Research Objects

More than the Sum
of the Many Parts

Carole Goble
The University of Manchester, UK
EU Infrastructures ELIXIR-UK, FAIRDOM, BioExcel, ISBE ...
Software Sustainability Institute UK
Digital Objects Wholes & Parts in an Expanding Ecosystem

A Digital Package Object Type composed of many interrelated elements

A Digital Object that represents properties in common across all research artefact types, Common PIDs and Metadata

A Digital Package Object that bundles together and relates digital resources of a scientific investigation with context

Citable Reproducible Packaging

Nested content
Heterogeneous elements.
Distributed and embedded content.
Externally stewarded content.
Checklists + Checksums
Workflow driven Data Analytics: Research Components are Many and Various
Collection in a Data Catalogue

Workflows of local or remotely executed codes


16 datafiles (kinetic, flux inhibition, runout)

19 models (kinetics, validation)

13 Standard Operating Procedures

3 studies (model analysis, construction, validation)

24 assays/analyses (simulations, model characterisations)

Systems and Synthetic Biology: Research Components are Many and Various
Citation G. Penkler; F. du Toit; W. Adams; M. Rautenbach; D. C. Palm; D. D. van Niekerk; J. L. Snoep; (2014): Glucose metabolism in Plasmodium falciparum trophozoites; FAIRDOMHub. 
http://dx.doi.org/10.15490/seek.1.investigation.56
Multi-results & Versions
Data of many types...Primary, secondary, tertiary...
Methods, models, scripts ...
Physical objects: samples, strains, specimens

Structured organisation

Distributed: Span repository silos, regardless of location and ownership
In house and External
Multi-site + multi-stewardship

Retaining context over fragmentation
Spanning across the Ecosystem

Publishing & Exporting
Snapshot versions and elements
DOI proliferation

Context
Digital Object Perspectives

Artefact Context
- workflow engine
- execution environment
- Inputs
- Outputs
- Parameters
- Configs
- Intermediates
- Workflow Run
- Provenance

Repository Context
- Pubmed
- Github
- figshare
- Docker image
- Slideshare
- Arxiv.org
- CommunityDB

Narrative
- Document References

Experimental Context
- Study
  - Investigation
  - Analyse(s)
  - Assay(s)
  - Research Object
  - ISA

Digital Object Perspectives

...
preserved, portable research products.
inter-platform exchange
multi-platform content & context dependencies

The Power of the Commons:
To dramatically decrease the length of inter-lab knowledge turns and optimize the pace of discovery

New Treatment!

New Discovery

[Josh Sommers]
- **Data** used and results produced in experimental study
- **Methods** employed to produce and analyse that data
- **Provenance** and settings for the experiments
- **People** involved in the investigation
- **Annotations** about these resources, to improve understanding & interpretation

Bechhofer et al (2013) Why linked data is not enough for scientists [https://doi.org/10.1016/j.future.2011.08.004](https://doi.org/10.1016/j.future.2011.08.004)
Research Objects
Analogous to Software artefacts and practices rather than Data or Articles

Atomicity, Granularity, Aggregation Composition Fragmentation Versioning Forking Cloning Portability Dependency management
Research Objects
4+1 Architecture Framework

Drivers

“logical” view
METADATA
PROFILES

“process” view
POLICIES &
GOVERNANCE

“development” view
LIFECYCLE
STEWARDSHIP

“physical” view
SERVICES &
TOOLS
Drivers within the Ecosystem

Commons & Catalogues
Publishing,
Exchange between people and platforms
Sharing,
Training

Replication
Reproducibility
Preservation
Portability

Active Research Release
Evolution & Snapshots
Remixing, Comparison, Review
Automated processing

Conservation
Repair
Archive

Goble, De Roure, Bechhofer, Accelerating Knowledge Turns, DOI: 10.1007/978-3-642-37186-8_1
ROs working across the ecosystem

(Inter)National Commons
Public Community Archives
Publishers

Institutional repositories

Personal Electronic Lab Notebooks,
Project / Group Commons
Research Context

Boundary Objects


The ‘last mile’ challenge for European research e-infrastructures https://doi.org/10.3897/rio.2.e9933
FAIR ROs Desiderata

Use Open Standards. Incremental.

Technology Independent.


Multi-platform.

Commodity tooling.

The least possible.
The simplest feasible. Low tech.
Low user overhead and thin client.
Manifests of Metadata

Manifest Construction
- Identification: to locate and cite
- Aggregates: to enumerate & link together
- ROs and Elements
- Annotations: about RO, elements & their relationships

Manifest Profile Description
- Type Checklists: what should be there
- Provenance: where it came from
- Versioning: its evolution
- Dependencies: what else is needed

Container
researchobject.org
Standards & COTS Platforms

**Identifiers**: URI, RRI, DOI, ORCID

Open Archives Initiative
Object Exchange and Reuse
Aggregation

W3C Web Annotation Vocabulary
Annotation

Manifest

Construction

Open Archives Initiative
Object Exchange and Reuse
Aggregation

W3C Web Annotation Vocabulary
Annotation

Container

BagIt

research

Object

.org

ZIP

docker

VAGRANT

conda

bioboxes

bioconda
Linking across ROs and into the Linked Open Data Cloud

- Recording & linking together the components of an experiment
- Linking across experiments.
- Linked ROs
- Semantic Web + Digital Objects
Goldilocks Profiles & Progression Levels to define and interpret content

- its evolution
- where it came from
- what else is needed
- what should be there for types

<table>
<thead>
<tr>
<th>Versions</th>
<th>Provenance</th>
<th>Dependencies</th>
<th>Checklists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fedora</td>
<td>Apache Taverna</td>
<td>puppet, CHEF</td>
<td>FAIRDOM</td>
</tr>
<tr>
<td>Mendeley Data</td>
<td>KNIME</td>
<td>Galaxy Project</td>
<td>Galaxy</td>
</tr>
<tr>
<td>Digital Commons</td>
<td>Galaxy Project</td>
<td>Galaxy Project</td>
<td>Galaxy</td>
</tr>
<tr>
<td>openBIS</td>
<td>Galaxy</td>
<td>Galaxy</td>
<td>Galaxy</td>
</tr>
<tr>
<td>SEEK</td>
<td>ProvONE</td>
<td>BioContainers</td>
<td>BioContainers</td>
</tr>
<tr>
<td>GitHub</td>
<td>wfprov</td>
<td>BIOMD</td>
<td>BIOMD</td>
</tr>
<tr>
<td>PAV</td>
<td>ReproZip</td>
<td>Combine Archive</td>
<td>Combine Archive</td>
</tr>
<tr>
<td>Zenodo</td>
<td>ProVOL</td>
<td>BioCompute Objects</td>
<td>BioCompute Objects</td>
</tr>
<tr>
<td>W3C PROV</td>
<td>VoID</td>
<td>&lt;DATS&gt;</td>
<td>&lt;DATS&gt;</td>
</tr>
</tbody>
</table>

- Project / Lab Specific
- Community-based Types, Context
- All

Profiles & Progression Levels
simplify the solution space
but still encode data types

Common properties among content types

Minimum information for one content type

Minim model for defining checklists

Profiles & Progression Levels

Profiles

Progression Levels

Minim model for defining checklists


http://purl.org/minim/description

W3C Shape Specs

http://purl.org/minim/description
Validation and Monitoring Tools
interpret the content

http://www.rohub.org/
Making use of these various objects will depend on available infrastructure & tools etc.

[Lilian Gorea, Oluwatomiode Fasugba]
Research Object Manifest Model
http://www.researchobject.org/specifications/

RDA Data Foundation and Terminology WG Core model.

Bioschemas.org
schema.org
Data Package

Dockeronto

Lots of roads from A to B
A publishing trend…. JSON(-LD) + schema.org

https://dokie.li/

https://linkedresearch.org

Manifest: Schema.org, JSON-LD, RDF
Archive: .tar.gz

Reproducible Document Stack project
eLife, Substance and Stencila

DataCrate Specification
BagIT data profile + schema.org JSON-LD annotations
schema.org tailored to the Biosciences

simple structured metadata markup on web pages & sitemaps

Finding, Citing, Metadata Exchange

Bioschemas Profiles
Schema.org Types

Data
Tools
Training
...

Bioschemas.org

Data repository
Training Resource
Data repository

Search engines
Registries
Data Aggregators

Standardised metadata mark-up

Metadata published & harvested without APIs or special feeds

FORCE11
NIH
NIH Big Data to Knowledge (BD2K)
elixir

InterPro
Pfam
Reactome
UniProt
PDB
UniProtKB/COPaKB
CATH/Gene3D
AGI
transPLANT
Brassica
European genome-phenome archive
Human Protein Atlas
MarRef
Identifiers.org
FAIRsharing.org
OOG
InterMine
BioSamples
RD Connect
Sample Catalogue
molgenis

Finding, Citing, Metadata Exchange
Bioschemas.org

schema.org tailored to the Biosciences

**simple structured metadata markup** on web pages & sitemaps
don’t register – harvest & index

**Bioschemas Data Profiles**

- Specific for life sciences
- Extends existing Schema.org types
- Focused on few types and well defined relationships
- Minimum properties for finding and accessing data
- Best practices for selected properties
- Managed by Bioschemas.org

**New proposed Bioschemas profiles**
- Chemical
- Beacon
- Protein
- Sample

**Bioschemas profiles**
- Catalogue
- Dataset
- Sample
- LabProtocol

**Schema.org types**
- Catalogue
- Dataset
- Record
- Entity
- LabProtocol

**Types proposed to Schema.org by Bioschemas**

**Schema.org Types**

- Generic data model
- Generous list of properties to describe data types
- Managed by Schema.org
Research schemas

Common Research **Types**

**Common Research Profiles**

**Specific Research Profiles**

- Bioschemas
- Agroschemas
- Astroschemas
- Earthschemas
- Biodiversitieschemas...

**Maintain common profiles** across scientific domains focused on finding and accessing data and exchanging metadata in catalogues. **Serving Cloud Services & Supporting Boundary Objects**
**Research Object Bundles for Data Releases**

- **Dataset “build” tool**
  - Standardised packaging of Systems Biology models
  - European Space Agency RO Library Everest Project
  - Metagenomics pipelines and LARGE datasets
  - European Space Agency RO Library Everest Project
  - U Rostock

- **Asthma Research e-Lab sharing and computing statistical cohort studies**
  - Public Health Learning Systems

- **ISA based packaging, snapshotting, exporting and publishing for Systems Biology models**
  - FAIRDOM
  - Precision medicine NGS pipelines regulation
  - STELAR

- **Manifest description of CWL workflows**
  - Open PHACTS

- **Research Object Bundles for Data Releases Dataset “build” tool**
  - NIH Big Data to Knowledge (BD2K) BDS
  - ISI, USC

- **Precision medicine NGS pipelines regulation**
  - Cannes, France
Easy to make
Hard to consume
Generic vs Specific
Don’t be too flexible!

Complex Objects types
Multi-artefact Objects

Seeding critical mass
research
object.org
Concrete Focus

Reproducibility
Commons
Portability

Tools

Driver

Platform & user buy-in from the get-go
Passionate, dedicated leadership

Community

ROs acceptance in the ecosystem

Computational Workflows & Pipelines
Multi-disciplinary investigations
Fast sampling for quantitative microbial metabolomics: new aspects on cold methanol quenching: metabolite co-precipitation

Abstract
The intra- and extracellular concentrations of 13 metabolites were measured in chemostat (O.1 1·h⁻¹) anaerobic cultures of high-yield Saccharomyces cerevisiae CEN.PK 113-7D growing on minimal media. Two independent sampling workflows were employed: (i) conventional approach. Metabolites were quantified in different sample fractions (total, extracellular, quenching supernatant, metabolite mass balance). The different method in combination with absolute metabolite quantification by gas chromatography (GC–MS) was used as a benchmark to assess quality of the cold methanol quenching procedure. Quantitative conclusions derived from different quenching techniques indicate asymmetric loss of the total mass of volatile metabolites. Pterosterol resulting from the cold methanol quenching process is present in considerable amounts of precipitated lipids. Quantitative analysis has revealed significant co-precipitation of some extracellular metabolites together with those in the lipids, with large extracellular mass-fracture. The precipitation of the lipids is a typical non-dissolved phenomenon. The co-precipitation is a highly significant phenomenon in culturing conditions (i.e., media composition) and chemical properties of the particular metabolites. Thus, methods of cold methanol quenching might be uncertain and variable based on co-precipitation described phenomena.
Computational Workflow Research Objects

Community led standard way of expressing and running workflows and the command line tools they orchestrate, supporting containers for portability.

Gathers CWL workflow descriptions together with rich context and provenance using multi-tiered descriptions. Snapshots the workflow. Relates to other objects.
Hettne KM, et al (2014), Structuring research methods and data with the research object model: genomics workflows as a case study. J. Biomedical Semantics 5: 41
Download as a Research Object Bundle

Permalink to snapshot the GitHub entry and RO identifier

Over an active GitHub entry for an actively developing workflow

Special Tooling: Common Workflow Language Viewer

https://view.commonwl.org/
Precision Medicine
High Throughput Sequencing, from a biological sample to biomedical research and regulation

[Image of a flowchart showing the process of using high throughput sequencing in precision medicine.]

sample

sequencing run
produced files are massive in size
transfers are slow

file transfer

archival
data retrieval

difficult to validate/verify

knowledge extraction

computation pipelines

hard to keep forever, not standardized

experiment

sample

how can industry and FDA work together to avoid mistakes?

FDA

regulation

analysis and review

[Vahan Simonyan]

FDA BAA contract HHSF223201510129C (PI: Raja Mazumder)
Precision Medicine High Throughput Sequencing, from a biological sample to biomedical research and regulation

Emphasis on parametric domain and robust, safe reuse.

Linked Data, JSON-LD, Ontologies (EDAM, SWO)

researchobject.org

COMMON WORKFLOW LANGUAGE

data formats, elements and APIs for EHR & genomics

Alterovitz, Dean II, Goble, Crusoe, Soiland-Reyes et al Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results, biorxiv.org, 2017, https://doi.org/10.1101/191783
Ecosystem of tools and services for big data analysis and sharing in an ecosystem

Assemble, share, and analyze large and complex multi-element datasets to integrate into biomedical HTS analytic pipelines
Secure large scale moving of patient data

1000s of images and genome sequences assembled from diverse repositories, data distributed across multiple locations, referenced because big and persisted, efficiently moved by Grid technologies

Chard et al I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets, https://doi.org/10.1109/BigData.2016.7840618
Chard et al: I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets, https://doi.org/10.1109/BigData.2016.7840618
Management in the Ecosystem

whole > $\sum$(independent)parts

Inspirations:
software development &
old skool data integration
The challenge for all the data-commons initiatives — is that many online datasets are annotated with metadata that are simply terrible…. Creating good metadata takes considerable work ....

When investigators act in their own self-interest, taking short cuts to generate metadata as quickly as possible, we should expect that the overall utility of the resource will decline.

The creation of a data commons requires the ability to deal with extremely varied — and often unanticipated — metadata patterns and data types .... a need for easy-to-use solutions that are generic to provide guidance over the entire life cycle of metadata — streamlining metadata creation, discovery, and access, as well as supporting metadata publication to third-party repositories”
Stewardship in a multi-component, evolving ecosystem

Dependencies & Responsibilities with multi-stewardship at different granularities

Who manages the RO and who manages and governs the parts?

Who maintains the manifests?

Delegation and trust!
Expect component rot
A new career?
Multi-Stewardship of ROs and Elements and the stewardship of manifests.....

Different granularities
Domain and type specific standards, lifecycles, behaviours

Atomicity
Composition
Dependency

Fixity, verifying intended versions of contents, element change detection, snapshots

Stewardship hand-offs
Multi-stewardship guarantees
- Content change
- Content decay
- PID and Resolution services
- Provenance attribution
- Credits

Who is responsible? Spectrum of governance?
Delegation and degradation.
Creation, Credit, Curation

Authenticity, Tamper-proofing
- Hashing & Checksums
- Secure signature & probity services
- Block chain & Ethereum

DOI proliferation
- Channelling for Counting
- Landing Pages

Katz and Smith “Contriponents”
- Micro-credit and citation aggregation
- Tracking RO usage & indirect contributions
- Awarding fractional weighted credit to contributors
- Networked Credit maps*

Trend - bottom up initiatives sheltered by big umbrellas

- Grassroots community activities
- Fostered by Infrastructure Initiatives
- Don’t swash the start up!
- Open standards and lightweight
- Practical engineering
- Keeping it simple and real
- **Ramps** rather than Revolution
Thank You!

Barend Mons
Sean Bechhofer
Matthew Gamble
Raul Palma
Jun Zhao
Josh Sommer
Matthias Obst
Jacky Snoep
David Gavaghan
Stuart Owen
Finn Bacall
Paolo Missier
Phil Crouch
Oscar Corcho
Dan Katz
Arfon Smith

Mark Robinson
Alan Williams
Jo McEntyre
Norman Morrison
Stian Soiland-Reyes
Paul Groth
Tim Clark
Alejandra Gonzalez-Beltran
Philippe Rocca-Serra
Ian Cottam
Susanna Sansone
Kristian Garza
Daniel Garijo
Catarina Martins
Alasdair Gray
Rafael Jimenez
Iain Buchan
Caroline Jay
Michael Crusoe
Katy Wolstencroft

http://www.researchobject.org
http://wf4ever.org
http://www.fair-dom.org
http://seek4science.org
http://rightfield.org.uk
http://www.bioschemas.org
http://www.commonwl.org
http://www.bioexcel.eu

All the members of the Wf4Ever team
Colleagues in Manchester’s
Information Management Group,
ELIXIR-UK, Bioschemas
researchobject.org
Semantic Bindings for non-embedded metadata
Bind Grid Entities and Knowledge Entities

An overview of S-OGSA: A Reference Semantic Grid Architecture
One of the mistakes that’s been made repeatedly in open scholarly communication projects has been the attempt to create the bucket of everything.
RO based ecosystems

Beyond islands with the odd special ferry-services (e.g. GitHub software releases archived in Zenodo to get DOI)

Bridges to connect anything-toanything. ROs carried between or collected in depots

Ecosystem evolution
• Mitigate
• Adapt
• Migrate
• Die

Ecosystem incentivisation
• Love
• Money
• Fame
• By Side effect