Digital Objects – the core of the complex Data Market

Peter Wittenburg
Max Planck Computing & Data Facility
RDA Europe/Germany

1. Digital Objects
2. Complex Data Market for sharing/trading
3. Persistent Identifiers

Why this focus on Digital Objects?

- obviously many are concerned about how to build a manageable and easy to use data domain
- some argue that we have the **web** - was it made for such a data domain?
- some argue that we have the **clouds** - do the millions of cloud solutions address interoperability?
- some argue that we have the **FAIR** principles - do they help to build data infrastructures?

- 75/80% of scientists' time is lost for data management, etc.
- >60% of costs in industry are devoted to steps before analytics

Can a simple concept such as Digital Objects help?

„Object“ in Philosophy

- In English the word *object* is derived from the Latin *objectus* (pp. of *obicere*) with the meaning "to **throw, or put before or against**“. (Wikipedia)
- An **object** is a technical term in modern philosophy often used in contrast to the term *subject*. A subject is an observer and an **object is a thing observed**. (Wikipedia)
- B. Cassin: Platon, Aristoteles had no specific words
- An **entity** is something that exists as itself, as a subject or as an object, actually or potentially, concretely or abstractly, physically or not.
- **A property** is that which belongs to or with something, whether as an attribute or as a component of said thing.

Objects in this Talk

- Objects are **meaningful** and have names allowing us to talk about and refer to them.
- An object has properties describing its characteristics.
- A Dollar bill is an object, but it does not have a name as individual object – it‘s the mass that requires it to get a „class name“ and a **number**.
  - This number is unique in the name space defined by the Federal Reserve.
- In this case properties describe the characteristics of the class.

Objects are central for human communication/interaction.

And we can identify them.

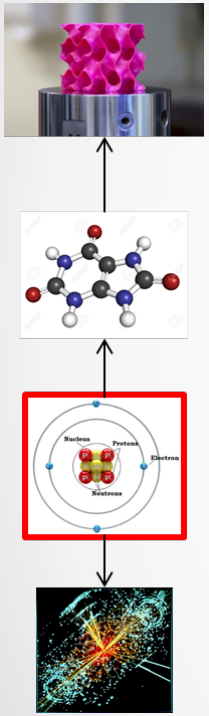
Digital Objects

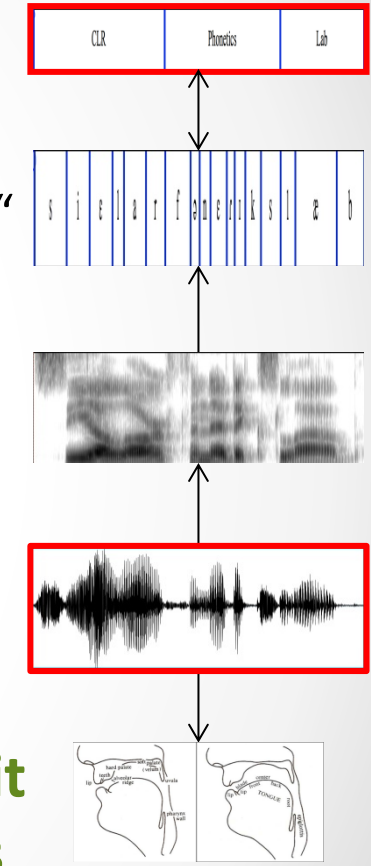
- Digital Objects are "**meaningful entities**" existing in the digital domain of bits.
- meaningful: some people want to talk about it, work with it, refer to it, cite it, etc.
- DOs can include data, collections, metadata, software, publications, queries, categories, assertions, etc.
- DOs have
  - some content represented by (structured) bit sequences (stored somewhere)
  - a name (class)
  - a „number" due to the amounts
  - properties which are described by different types of metadata

Digital Objects are central for human and machine communication. And we need to identify them.

Do we all agree?

- do we widely agree that DOs are central in the digital domain?
  - just now colleagues talking about putting „FAIR into practice“ found mentioning the term „Digital Object“ in a strategic document too „technical“ and "unknown" to include
- thus - answer is NO
  - if we see DO purely as a technical term we miss the point
  - it‘s about determining suitable conceptual layers





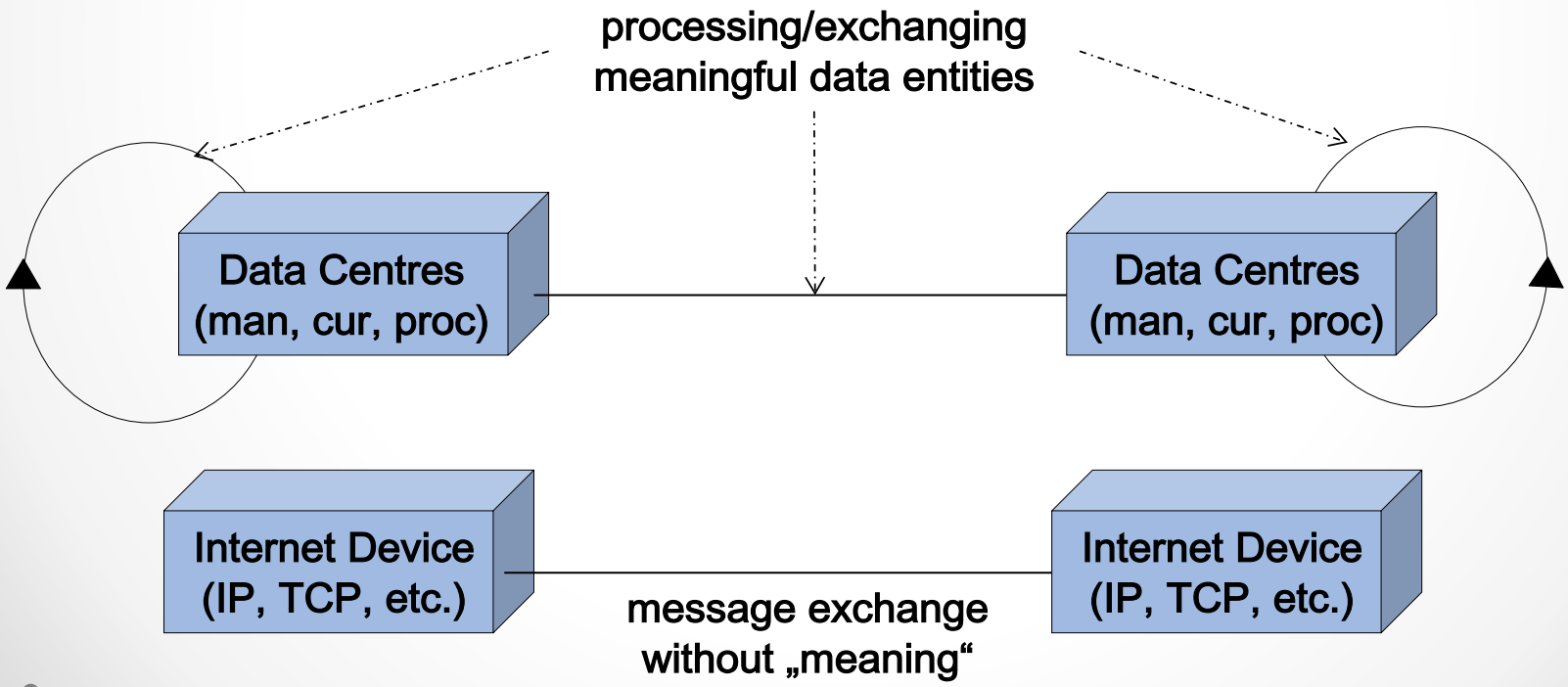
DOs are the „atoms“ of our digital domain, since it makes sense to associate relevant characteristics with them. It‘s conceptual and its time to dissiminate.

Digital Objects – looking back

- 1995 Kahn & Wilensky: DOs have structured bit sequence, persistent ID, key metadata

  (key metadata = one key-value pair to cover the PID)

- „something“ was missing after Internet



Digital Objects – looking back

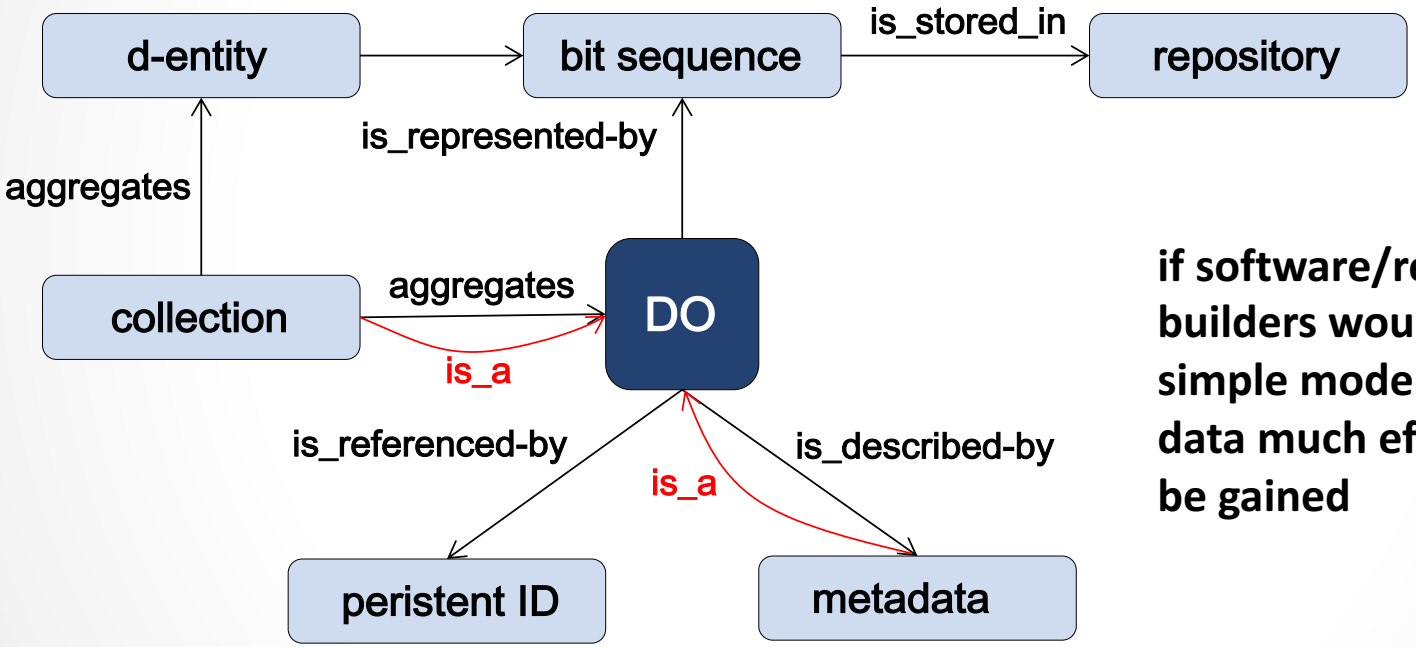
- 1997 Cross-Industry Working Team (XIWT)
  support for DO and operations on DO
- 1997+ Fedora Commons software
  (started as a joint Cornell/CNRI project, later software library for managing DOs)
- 1993+ World Wide Web took off & dominated scene
  (HTML, HTTP, URLs for referencing web information)

- 2000+ DOBES Archiving: DOA inspired and FAIR compliant
- 2006+ Amazon's Elastic Compute Cloud
  (private "Object Store", hash as PID, metadata in admin layer)

„Digital Object" concept has proven its strenghts.

RDA Data Foundation & Terminology (2013/2

- start at 1st RDA Plenary (March 2013 Gothenburg)
- all based on >20 use cases from various disciplines



if software/repository builders would follow this simple model for organising data much efficiency would be gained

Implemented by some communities to manage large collections from 2000 on (DOBES, ENES, etc.)

Digital Object Architecture (Kahn)

- **DO Repository**

  Systems where DOs are stored and give access to them
- **Identifiers/Handles Resolution System & Registration Agencies**

  Trustworthy global system to resolve Handles to “state” information.
- **DO Registry**

  A kind of metadata registry to maintain information about the DOs.
- **Security Considerations**

  PKI based security mechanism to protect Handles.

Complex Data Market

- essential drivers are billions of smart IoT sensors all producing continuous high resolutions streams
- need/wish to use data across borders/silos
- variety will be the most challenging dimension
- a few expected trends
  - data will be subject of massive exchange & processing
  - difficult to track – need new ways to identify usage/locations
  - sharing only when rights are defined and respected
  - need to separate between creators, aggregators, providers, brokers and users – currently aggregators sit on data preventing innovation
  - increasingly automatic processing of collections

Data Market and DOs

Data Market to be built on DOs

- clearly identified and described DOs allow us to know where we are talking about, what we are sharing or trading in this „gigantic data lake“
- we can search for them, access them, reuse them, archive them, etc.
- we can reference them from documents or workflows
- with sufficiently rich metadata (typing) we can automatically process them
- during processing we can create new metadata for new DOs from old metadata by adding provenance information

DOs are perfect vehicles for applying FAIR principles: to what do you want assign PIDs and MD?

Complex Data Market

need to separate between types of DOs

- **PIDs** – their resolution to be stable for very long period
- **metadata** should be open and offered
- **data** can be protected – different degrees required
- **transaction** information needs to be safe
- **smart contracts** to define usage

metadata to structure the market

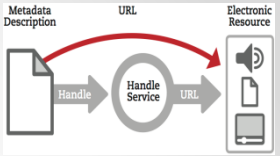
- machinery widely known (harvesting, aggregating, mapping, indexing)
- offer it with OAI ResourceSync (OAI PMH phasing out)
- need a registry of ResourceSync offers (repositories) with little metadata

Metadata Challenges

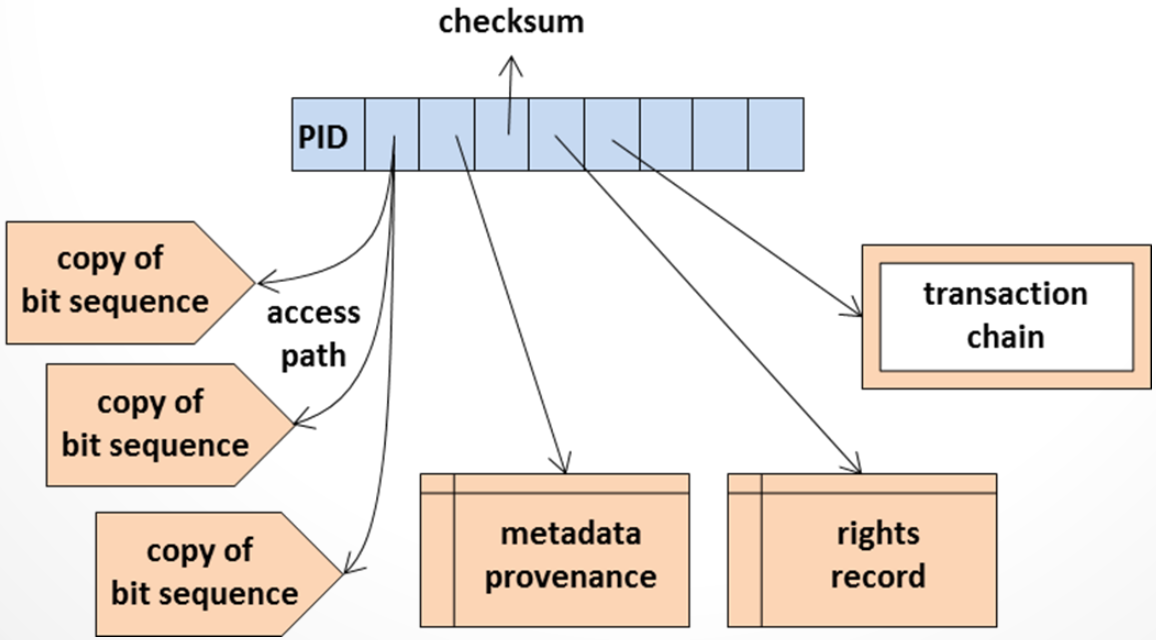
- FAIR requires rich metadata – what does this mean/ who will create it?
  - key-value pairs to describe the DOs content for others with different intentions (occasional user, scientific analysis, machine usage, etc.)
  - most problematic issues are duplication, bad quality & semantic mapping
  - usage of contextual information through relationships (LOD)
  - community standards define a familiar semantic space to help
  - strong typing incl. provenance required for automatic processing
    (CLARIN: Weblicht workflow tool for annotating texts)
  - Virtual Language Observatory: 800.000 records – how to use this?

- amounts of data require to use smart agents to find useful DOs
- brokers with specific interests will harvest and offer services – without smart mediators the data market will not take off

Persistent Identifiers are crucial



- PIDs need to be persistent – we need to make them persistent (!)
- PIDs can help to identify, check authenticity, find copies, etc.
- PID record attributes can lead us to all entities of a DO, i.e. they can take a "binding role"
- PIDs can open the way to global virtualisation (-> Larry)
- just finished a paper on PIDs agreed by delegates from 47 large EU research infrastructures (GEDE) with wide agreements



developed in RDA Data Fabric IG

worked on by RDA Kernel Information WG

DOs and the Web

- The Web is a huge information system with an enormous global impact.
- The Semantic Web started tackling the issue of variability of semantic scopes and developed great tools (XML, RDF, OWL, SCOS, etc.).
- In particular RDF assertions are extremely useful to specify and store semantic relations (–> LOD, nano publications)
- The Semantic Web uses Cool URIs to point to its entities since they are all html "pages".
- Many in data community do not rely on Cool URIs as PIDs, want a "true" identity and accept the indirection (->binding)
- Many rely on Handles for data and on DOIs for ePublications and Published Data.

It‘s time to integrate the two worlds.

Let‘s make DOs real

How to establish a functioning data market?

How to create a momentum towards more efficient integration of data from the many silos globally?

It seems to be time to implement and test a global DO centric domain of data to make us fit for the IoT challenges.

PID systems are becoming so crucial that we need to see them as „public good“.

Thanks for the attention.

Barbara Cassin: Vocabulaires Européen des Philosophies
K&W 95: http://www.cnri.reston.va.us/k-w.html
K&W 06: https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf
DOA: http://www.cnri.reston.va.us/papers/OverviewDigitalObjectArchitecture.pdf
XIWT: http://www.xiwt.org/documents/ManagAccess-1.pdf
Fedora: http://fedorarepository.org/
Amazon ECC: https://aws.amazon.com/de/ec2/
RDA DFT: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
ResourceSync: http://www.openarchives.org/rs/1.1/resourcesync
RDA GEDE: https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
RDA DF: https://www.rd-alliance.org/group/data-fabric-ig.html
RDA Kernel: https://www.rd-alliance.org/groups/pid-kernel-information-wg
Dobes: http://dobes.mpi.nl/
Weblicht: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page
VLO: https://www.clarin.eu/content/virtual-language-observatory-vlo