GEDE-DO/C2CAMP
Scientific Use Cases for DOs

Peter Wittenburg
Max Planck Computing & Data Facility

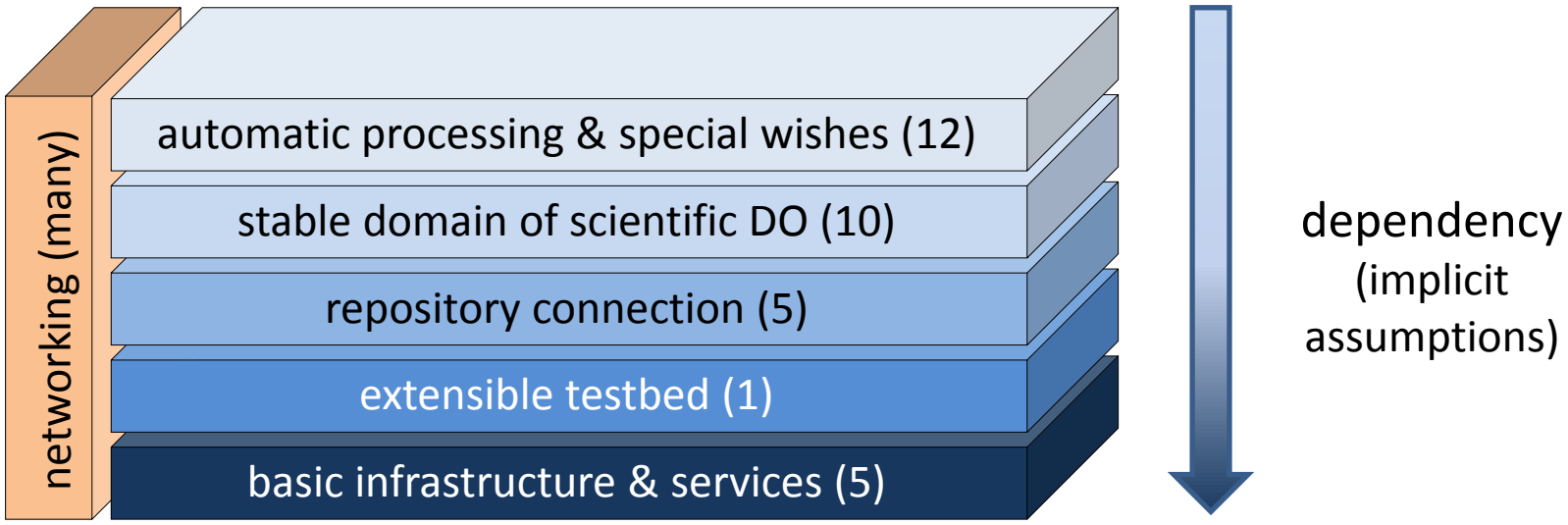
Background

- better understanding how FAIR DOs can help DIS
- 5 papers elaborating on FAIR DOs have been written
  - Wittenburg & Strawn: Common Patterns in Revolutionary Infrastructures and Data
  - Wittenburg, Strawn, Mons et.al.: Digital Objects as Drivers towards Convergence in Data Infrastructures
  - Strawn: Open Science, Business Analytics, and FAIR Digital Objects
  - Wittenburg: Commenting on “Digital Object” Aspects
  - Schultes & Wittenburg: FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure (not yet published)
- DONA: DOIP v2.0 Document

- missing is a paper about what FAIR DOs can do for science
- Request for use cases: 34 responses, some already active
- Koenraad, Dimitris & Peter now editing a paper

Response Classification

- 34 reponses and 3 former contributions from 23 different RI



- EUDOn networking proposal resubmission in September
  (about 150 mentioned interest)

Basic Infrastructure, Services and Operation

- >6 initiatives are interested to offer stable services allowing others to build on FAIR DOs: DONA, CNRI, CNIC, GWDG, RDA, IDF, etc.
  - maintenance and dissemination of basic concepts (FAIR-DO, DTR, etc.)
  - DO Interface Protocol (now V2.0)
  - Handle System
  - Handle Service Providers
  - CORDRA
  - Kernel Information Types
  - Data Type Registry Instances
  - etc.

  - in addition many repositories and registries

Extensible Testbed

- Indiana U (US)
  - already get funds to extend their pilot towards an extensible testbed for DO technology and approaches
  - collaborating closely with CNRI
- seems that CNIC is also interested to setup a testbed

Connection of Repositories

- explicitly mentioned by ENES, CLARIN, NOMAD, VAMDC, ICOS
  - the connection would be a basic step to join the DO domain
  - initiatives will not change their repository setups
  - repositories are organised in very different ways
    - data model (rel DB, no-SQL DB, files, cloud objects, etc.
    - data organisation (data, different types of metadata, PIDs)
    - organisation of work (storage, content management/curation, etc.)
    - types of services for ingest, management, access, etc
- started a separate overview (see later)

Stable Domain of Scientific DOs I

- 10 responses mostly ESFRIs with different similar intentions:
  - can DOs help to structure the domain of digital entities in the disciplines and to create stability over decades
  - much similarity beyond just connecting to DOIP
- DISSCO (biodiversity)
  - have huge arsenal of specimen which are the basis
  - need to establish various classifications and annotations systems on top which is a representation of the field knowledge
  - need to rely on stable and persistent relationships based on clear identities
- ECRIN (medical trials)
  - highly sensitive medical trial data are exchanged for research purposes
  - a big interest in tracing state of DOs and reusage combining DO and blockchain technology
- EISCAT (athmospheric obervations with antenna fields)
  - measurements of antenna field will be combined in different ways
  - only clear identities and provenance tracking will prevent chaos

Stable Domain of Scientific DOs II

- ELIXIR (biomed domain)
  - looking for ainteroperability platform across all their domain-borders that has the potential to offer stability across decades
  - finally they see DO‘s potential for complex computational systems and workflows
- Wageningen U (agricultural domain)
  - looking at complete food chains from „cow“ to „milk powder“ engaging all kinds of activities and actors/components in this process
  - they also see DO‘s potential to build a well-structured domain of digital entities to manage the complexity of these processes
- E-RIHS (cultural heritage)
  - they see the potential of DO‘s to make digital humanities a data-driven discipline
  - utterly complex landscape of types and relationships (ontologies) where they also hope to come to a stable landscape of digital objects

Stable Domain of Scientific DOs III

- MIRRI (connecting microbial databases - biology)
  - want to integrate many fragmented databases all with different properties (structure, classifications, etc.) to have a corpus that can be analysed efficiently
  - see DO as a way to implement FAIRness and to make integration workable
- INSTRUCT (structural biology)
  - also want to integrate many fragmented databases all with different properties to have a corpus that can be analysed efficiently
  - see the potnetial of DOs to go towards automated workflows
- GESIS (social science)
  - have many ideas to combine social data with geo data and others
  - consider a stable domain of linked data based on FAIR DOs
- ForumX (experimental sciences and cross-disciplinarity)
  - looking for a stable integration approach across various exp. disciplines (economy, psychology, neuroscience, etc.)
  - see the cross-disciplinary potential of the DO approach and want to offer it as a general solution for the German National Research Infrastructure

Automatic Processing I

- 5 proposals explicitly mention concrete workflow plans
- CLARIN (language domain)
  - want to extend their current WF orchestration environment to a more flexible switchboard solution where MIME indicators are not sufficient
  - DO concept with its typing&binding is flexible enough to implement this
- ENES (climate modeling)
  - see 3 areas for advanced workflows to cope with mass of data: automated data management, automated support for processing stages, actionable digital collections
  - FAIR-DOs with binding & encapsulation seem to be the way to go
- NOMAD (material science)
  - offer currently a download kit for doing analyses on the aggregated calculations
  - see the potential to use FAIR-DOs for also offering an orchestration and workflow framework

Automatic Processing II

- DEWcom (computation)
  - want to extend the classical cloud computing concept to a distributed scenario – thus combining clouds with grids and have a software stack to run DIS jobs
  - want to extend their software to include FAIR-DOs and thus to solve issues such as proper provenance tracking based on clear identifications
- CNRI Eager (computation)
  - studying how types can unleash the rich domain of possible operations using type registries
  - also here the potential to improve context and provenance recording without human intervention is in the focus

Special Needs

- 5 proposals were submitted with special needs
- CLARIN (language domain)
  - have a toll to build virtual collections (could be used across disciplines)
  - want to extend this tool to support FAIR-DOs fully
- ENES (climate modeling)
  - want to design a VRE offering all kinds of tools and services where abstractions are essential as FAIR-DOs offer them – so should be based on DOs
- NOMAD (material science)
  - all material scientists are required to use lab-books to describe their experiments and document their actions
  - would like to combine these entries with blockchain technology using FAIR-DOs
- GOFAIR (broad initiative)
  - working hard on the knowlet concept to build machineries to better understand and analyse the complex domain of assertions
  - FAIR-DOs with their indetification, binding could form the needed stable basis
- VAMDC (atomic&molecular physics)
  - scientists do work based on data in existing databases – but no link back from results to these databaes
  - suggest to use FAIR-DOs for improved provenance tracing, error tracking etc.

Summary

- there is an interest to experiment with FAIR-DOs and to use their structuring and persistence potential
- there is a need to have operational basic services for everyone (24/7)
- there are common patterns in many suggestions

- the plans are at varying stages – some already started, some designing systems, many wait on funds

GEDE-DO/C2CAMP Repository Adaptation

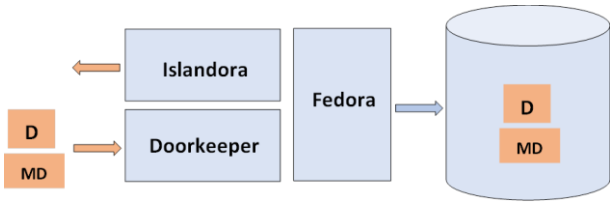
Peter Wittenburg
Max Planck Computing & Data Facility

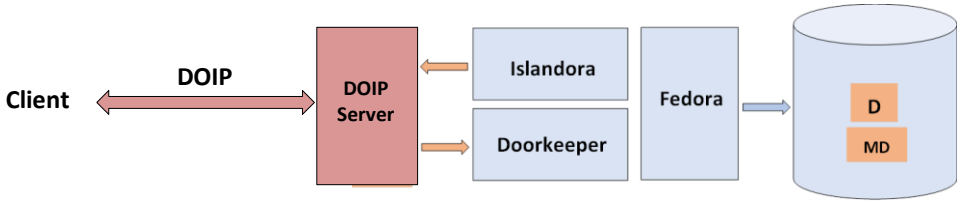
Overview

- 4 initiatives sent responses – time was too short, hopefully more to come
  - DOBES Archive (language domain)
  - NOMAD Archive (material science)
  - ENES Repository infrastructure (climate modelling)
  - U Sheffield (space science & solar physics)
- goals are
  - repositories are key pillars in most research infrastructures
  - high investments during the last two decades, i.e. chosen structures cannot be changed over night
  - efforts need to be done to connect existing structures to DOIP
  - questions:
    - can we reduce the adaptation effort to describe structure typesand develop software packages?
    - is there a typical approach for distributed repository landscapes with different actors (see ENES)?
    - etc.

DOBES Case with Central Archive

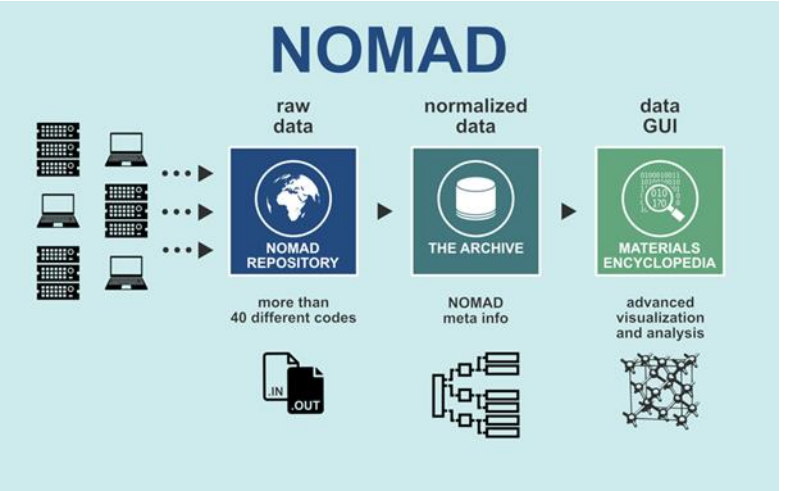
- principle data structure is based on „bundles“, small collections of streams that share a time axis and are often processed jointly
  - the bundle has a Handle and metadata
  - each stream incl. metadata has a PID





- archive is now based on Fedora Commons
  - using a special doorkeeper taking various actions at upload time and knowing about bundles
  - for access the Islandora package is being used
- DOIP is simple and does not know about bundles, i.e. bundle logic needs to be in the client, Doorkeeper functionality needs to be integrated into the DOIP server
- adaptation is fairly straightforward, to get it right 6 PM should be sufficient

NOMAD Case with Central Archive)



- many labs are contributing to the archive by uploading their calculations mostly in form of aggregated ZIP files
- dependent on the calculations metadata is complex structured in form of trees
- at upload time only minimal metadata
- step 1: upload of calculations into workspace
- step 2: managers put raw files into repository extended by first metadata
- step 3: after normalisations and metadata extension all into archive
- step 4: access is supported by encyclopedia guiding the users
- for DOIP adaptation we need to separate these steps
  - step 1: no real use for DOIP
  - step 2: upload into repository could be done with help of DOIP
  - step 3: upload into archive could be done with help of DOIP
  - step 4: encyclopedia could include DOIP to access data + metadata
- adaptation would be straight forward, 6 pm should be sufficient

ENES Case with Distributed Repositories

- adaptation is not only a technical, but in particular a social problem
- many federated repositories and services developed by different players
- challenge 1: how to understand all aspects in such scenario, no central point, different services created by different teams (in copy action metadata needs to be modified at both sides)
- challenge 2: transport of huge payloads (big fils) – have their solutions
- challenge 3: protocol wrapper (DOIPV2.0 with JSON, all solutions as REST, i.e. no direct solutions for standard libraries (Java, Python, etc.)

- DOIP V2.0 adoption not straightforward
- a detailed analysis would require 2-3 PM
- some core functions such as ingest may take 4-5 PM
- things need to be carefully tested and then deployed – more PMs

- no chance to do real work without special funding

Sheffield U Case

- collect massive amounts of data from space and solar measurements
- complex workflows are in use to extract features to determine space weather
- they need to move towards PIDs and FAIR-DO to improve efficiency in their data work due to the many different relationships between data sets

- need deeper insight in DOIP
- they have already several software stacks for managing these data flows and to extract knowledge

- there is an interest to adopt the DO concept, but too early for them to make estimates of efforts.

Summary

- too few cases to make realistic statements
- in centralised setups it might be straightforward to adopt DOIP, but some intelligence needs to be dealt with at clients since DOIP supports atomic cases
- an effort of about 6 pms is estimated to get the adaptation done
- in distributed setups with different players the social aspects come into the focus, i.e. developing demonstrators to convince all partners, synchronising work at different software stacks is required, etc.
- not realistic effort estimates can be made without a careful analysis

- **need far more adaptation cases ☺**