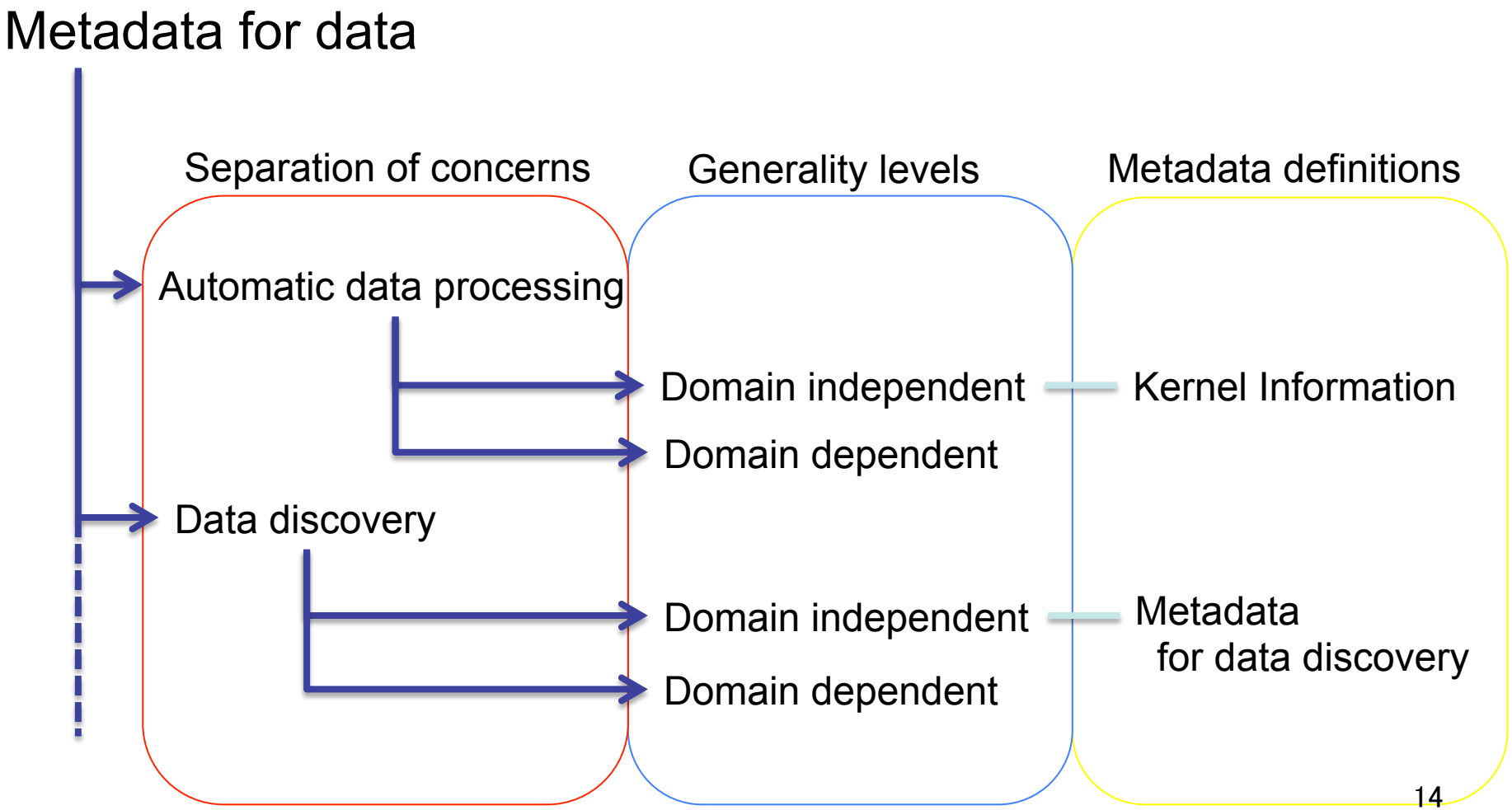
PID centric information sequence



Data processing paradigm shift



Metadata splitting



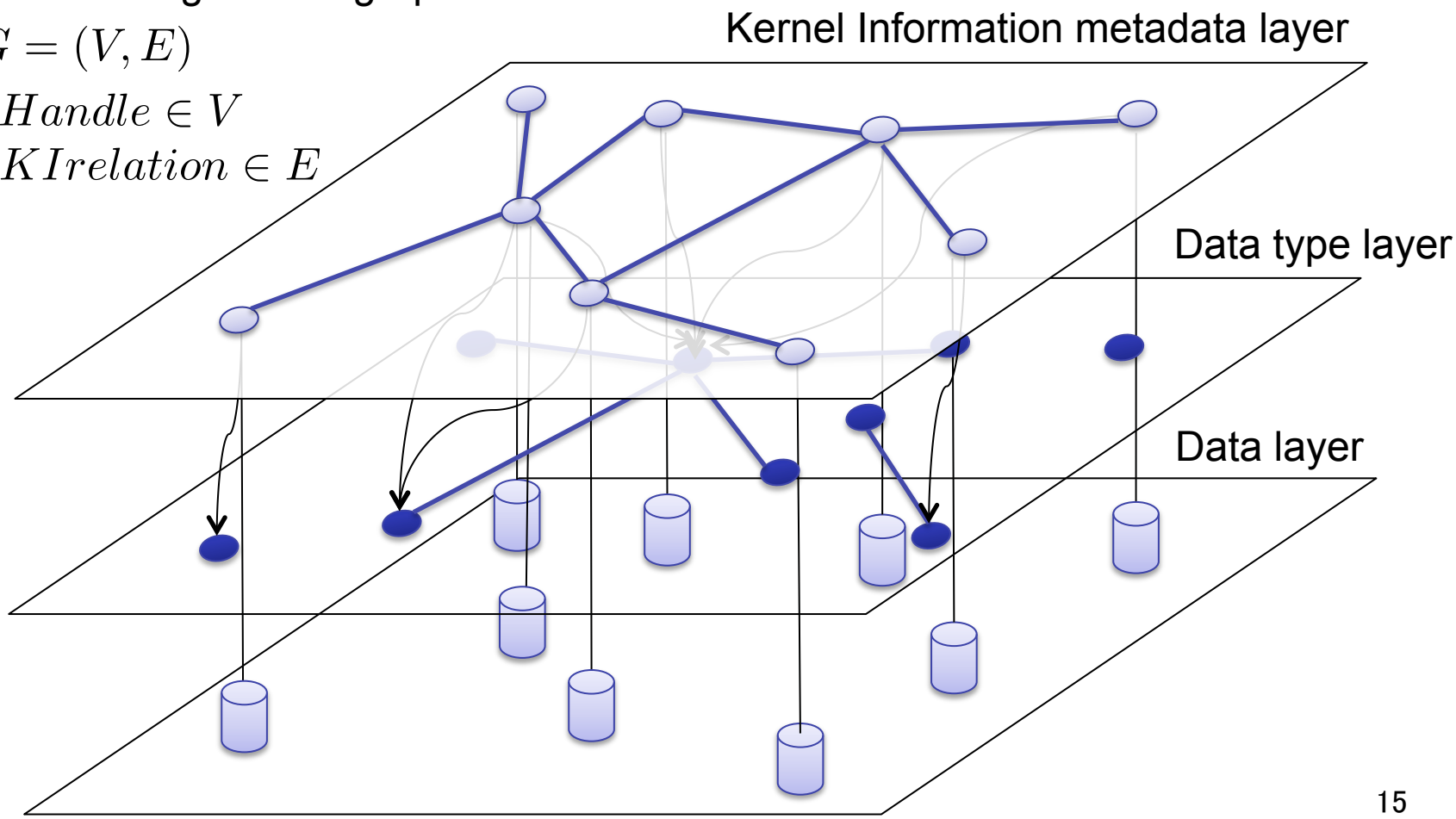
Handle-centered networks on Kernel Information metadata layer

Attribute augmented graph

$G = (V, E)$

$\mathrm{a}Handle \in V$

$\mathrm{a}KIrelation \in E$



Future directions

- Domain agnostic functionalities
- Data science approach
  - Analysis of data
  - Classification of data
  - Recommendations of data
  - Prediction of data
- On Kernel Information metadata layer,
  - Trustworthy and traceability analysis before download the data

Summary

- The objective is
  - Domain-independent automatic data processing environment on the PID centric information model for very large collections of distributed scientific data
- We introduced
  - Kernel Information as a RDA working output
    - Data
    - Data type
    - Structural relation between data
- We viewed
  - Handle-centered networks on the Kernel Information metadata layer
- Domain agnostic functionalities is emerging from
  - Graph based reasoning on the framework.

Acknowledgement

- This work is supported by the open collaborative research at National Institute of Informatics (NII) Japan (FY2017).
- The authors are thankful to all RDA Kernel Information WG members for their great discussions on remotely and in-person meetings.