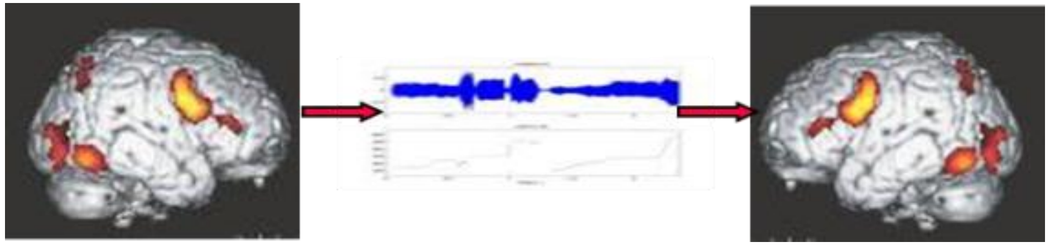
Needs and Challenges for Putting FAIR into Practice

Peter Wittenburg

Max Planck Compute & Data Facility

Research Data Alliance GEDE

Who am I – short Intro



- Responsibility for Methodology and Technology at Max Planck Institute for Psycholinguistics

  (what happens in the brain while listening, speaking, acquainting language)
- Responsibility for some large Research Infrastructures (DOBES, CLARIN, EUDAT)
- Co-Founder of Research Data Alliance and co-chairing groups

  (Data Foundation&Technology, Data Fabric, Group of European Data Experts)
- Pushing the Concept of (FAIR) Digital Objects
- Co-Editor of some “relevant” Papers

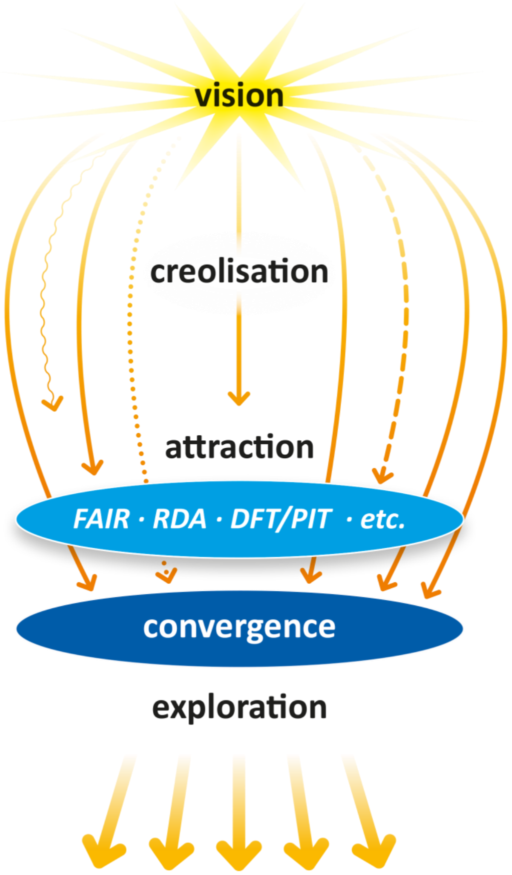
  (Riding the Wave, FAIR Principles, PID Usage, Turning FAIR into Practice, Revolutionary Infrastructures, etc.)

Kola-Sámi
Nenets/Enets
Tofa
Lower Sorbian
Hoocąk
Délı̨ne
Beaver
Archive
Svan/Udi/Tsova-Tush
Shiri/Sanzhi
Even
Salar/Monguor
Lakota
Wichita
Minderico
Khinalug
Gorani
Chintang/Puma
Tangsa/Tai/Singpho
Lacandon
Kyanga/Shanga
Semang
Chontal
Baïnounk
Kurumba Languages
Semaq Beri/Batek
Laal
Tima
Totoli
Aikanã/Kwazá
Bakairi/Mawé/Katxuyana
Oyda
Saliba/Logea
Tsafiki
Bezen
Akie
Bayso/Haro
Wooi
People of the Center
Wovia/Isubu
Sri Lanka Malay
Teop
Bakola
Cashinahua
Trumai
Central Papuan Languages
Baure
Kuikuro
Waima'a
Movima
Awetí
Ujir
Yurakaré
Aché
Taa
ǂAkhoe Haiǁom
Iwaidja
Uru-Chipaya
Tsambaan/Nsong/Ngong
Jaminjung
Marquesan
Chaco Languages
Nen/Tonda
Tuamotuan
Savosavo
Vurës/Vera'a
Ambrym Languages

MPCDF

GEDE
Group of European Data Experts

Dreams



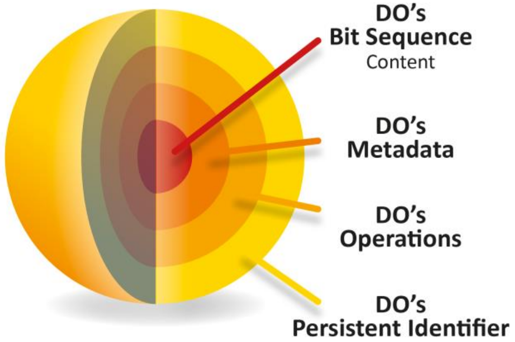
- TCP/IP brought us the world wide unified Computer Network (Internet)
- HTTP brought us the world wide unified Information Network (Web)
- ??? brought us the world wide unified Data Network (???)

- FAIR Principles – great summary of discussions (but paper work)
- RDA with about 10.000 experts working on data issues – grass-roots initiative and yet no systemic approach
- concept of FAIR Digital Objects implements FAIR principles but still no agreement about its usefulness

- but revolutionary inventions take much time (Internet ~ 30 years)

Taken from Wittenburg & Strawn
Common Patterns in Revolutionary
Infrastructures and Data

Digital Object Modell
from RDA DFT



GEDE
Group of European Data Experts

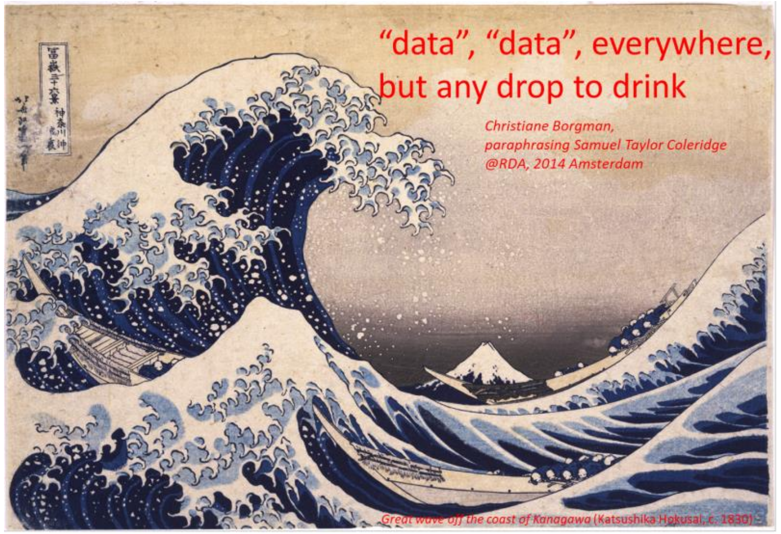
Reality I

- 80% of time & effort in data projects is spent on wrangling (science, industry, etc.)
- 60% of data projects in industry fail
- many researchers are excluded from data driven science (cross-silo/disciplinary)

- just studied ~60 RI reports deeply – some paradoxes
  - “Standards” are good for science, but researchers don’t want to change if no clear benefit.
  - Great FAIR Principles, but researchers shift changes to the end stage of a project.
  - Have huge number of tools, but they don’t help to create the unified FAIR domain.

Reality II

- just studied ~60 RI reports deeply – some paradoxes (ctnd)
  - Having increasing number of regulations (legal, ethical, formal, DMPs), but researchers shift to the start/end stage and hope on copy&paste
  - >90% of data is in the processes and little data will be published, but researchers shift actions to the last step, i.e. Open Science remains a myth – data sharing without metadata?
  - Discipline experts believe that their practices are unique, however, there are re-occurring patterns in data creation, management and processing



"data", "data", everywhere, but any drop to drink

Christiane Borgman, paraphrasing Samuel Taylor Coleridge @RDA, 2014 Amsterdam

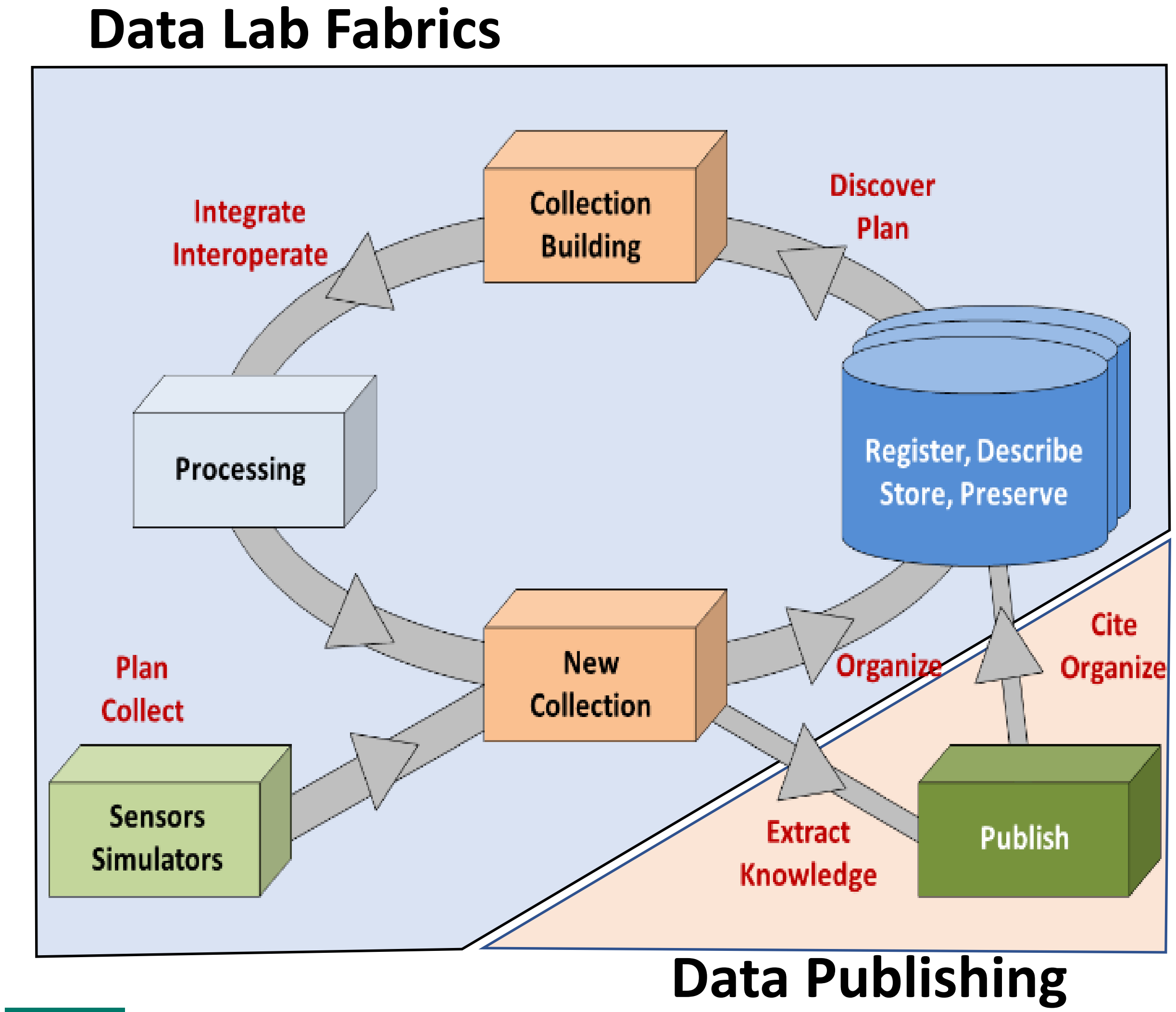
Great wave off the coast of Kanagawa (Katsushika Hokusai, c. 1830)

MPCDF

GEDE

Group of European Data Experts

Data Cycle Studied in RDA DF



The results confirmed RDA DF studies in 2014 that led to founding RDA Data Fabric:

- Much has been done to improve the last step: publication (Librarians & Publishers are very active)
- Practices in the Labs did not really change, but there is the mass of data to be re-used
- FAIR Digital Objects as a WayOut to improve practices in the labs !?

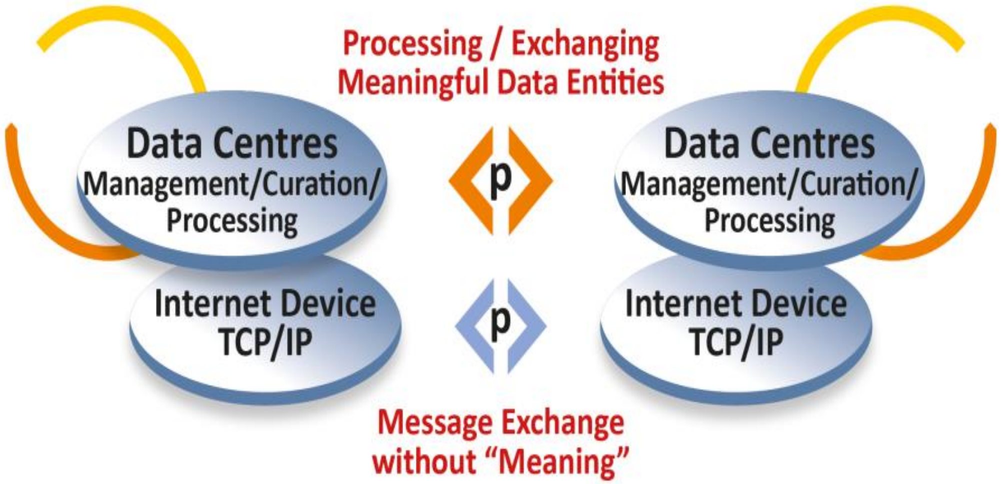
Digital Objects: Model Development I

some applications

↑

FTP
SMTP
GOPHER
etc.

early 80s



Processing / Exchanging Meaningful Data Entities

Data Centres Management/Curation/ Processing

p

Data Centres Management/Curation/ Processing

Internet Device TCP/IP

p

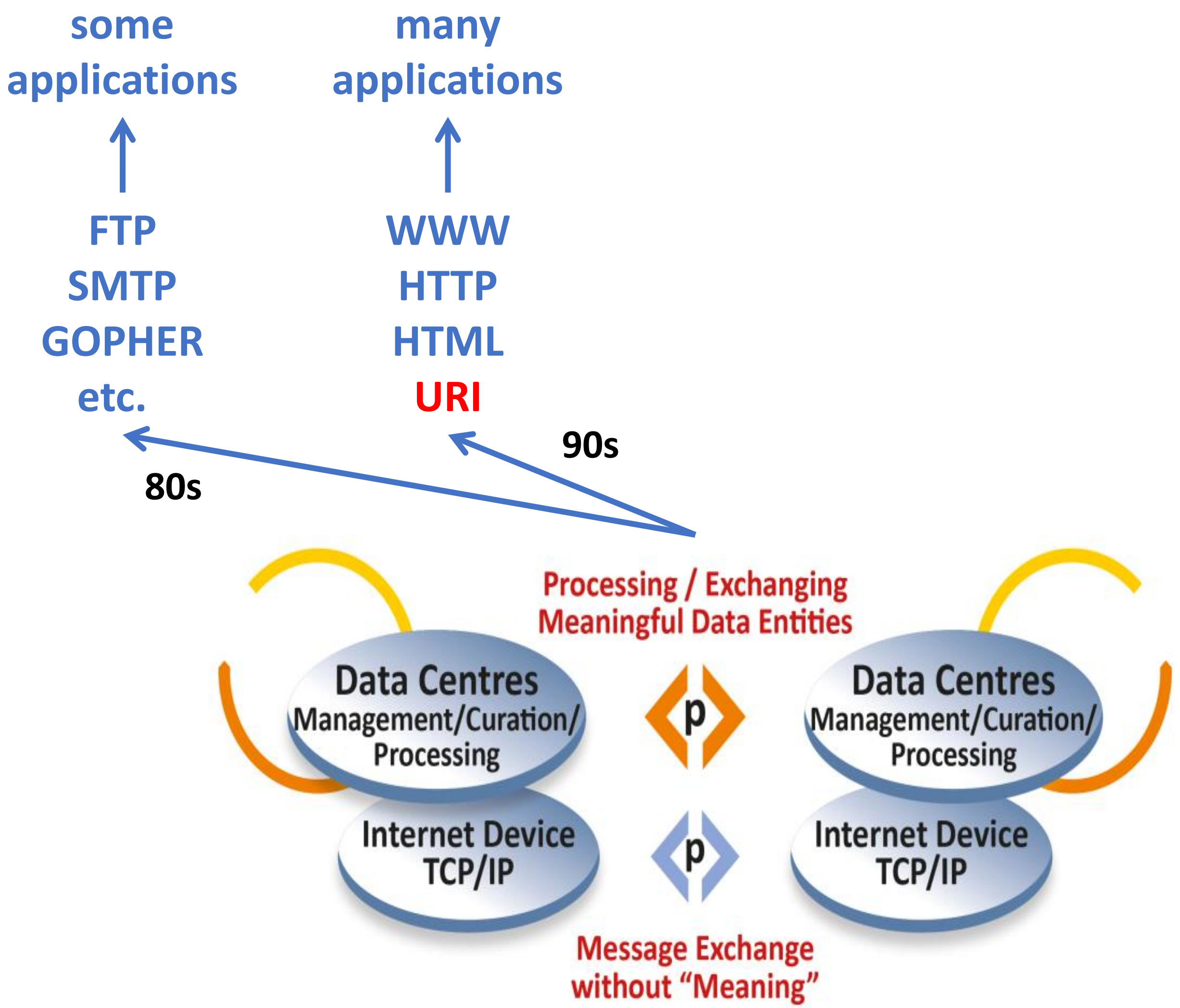
Internet Device TCP/IP

Message Exchange without “Meaning”

complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

DO: Model Development II



some applications

FTP
SMTP
GOPHER
etc.

many applications

WWW
HTTP
HTML
URI

80s

90s

Processing / Exchanging Meaningful Data Entities

Data Centres
Management/Curation/
Processing

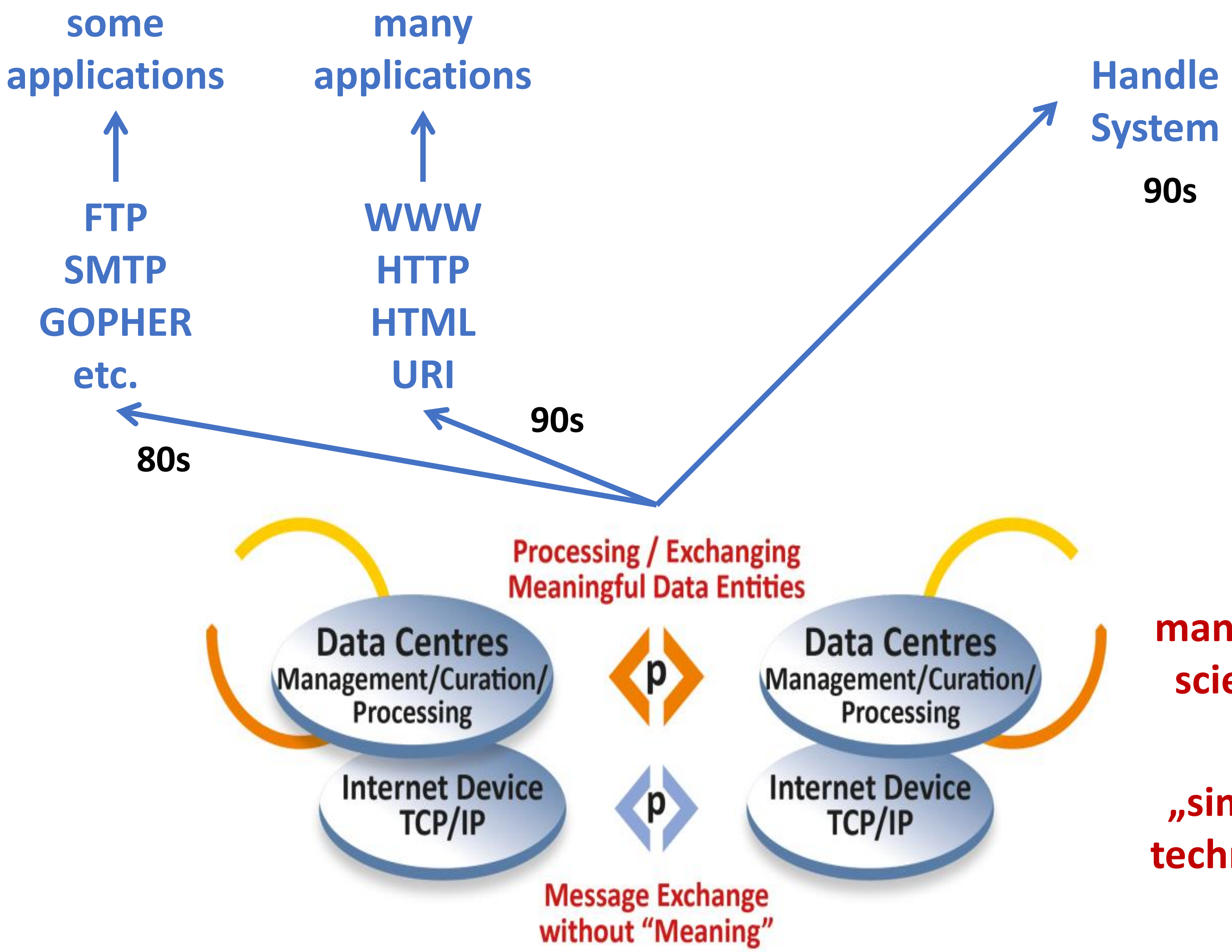
Internet Device
TCP/IP

Message Exchange without “Meaning”

complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

DO: Model Development III



some applications

FTP
SMTP
GOPHER
etc.

80s

many applications

WWW
HTTP
HTML
URI

90s

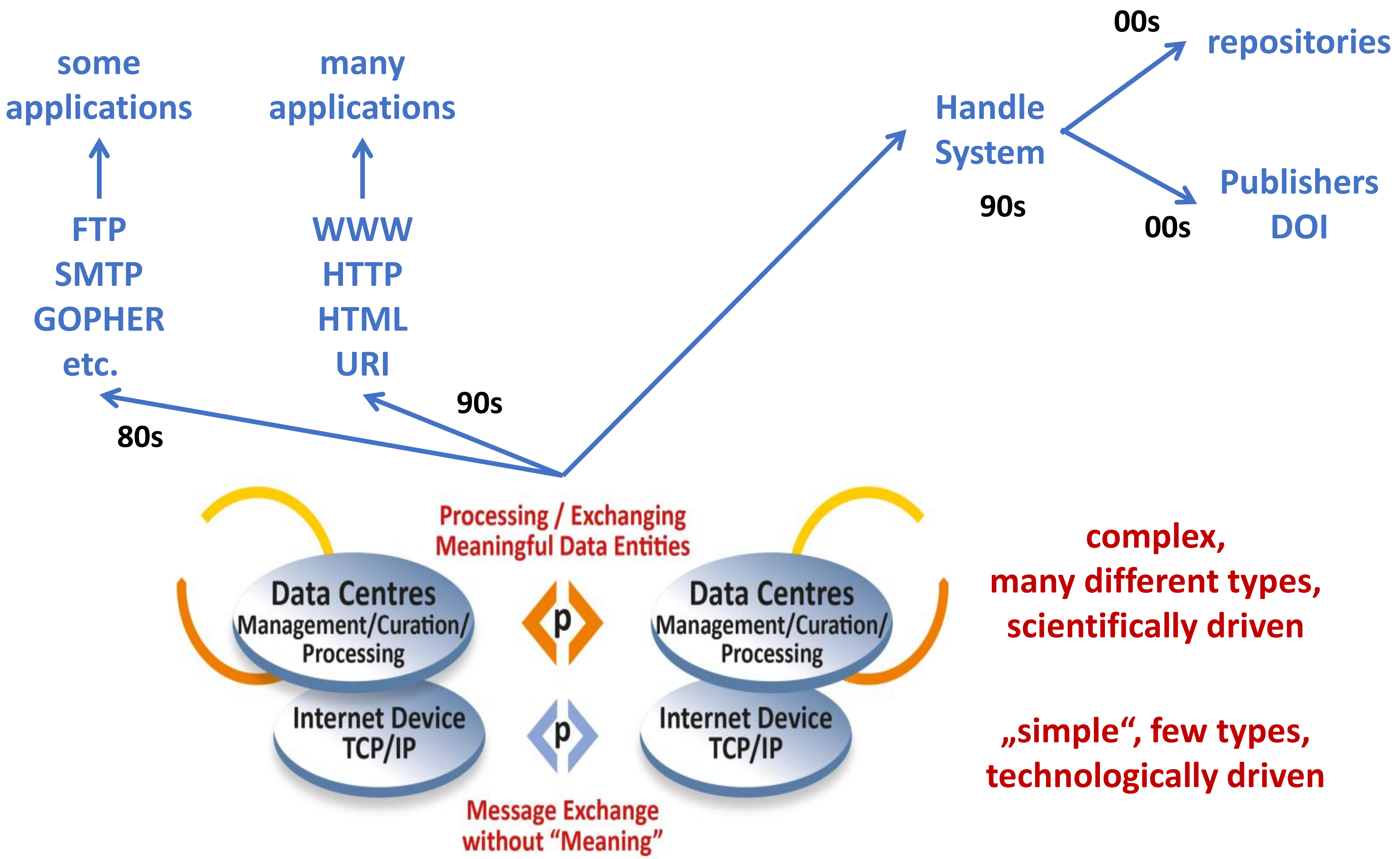
Handle System

90s

complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

DO: Model Development IV



some applications

↑

FTP
SMTP
GOPHER
etc.

80s

many applications

↑

WWW
HTTP
HTML
URI

90s

Handle System

90s

00s repositories

00s Publishers DOI

Processing / Exchanging Meaningful Data Entities

Data Centres Management/Curation/ Processing

Internet Device TCP/IP

Message Exchange without "Meaning"

complex,
many different types,
scientifically driven

„simple", few types,
technologically driven

DO: Model Development V

80s

FTP
SMTP
GOPHER
etc.

↑

some applications

90s

WWW
HTTP
HTML
URI

↑

many applications

95/06

A Framework for Distributed Digital Object Services (Kahn & Wilensky)

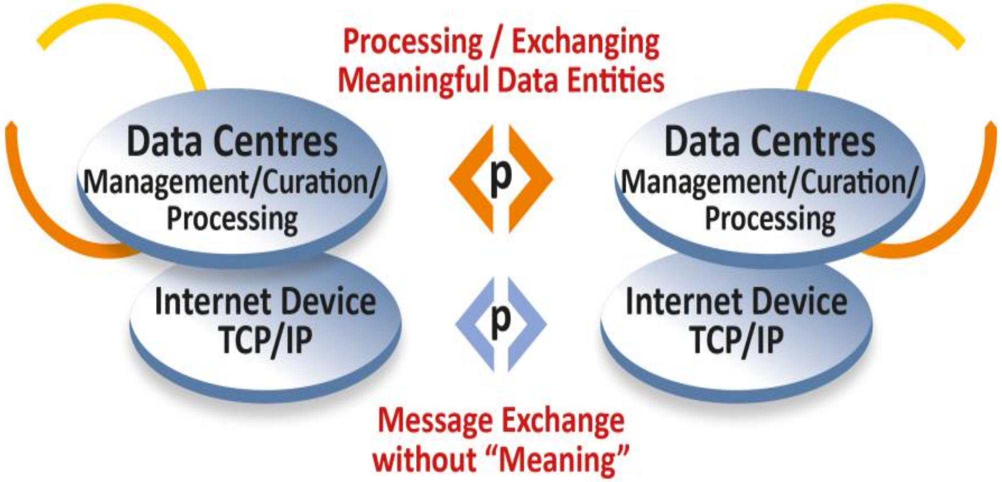
00s

DO Architecture

Handle System

repositories

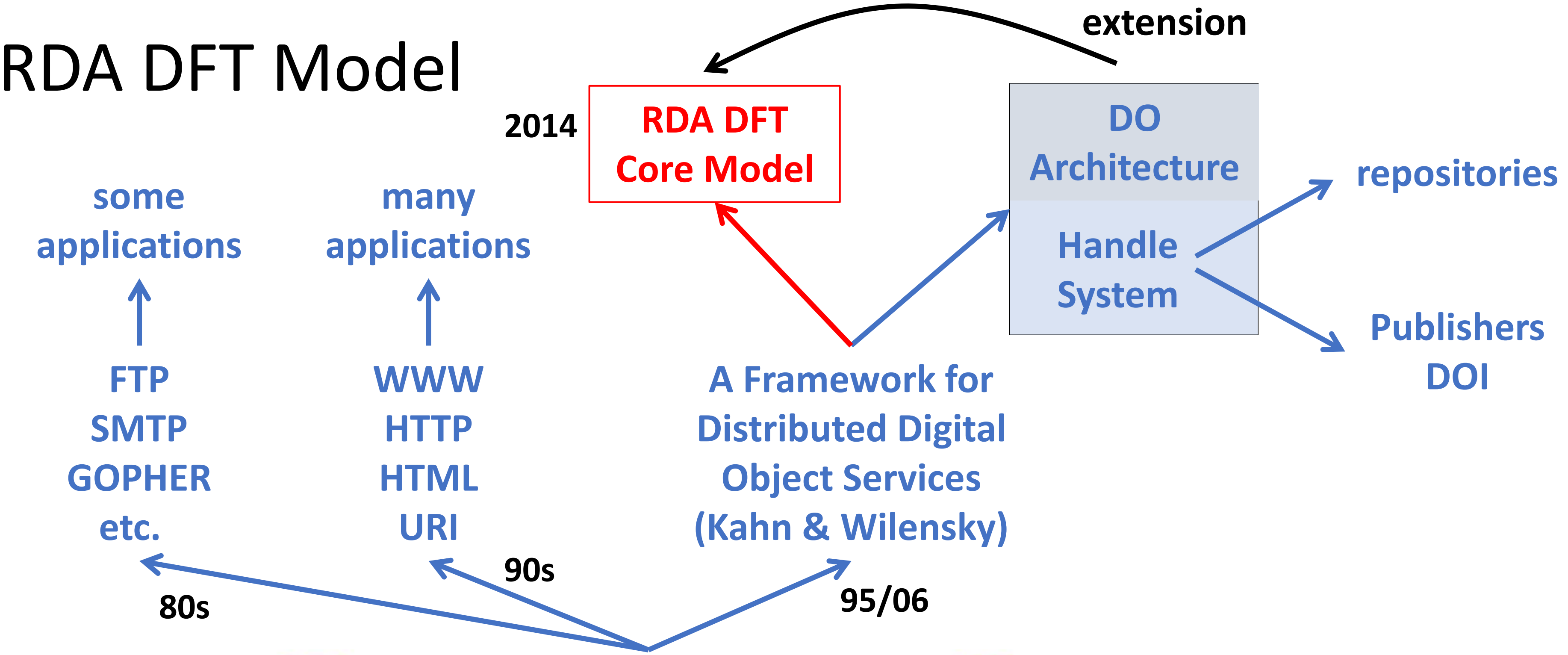
Publishers DOI



complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

DO: RDA DFT Model



extension

2014

RDA DFT Core Model

DO Architecture

Handle System

repositories

Publishers DOI

some applications

many applications

FTP
SMTP
GOPHER
etc.

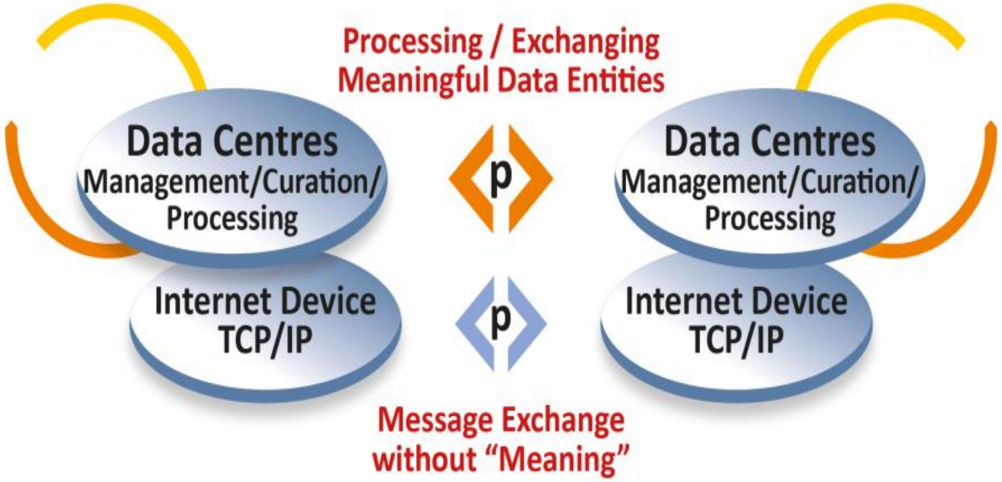
WWW
HTTP
HTML
URI

A Framework for Distributed Digital Object Services (Kahn & Wilensky)

80s

90s

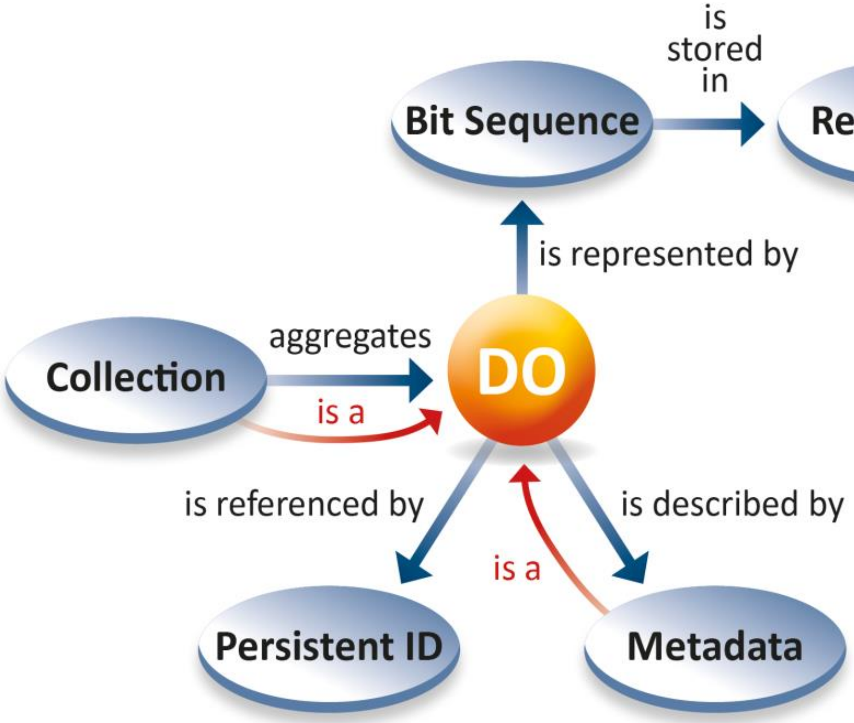
95/06

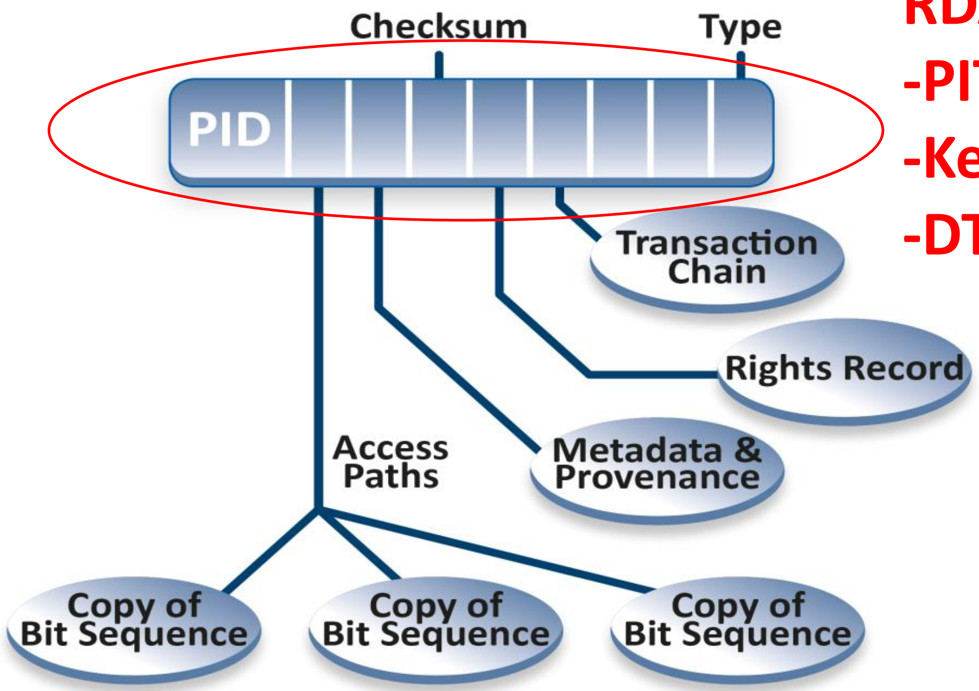

Processing / Exchanging Meaningful Data Entities
Data Centres Management/Curation/ Processing
p
Data Centres Management/Curation/ Processing
Internet Device TCP/IP
p
Internet Device TCP/IP
Message Exchange without "Meaning"

complex,
many different types,
scientifically driven

„simple“, few types,
technologically driven

DO: RDA Data Foundation & Terminology (2014)





RDA Specs

-PIT

-Kernel

-DTR

RDA DFT: a DO has a structured bit sequence stored in some repositories, is assigned a PID and is described by metadata.

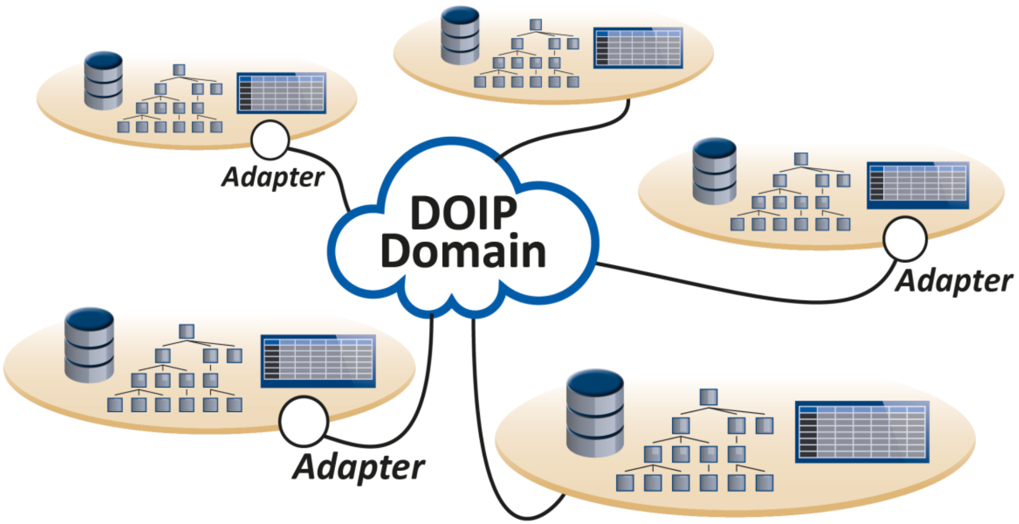
DOs can be aggregated to collections which are also DO. Metadata descriptions are DOs.

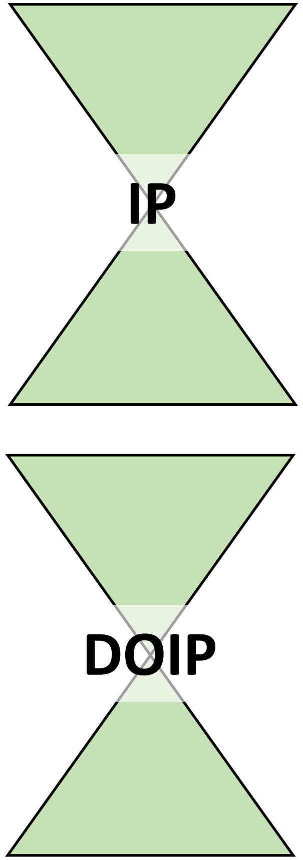
DO‘s PID Record is resolved to machine-actionable attributes enabling human/machine actions.

PID = globally unique persistent resolvable identifier (Handle, DOI)

also Software available: DOIP V2.0 (DONA)

- improved specification and implementation of DO Architecture
- DOIP V2.0 specifying unified client – DO Server interaction
  - CORDRA reference implementation ready
  - DOIPV2.0 SDK ready
  - all based on PIDs





Cannot expect people to start from Scratch

Do DOs support FAIR?

some applications ↑ FTP SMTP GOPHER etc.

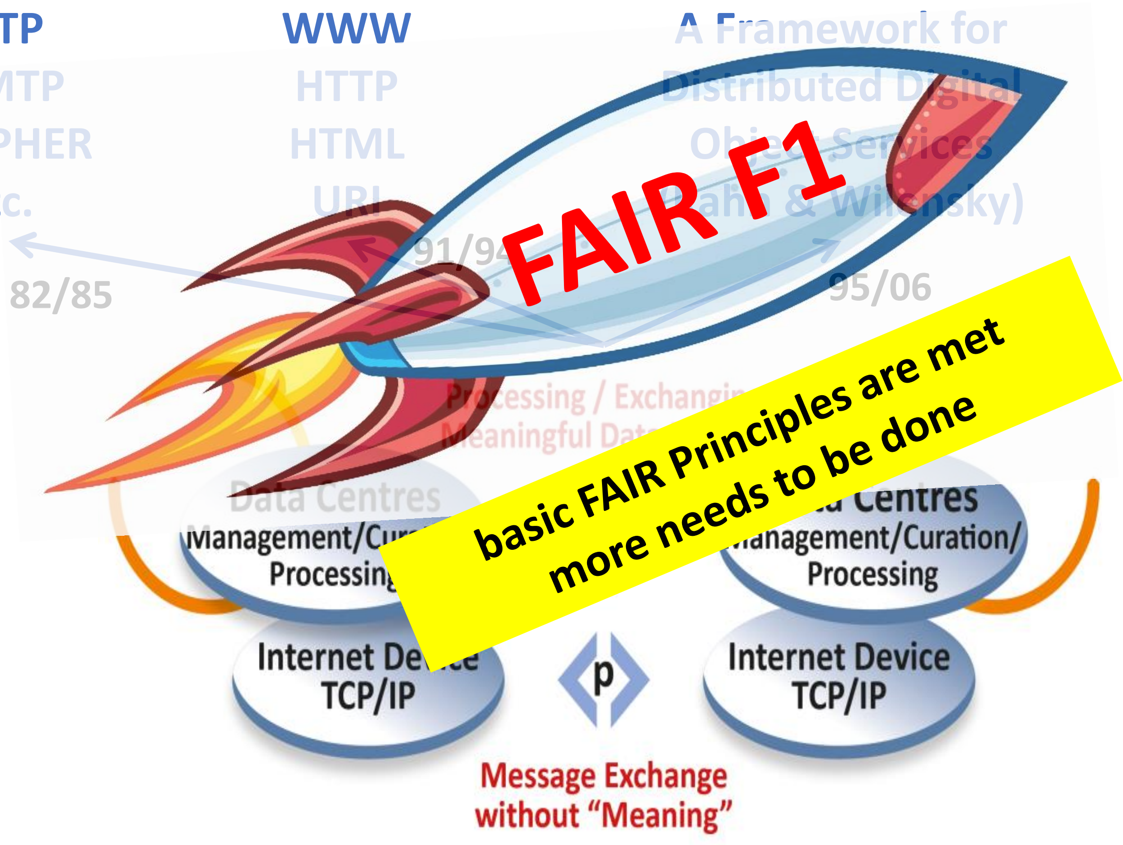
many applications ↑ WWW HTTP HTML URI

RDA DFT Core Model

DO Architecture

Handle System → „large“ repositories

Handle System → Publishers DOI


A Framework for Distributed Digital Object Services (Kahn & Wilensky)
82/85
91/94
95/06
FAIR F1
Processing / Exchanging Meaningful Data
Data Centres Management/Curation/Processing
Internet Device TCP/IP
p
Message Exchange without "Meaning"

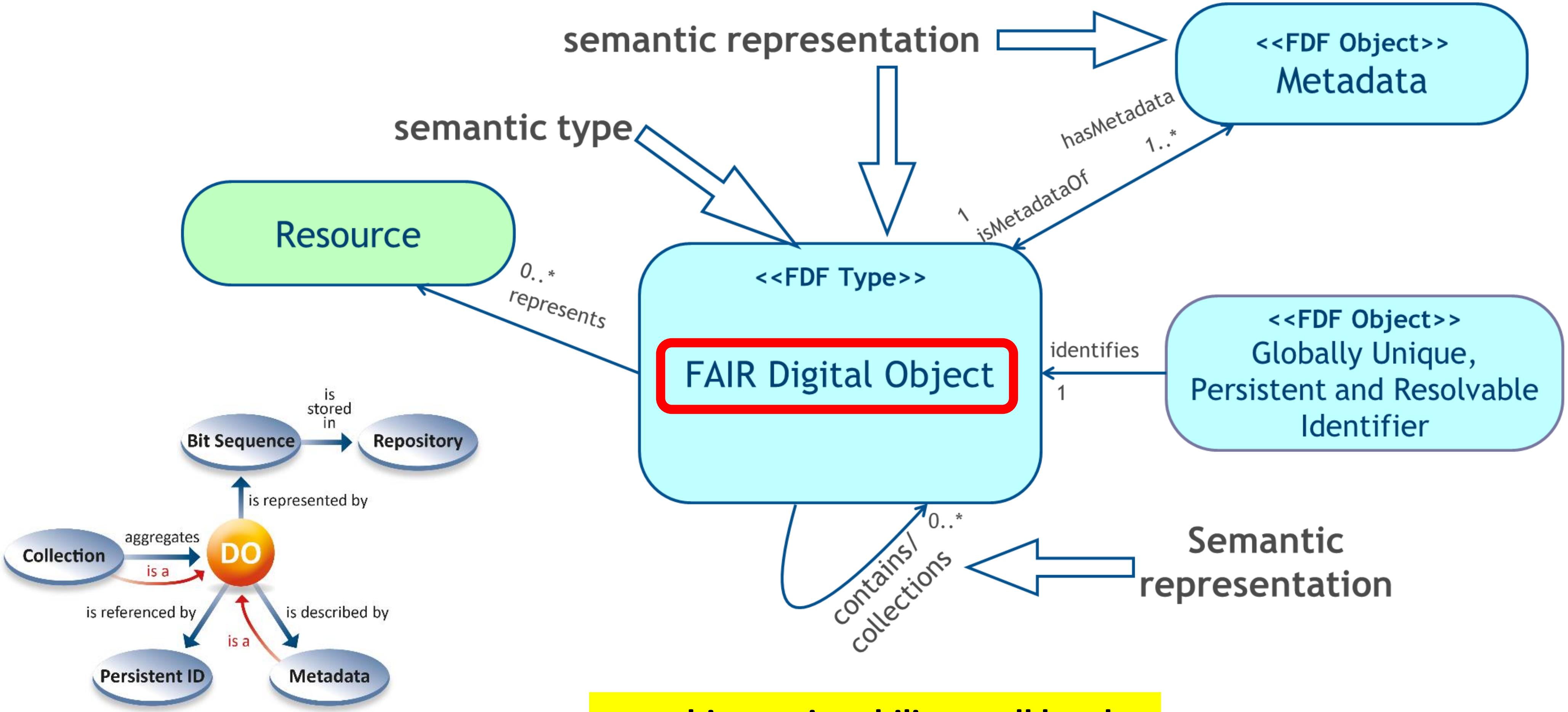
basic FAIR Principles are met more needs to be done

complex, many different types, scientifically driven

„simple“, few types, technologically driven

FAIR requires Semantic Explicitness

(in close collaboration with Luiz Bonino, applying mechanisms from LD)

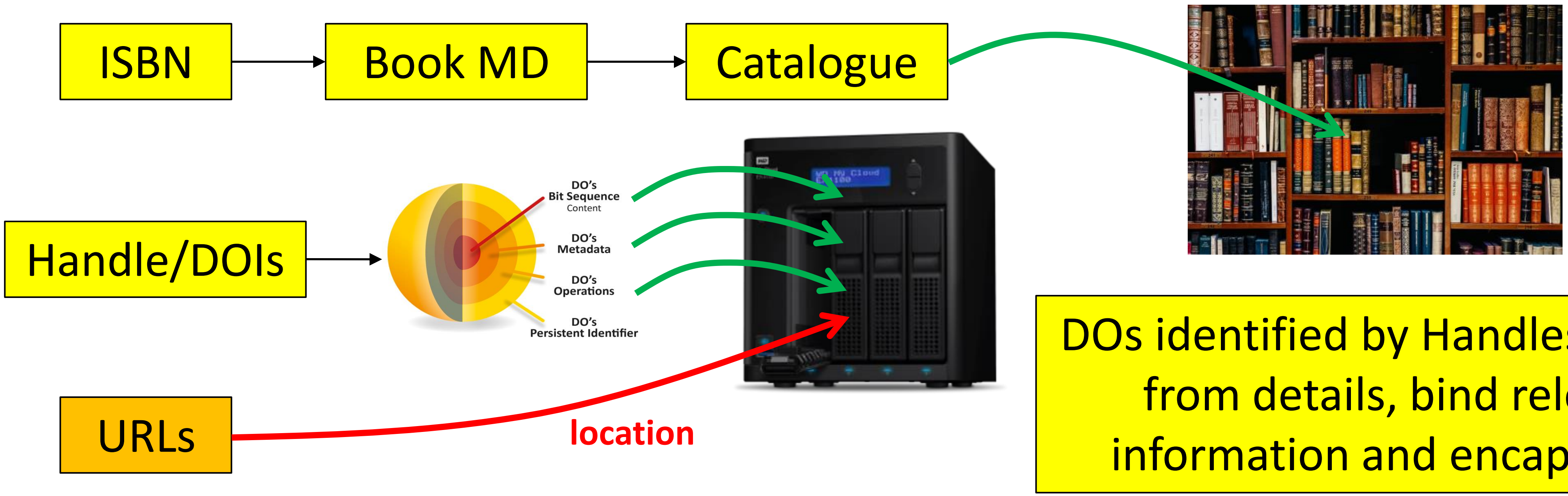


machine actionability at all levels
what about metadata ???

Long Term Vision & Identification (FAIR F1)

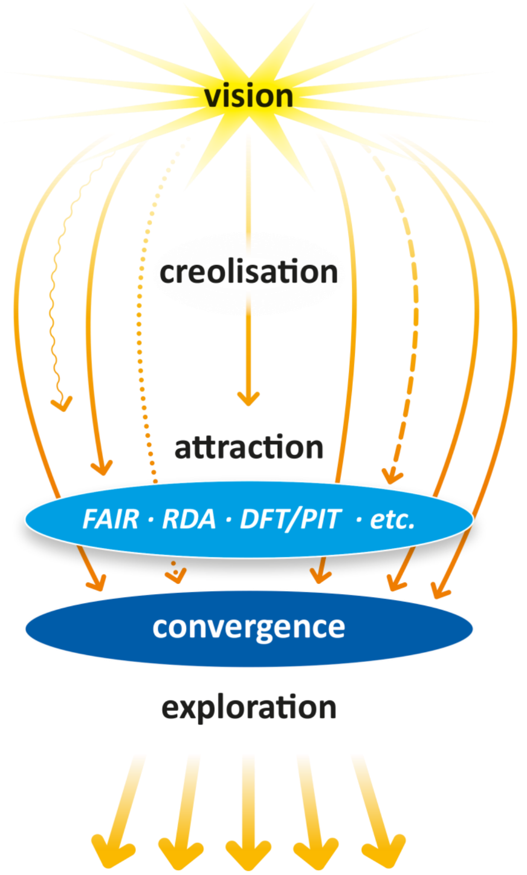
- V. Cerf: warning for a dark digital age

- why?
  - it‘s about persistence of relevant bit-sequences, describing metadata AND **relations** for **100+** years
  - and relations will express much of our cumulative scientific knowledge



DOs identified by Handles abstract from details, bind relevant information and encapsulate!

All Ready for a Big Change?



- **NOOooOooo**
- FDO not yet accepted broadly – many different voices how to build a global unified data infrastructure (yet no help from EOSC)
- Researchers are right to be careful:
  - no stability yet – still much dynamics in convictions, trends
  - miss supporting software to reduce the load for researchers
- Thus, if we want to change practices
  - need to take the researchers with us who are not interested in technicalities
    - offer the obvious (Zenodo, B2Share, Handle/DOI, etc.)
    - address data sovereignty
    - need to be patient, nevertheless work hard on DO SW components
  - Interested? – Join the GEDE DO and CWFS discussions (Canonical Workflow Frameworks for Science)

Thanks for the attention.

all can be found under GEDE – Github: https://github.com/GEDE-RDA-Europe/GEDE
just search for „Github GEDE“