CORBEL shared services for life-science

4th AGM, 02 March 2020, Brussels, Belgium

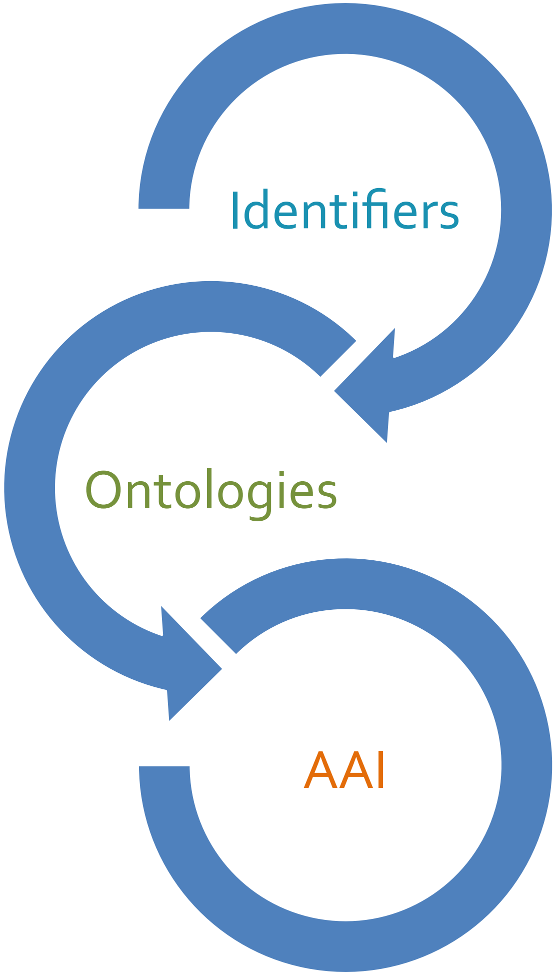
WP6 – Data access, management and integration

Carole Goble on behalf of **Helen Parkinson**, Morris Swertz, Jason Swedlow, Thomas Keane, Ilkka Lappalainen and WP6 Partners

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654248.

WP6 GOALS & OBJECTIVES

Improve interoperability and standards across BMS ESFRIs



identifier best practices, resolution and services

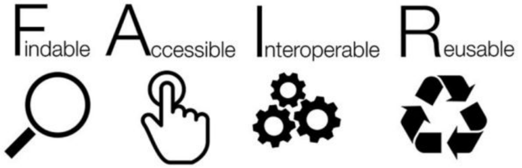
standards compliant web and programmatic based access to ontologies

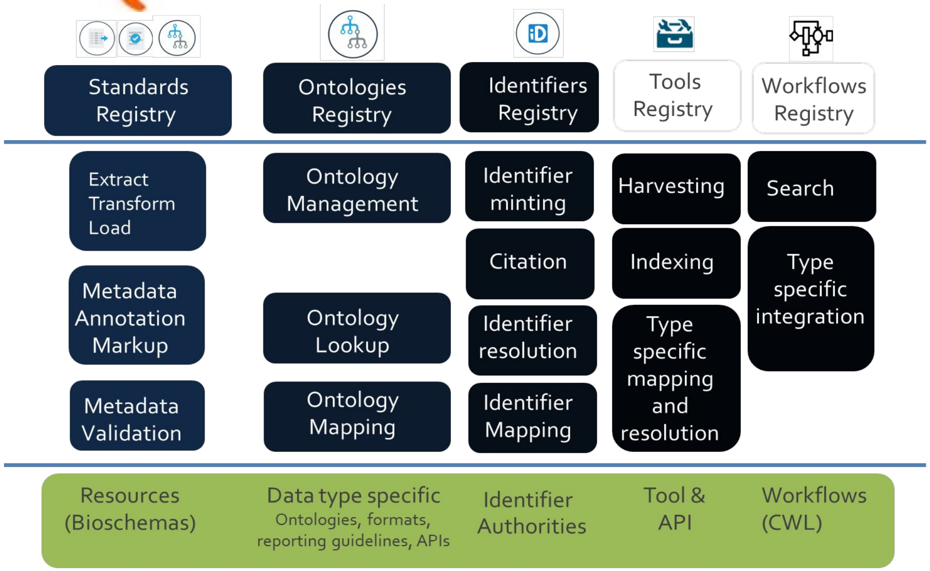
validated data-ontology maps

ontology-ontology mappings

federated authentication and access

KEY CO-OPERATIONS AND LEGACY IMPACT





Standards Registry
- Extract Transform Load
- Metadata Annotation Markup
- Metadata Validation

Ontologies Registry
- Ontology Management
- Ontology Lookup
- Ontology Mapping

Identifiers Registry
- Identifier minting
- Citation
- Identifier resolution
- Identifier Mapping

Tools Registry
- Harvesting
- Indexing
- Type specific mapping and resolution

Workflows Registry
- Search
- Type specific integration

Resources (Bioschemas) | Data type specific Ontologies, formats, reporting guidelines, APIs | Identifier Authorities | Tool & API | Workflows (CWL)

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human

European Commission

Final Report and Action Plan from the European Commission Expert Group on FAIR Data

TURNING FAIR INTO REALITY

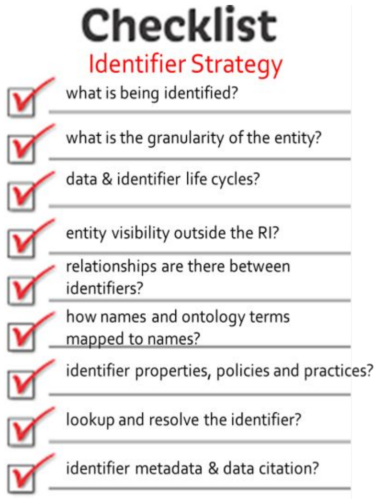
2018

Research and Innovation

WP6 ACTIVITIES

Year 1

Identifier best practice and checklists



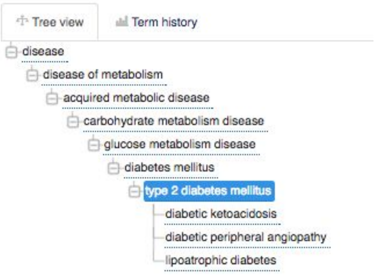
Checklist

Identifier Strategy

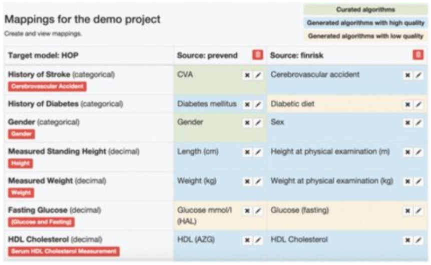
- ☑ what is being identified?
- ☑ what is the granularity of the entity?
- ☑ data & identifier life cycles?
- ☑ entity visibility outside the RI?
- ☑ relationships are there between identifiers?
- ☑ how names and ontology terms mapped to names?
- ☑ identifier properties, policies and practices?
- ☑ lookup and resolve the identifier?
- ☑ identifier metadata & data citation?

Year 2

Semantic infrastructure and ontologies



- disease
  - disease of metabolism
    - acquired metabolic disease
      - carbohydrate metabolism disease
        - glucose metabolism disease
          - diabetes mellitus
            - type 2 diabetes mellitus
              - diabetic ketoacidosis
              - diabetic peripheral angiopathy
              - lipoatrophic diabetes



Mappings for the demo project

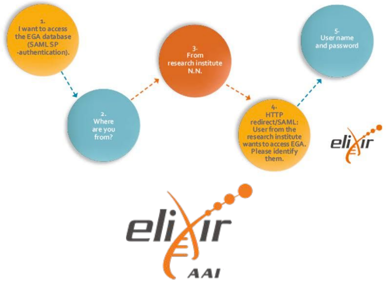
Create and view mappings.

- 🟩 Curated algorithms
- 🟦 Generated algorithms with high quality
- 🟨 Generated algorithms with low quality

Target model: HOP	Source: prevend	Source: finrisk
History of Stroke (categorical) Cerebrovascular Accident	CVA	Cerebrovascular accident
History of Diabetes (categorical)	Diabetes mellitus	Diabetic diet
Gender (categorical) Gender	Gender	Sex
Measured Standing Height (decimal) Height	Length (cm)	Height at physical examination (m)
Measured Weight (decimal) Weight	Weight (kg)	Weight at physical examination (kg)
Fasting Glucose (decimal) (Glucose and Fasting)	Glucose mmol/l (HAL)	Glucose (fasting)
HDL Cholesterol (decimal) Serum HDL Cholesterol Measurement	HDL (AZG)	HDL Cholesterol

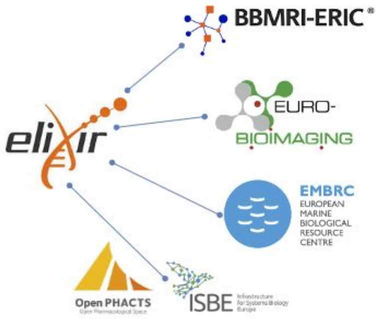
Year 3

Secure data access technology


1.
I want to access the EGA database (SAML SP -authentication).
2.
Where are you from?
3.
From research institute N.N.
4.
HTTP redirect/SAML: User from the research institute wants to access EGA. Please identify them.
5.
User name and password
elixir
elixir AAI

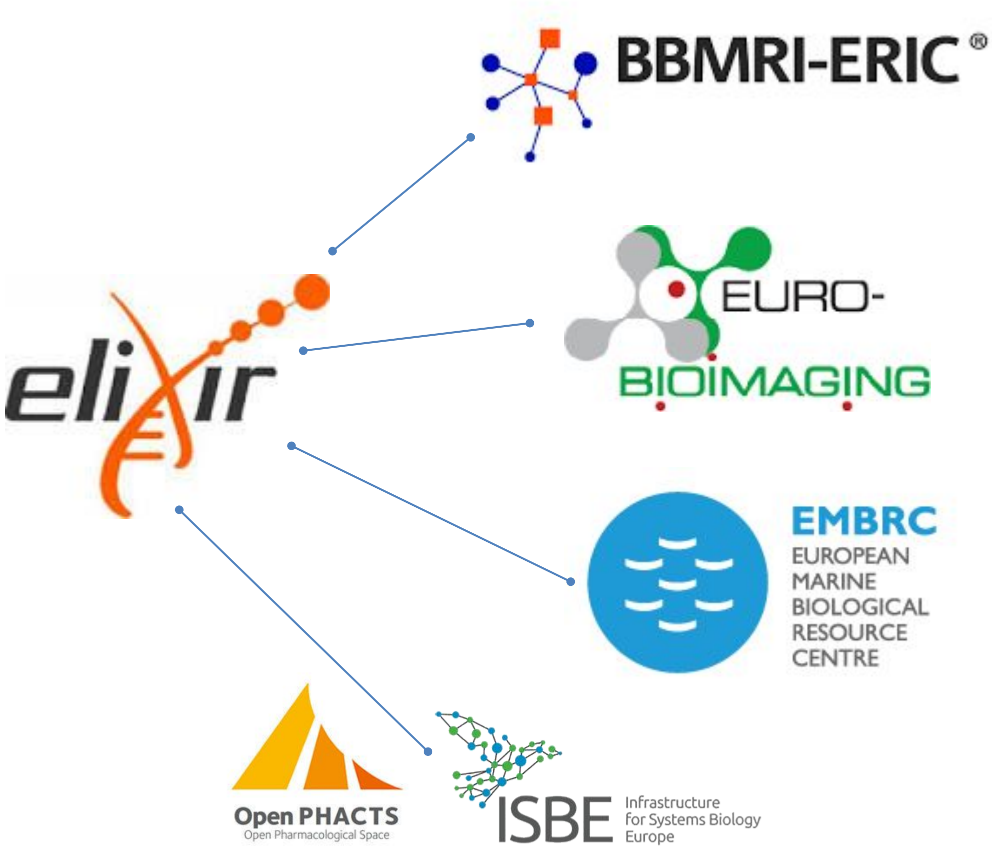
Year 4

Pan infrastructure deployment / enhancement -> EOSC Life



TASK 6.1 IDENTIFIER BEST PRACTICE, CHECKLISTS AND INFRASTRUCTURE

A **checklist based framework** for systematic documentation, gap analysis, recommendations and actions developed through **6 case studies**


elixir
BBMRI-ERIC®
EURO-BIOIMAGING
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
Open PHACTS
Open Pharmacological Space
ISBE
Infrastructure for Systems Biology Europe

1 Rare Disease

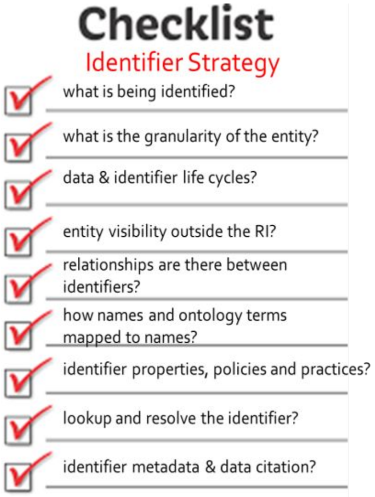
2 Biobanking

3 Imaging data

4 Marine metazoan models

5 Ocean sampling

6 Genes, proteins and drugs



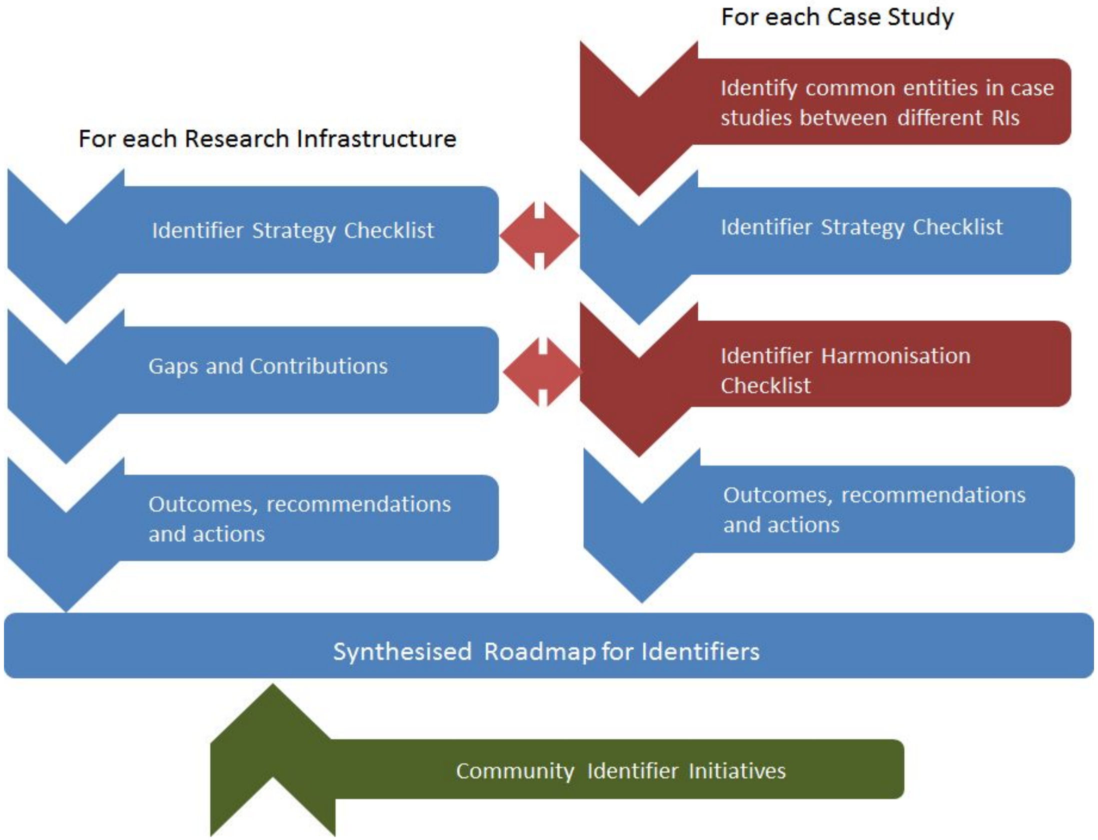
Checklist

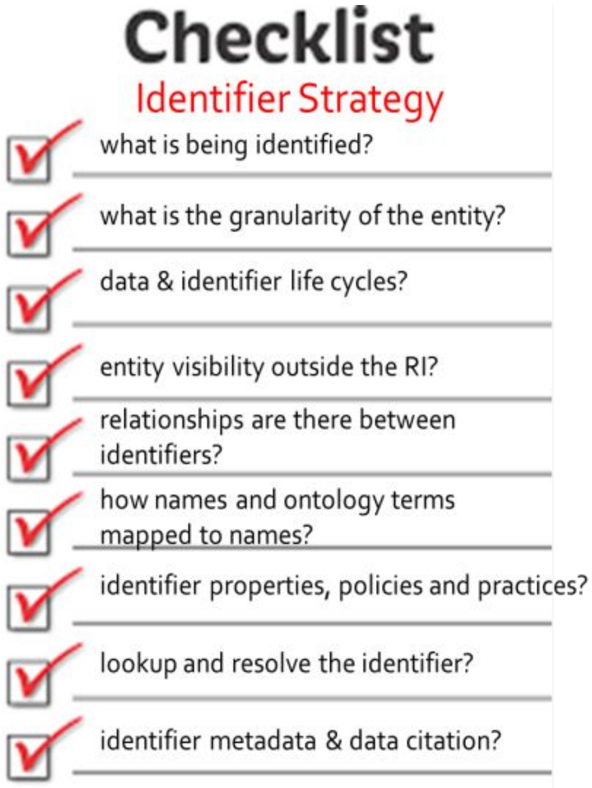
Identifier Strategy

- ☑ what is being identified?
- ☑ what is the granularity of the entity?
- ☑ data & identifier life cycles?
- ☑ entity visibility outside the RI?
- ☑ relationships are there between identifiers?
- ☑ how names and ontology terms mapped to names?
- ☑ identifier properties, policies and practices?
- ☑ lookup and resolve the identifier?
- ☑ identifier metadata & data citation?

TASK 6.1 IDENTIFIER BEST PRACTICE, CHECKLISTS AND INFRASTRUCTURE

A **checklist based framework** for systematic documentation, gap analysis, recommendations and actions


For each Research Infrastructure
Identifier Strategy Checklist
Gaps and Contributions
Outcomes, recommendations and actions
For each Case Study
Identify common entities in case studies between different RIs
Identifier Strategy Checklist
Identifier Harmonisation Checklist
Outcomes, recommendations and actions
Synthesised Roadmap for Identifiers
Community Identifier Initiatives



Checklist

Identifier Strategy

☑ what is being identified?

☑ what is the granularity of the entity?

☑ data & identifier life cycles?

☑ entity visibility outside the RI?

☑ relationships are there between identifiers?

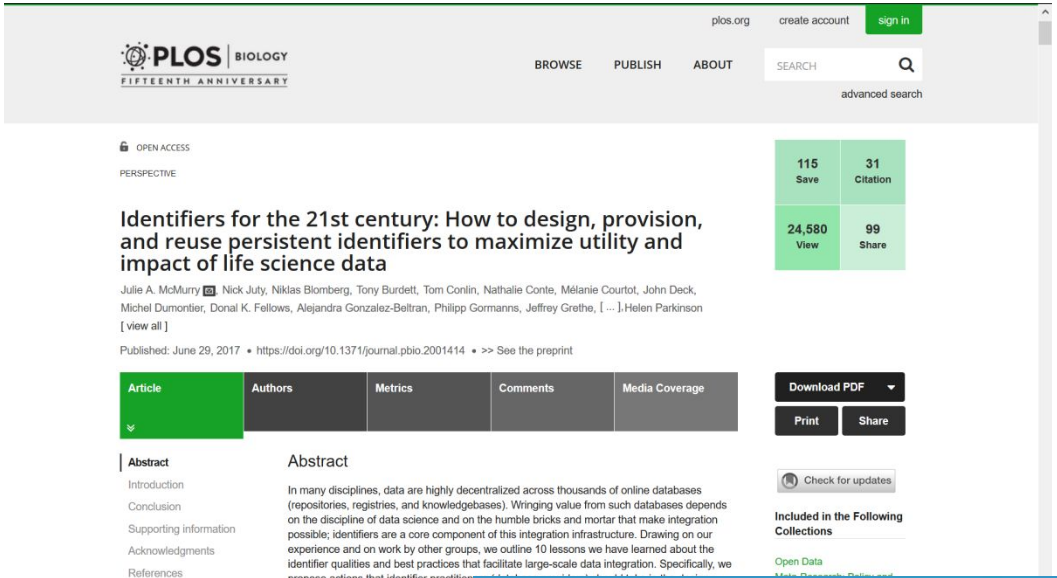
☑ how names and ontology terms mapped to names?

☑ identifier properties, policies and practices?

☑ lookup and resolve the identifier?

☑ identifier metadata & data citation?

TASK 6.1 IDENTIFIER BEST PRACTICE, CHECKLISTS AND INFRASTRUCTURE



plos.org   create account   sign in

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

BROWSE   PUBLISH   ABOUT   SEARCH
advanced search

OPEN ACCESS

PERSPECTIVE

Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data

Julie A. McMurry, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot, John Deck, Michel Dumontier, Donal K. Fellows, Alejandra Gonzalez-Beltran, Philipp Gormanns, Jeffrey Grethe, [...], Helen Parkinson
[view all]

Published: June 29, 2017 • https://doi.org/10.1371/journal.pbio.2001414 • >> See the preprint

115 Save	31 Citation
24,580 View	99 Share

Article   Authors   Metrics   Comments   Media Coverage

Download PDF   Print   Share

Abstract
Introduction
Conclusion
Supporting information
Acknowledgments
References

Reader Comments (2)
Media Coverage (1)
Figures

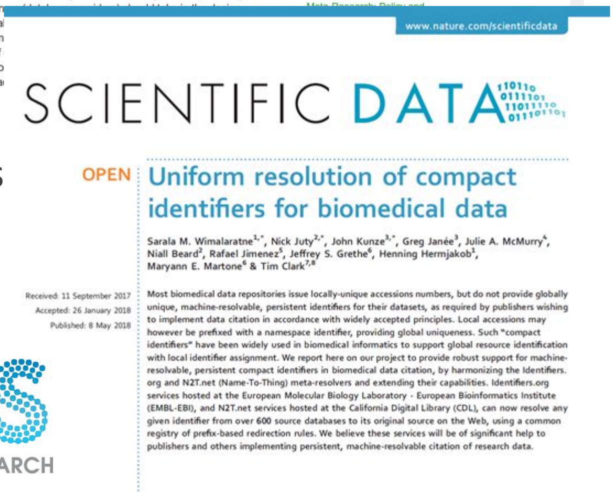
Abstract

In many disciplines, data are highly decentralized across thousands of online databases (repositories, registries, and knowledgebases). Wringing value from such databases depends on the discipline of data science and on the humble bricks and mortar that make integration possible; identifiers are a core component of this integration infrastructure. Drawing on our experience and on work by other groups, we outline 10 lessons we have learned about the identifier qualities and best practices that facilitate large-scale data integration. Specifically, we propose actions that identifier practition
provision and reuse of identifiers. We al
referencing identifiers in various circum
While the importance and relevance of
increased awareness about how to avo
those related to persistence and web-a

Check for updates

Included in the Following Collections

Open Data



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Uniform resolution of compact identifiers for biomedical data

Sarala M. Wimalaratne[1,*], Nick Juty[2,*], John Kunze[3,*], Greg Janée[3], Julie A. McMurry[4], Niall Beard[2], Rafael Jimenez[5], Jeffrey S. Grethe[6], Henning Hermjakob[1], Maryann E. Martone[6] & Tim Clark[7,8]

Received: 11 September 2017
Accepted: 26 January 2018
Published: 8 May 2018

Most biomedical data repositories issue locally-unique accessions numbers, but do not provide globally unique, machine-resolvable, persistent identifiers for their datasets, as required by publishers wishing to implement data citation in accordance with widely accepted principles. Local accessions may however be prefixed with a namespace identifier, providing global uniqueness. Such "compact identifiers" have been widely used in biomedical informatics to support global resource identification with local identifier assignment. We report here on our project to provide robust support for machine-resolvable, persistent compact identifiers in biomedical data citation, by harmonizing the Identifiers.org and N2T.net (Name-To-Thing) meta-resolvers and extending their capabilities. Identifiers.org services hosted at the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), and N2T.net services hosted at the California Digital Library (CDL), can now resolve any given identifier from over 600 source databases to its original source on the Web, using a common registry of prefix-based redirection rules. We believe these services will be of significant help to publishers and others implementing persistent, machine-resolvable citation of research data.

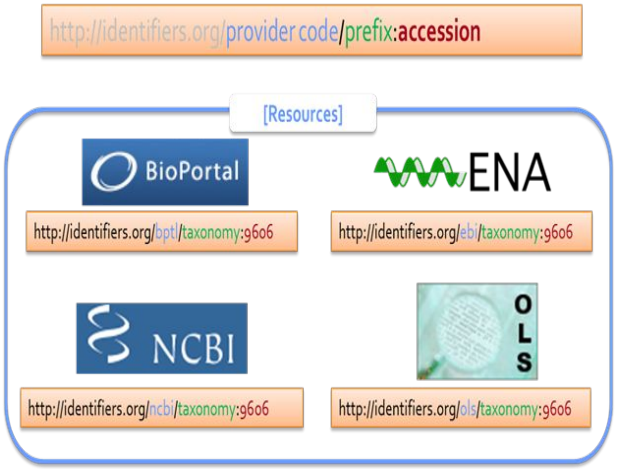
Aligned ontology ids and Identifier.org -> Ontology Lookup Service supports CURIEs

OLS ONTOLOGY SEARCH

FORCE11
The Future of Research Communications and e-Scholarship

identifiers.org

NIH
Big Data to Knowledge(BD2K)



http://identifiers.org/provider code/prefix:accession

[Resources]

BioPortal
http://identifiers.org/bptl/taxonomy:9606

ENA
http://identifiers.org/ebi/taxonomy:9606

NCBI
http://identifiers.org/ncbi/taxonomy:9606

OLS
http://identifiers.org/ols/taxonomy:9606

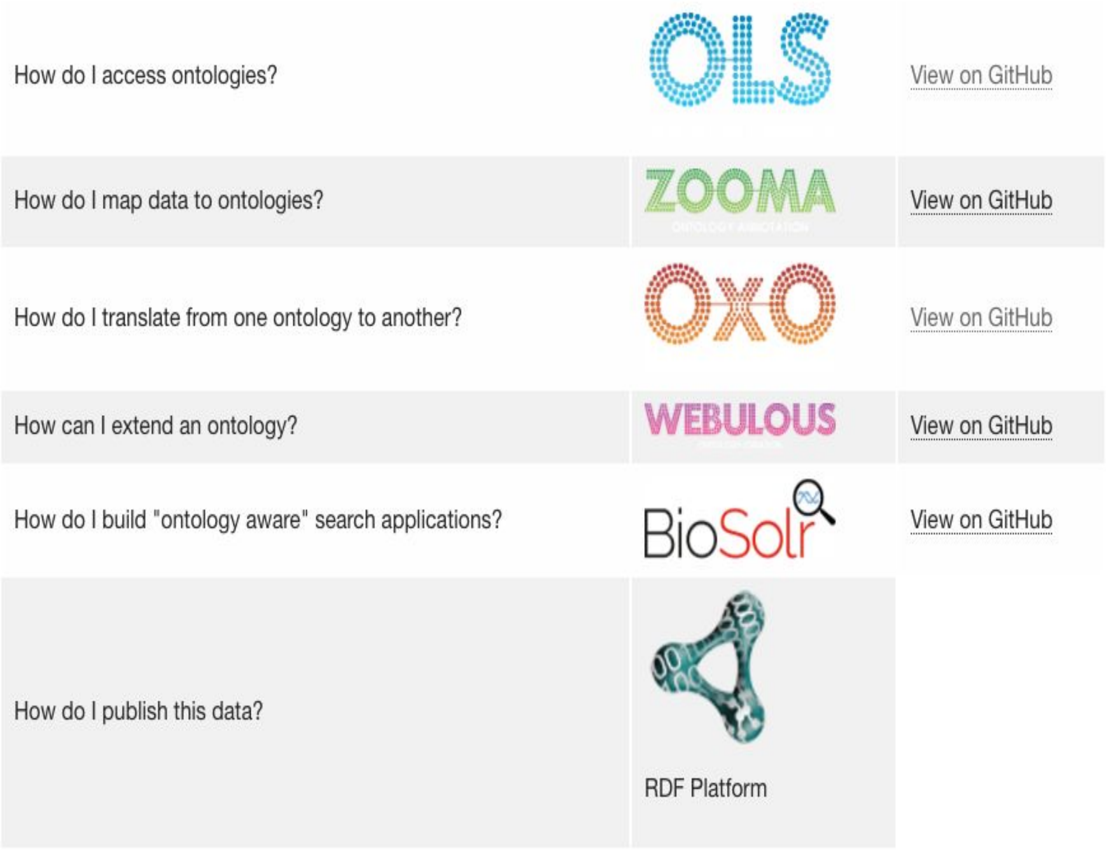
Provides unique stable, resolvable and location-independent compact URIs to identify and locate scientific data

N2T.net RESOLVER

Harmonisation with USA CDL name2thing (n2t.net) resolution service

TASK 6.2 A SEMANTIC TOOLKIT

- Standards compliant web and programmatic based access to ontologies and linked open data from an ontology access service
- Validated data-ontology maps with provenance between data and ontologies
- Ontology-ontology mappings with provenance supporting data integration across infrastructures
- Semantic infrastructure



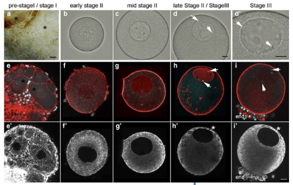
How do I access ontologies?	OLS	View on GitHub
How do I map data to ontologies?	ZOOMA	View on GitHub
How do I translate from one ontology to another?	OXO	View on GitHub
How can I extend an ontology?	WEBULOUS	View on GitHub
How do I build "ontology aware" search applications?	BioSolr	View on GitHub
How do I publish this data?	RDF Platform	

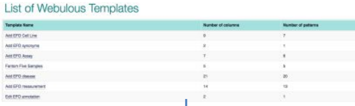
https://ebispot.github.io

TASK 6.2 A SEMANTIC TOOLKIT

Build/extend an ontology

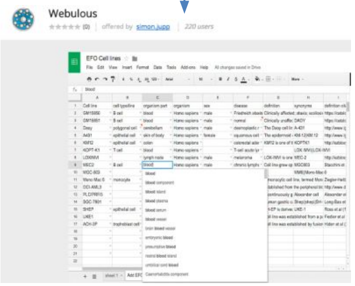
Webulous





List of Webulous Templates

Template Name	Number of columns	Number of patterns
Add EFO Cell Line	9	7
Add EFO synonyms	2	1
Add EFO Assay	7	6
Fantom Five Samples	5	5
Add EFO disease	21	20
Add EFO measurement	14	13
Edit EFO annotation	2	1



Webulous

☆☆☆☆☆ (0) | offered by simon.jupp | 220 users

EFO Cell lines

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

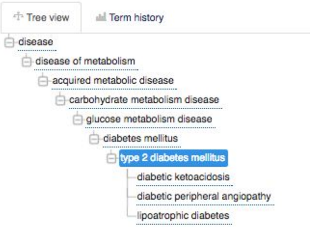
fx blood

	A	B	C	D	E	F	G	H	I
1	Cell line	cell type/line	organism part	organism	sex	disease	definition	synonyms	definition cit
2	GM15850	B cell	blood	Homo sapiens	male	Friedreich ataxia	Clinically affected; ataxia; scoliosis		https://catalo
3	GM15851	B cell	blood	Homo sapiens	male	normal	Clinically unaffec	DADY	https://catalo
4	Daoy	polygonal cell	cerebellum	Homo sapiens	male	desmoplastic c	The Daoy cell lin	A-431	http://www.
5	A431	epithelial cell	skin of body	Homo sapiens	female	squamous cell	The epidermoid	KM-12|KM 12	http://www.
6	KM12	epithelial cell	colon	Homo sapiens		colorectal ade	KM12 is one of t	KOPTK1	http://cellbar
7	KOPT-K1	T cell	blood	Homo sapiens		T-cell acute ly		LOX IMVI|LOX-IMVI	
8	LOXIMVI		lymph node	Homo sapiens	male	melanoma	LOX-IMVI is one	MEC-2	http://cellbar
9	MEC2	B cell	blood	Homo sapiens	male	chronic lympho	Cell line grew up	MGC803	Stacchini et
10	MGC-803							MM6|Mono Mac 6	
11	Mono Mac 6	monocyte					monocytic cell line, termed Mon		Ziegler-Heit
12	OCI-AML3						stablished from the peripheral bl		http://www.d
13	PLC/PRF/5						continuously g	Alexander cell	Alexander et
14	SGC-7901						man gastric c	Shep|shep|SH-	Long-Bao et
15	SHEP	epithelial cell					-EP is derive	UKE-1	Ross et al (1
16	UKE1						ll line was established from a p		Fiedler et al
17	ACH-2P	trophoblast cell					ll line was established by fusion		Hiden et al (
18									
19									
20									
21									
22									

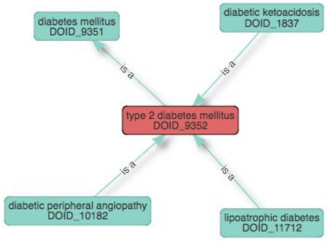
blood
blood component
blood island
blood plasma
blood serum
blood vessel
brain blood vessel
embryonic blood
presumptive blood
rostral blood island
umbilical cord blood
Caenorhabditis component

Access and visualise an ontology

Ontology Lookup Service



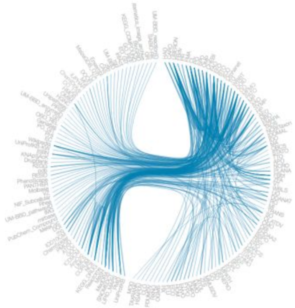
- disease
  - disease of metabolism
    - acquired metabolic disease
      - carbohydrate metabolism disease
        - glucose metabolism disease
          - diabetes mellitus
            - type 2 diabetes mellitus
              - diabetic ketoacidosis
              - diabetic peripheral angiopathy
              - lipoatrophic diabetes

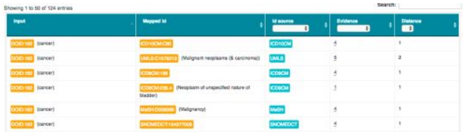


diabetes mellitus DOID_9351
diabetic ketoacidosis DOID_1837
is a
is a
type 2 diabetes mellitus DOID_9352
is a
is a
diabetic peripheral angiopathy DOID_10182
lipoatrophic diabetes DOID_11712

Map between ontologies 'xref'

Ontology Xref Service





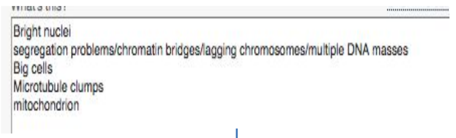
Showing 1 to 50 of 124 entries

Search:

Input	Mapped Id	Id source	Evidence	Distance
DOID:162 (cancer)	ICD10CM:C80	ICD10CM	4	1
DOID:162 (cancer)	UMLS:C1378213 (Malignant neoplasms (& carcinoma))	UMLS	5	2
DOID:162 (cancer)	ICD9CM:199	ICD9CM	4	1
DOID:162 (cancer)	ICD9CM:239.4 (Neoplasm of unspecified nature of bladder)	ICD9CM	1	1
DOID:162 (cancer)	MeSH:D009369 (Malignancy)	MeSH	4	1
DOID:162 (cancer)	SNOMEDCT:154677006	SNOMEDCT	4	1

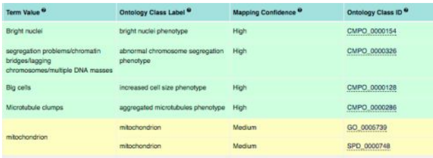
Map text to ontologies

ZOOMA



What's this?

Bright nuclei
segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses
Big cells
Microtubule clumps
mitochondrion



Term Value	Ontology Class Label	Mapping Confidence	Ontology Class ID
Bright nuclei	bright nuclei phenotype	High	CMPO_0000154
segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses	abnormal chromosome segregation phenotype	High	CMPO_0000326
Big cells	increased cell size phenotype	High	CMPO_0000128
Microtubule clumps	aggregated microtubules phenotype	High	CMPO_0000286
mitochondrion	mitochondrion	Medium	GO_0005739
	mitochondrion	Medium	SPD_0000748

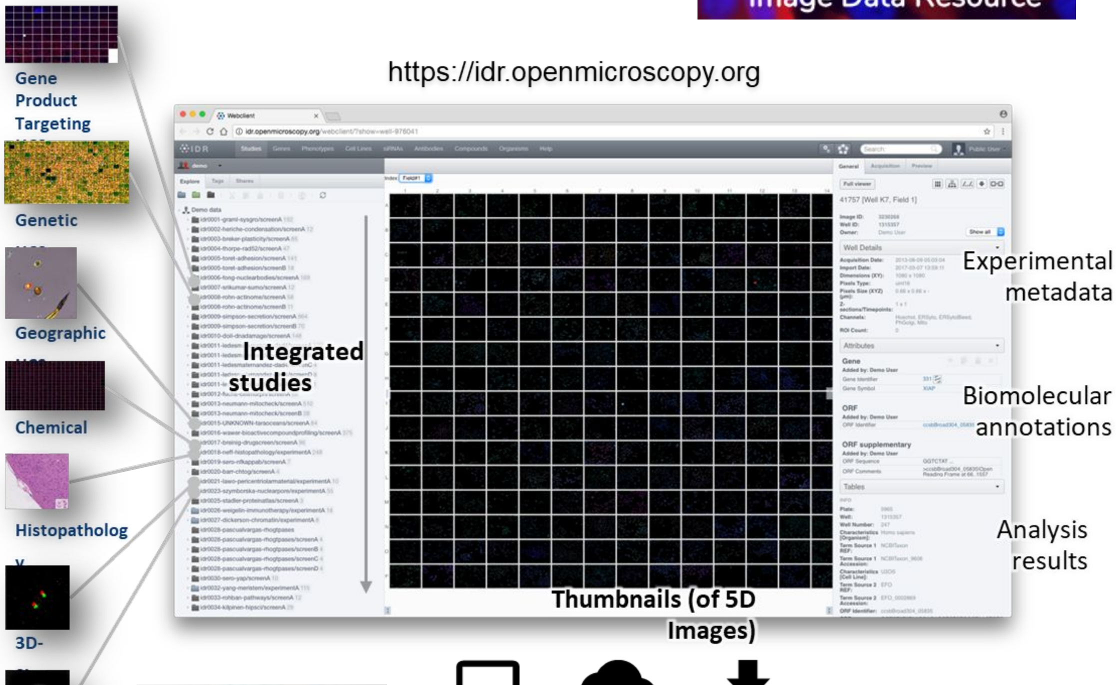
RI DEPLOYMENT IDENTIFIERS

"Euro-BioImaging has used identifiers for defining genes, proteins, antibodies, drugs, species, phenotypes, and organ systems and pathologies to publish ~170 TBytes of original image datasets

Identifiers are key for IDR to fulfil its function as an added value knowledgebase, making critical reference datasets well-annotated and ultimately linked, searchable and reusable.
Everything IDR has done has followed guidance and used tools developed within or related to CORBEL."

IDR Image Data Resource

https://idr.openmicroscopy.org



Jason Swedlow

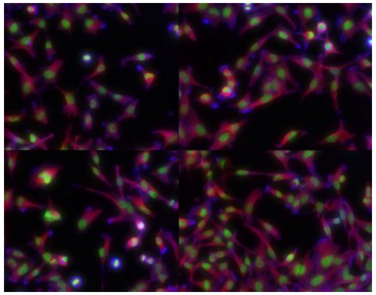
RI DEPLOYMENT

IMPROVED METADATA

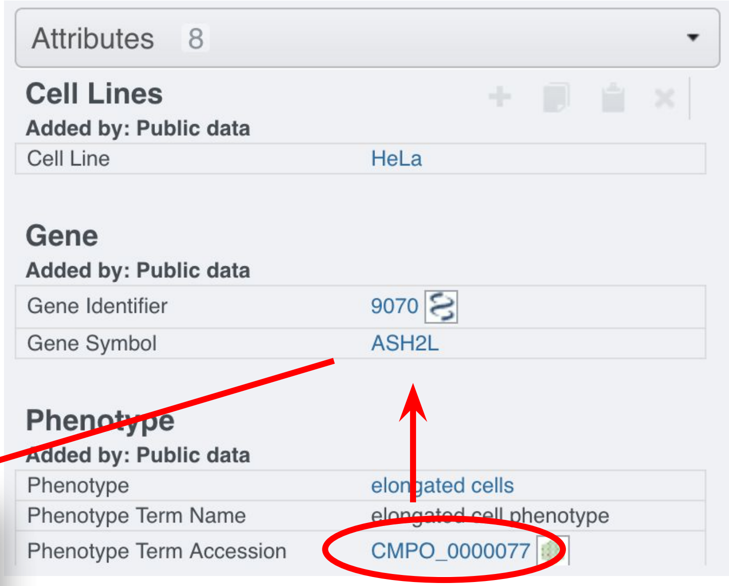
Improving the handling of image annotations

Using the Cellular Microscopy Phenotype Ontology and the OLS can group attributes

Now can link between studies through the ontology



idr0012



Attributes 8

Cell Lines
Added by: Public data

Cell Line	HeLa

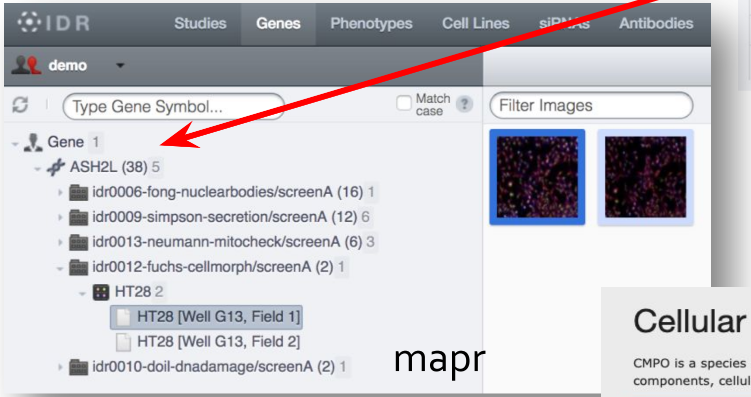
Gene
Added by: Public data

Gene Identifier	9070
Gene Symbol	ASH2L

Phenotype
Added by: Public data

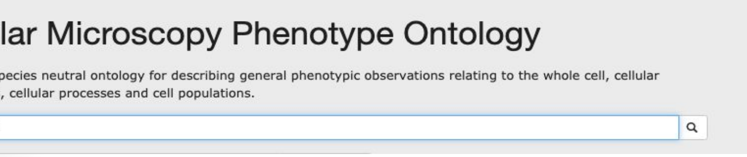
Phenotype	elongated cells
Phenotype Term Name	elongated cell phenotype
Phenotype Term Accession	CMPO_0000077

ONTOLOGY SEARCH



- Gene 1
  - ASH2L (38) 5
    - idr0006-fong-nuclearbodies/screenA (16) 1
    - idr0009-simpson-secretion/screenA (12) 6
    - idr0013-neumann-mitocheck/screenA (6) 3
    - idr0012-fuchs-cellmorph/screenA (2) 1
      - HT28 2
        - HT28 [Well G13, Field 1]
        - HT28 [Well G13, Field 2]
    - idr0010-doil-dnadamage/screenA (2) 1

mapr



Cellular Microscopy Phenotype Ontology

CMPO is a species neutral ontology for describing general phenotypic observations relating to the whole cell, cellular components, cellular processes and cell populations.

CORBEL

TASK 6.3 SECURE ACCESS TO SENSITIVE DATA

Federated authentication and access solution to data service provides selected by the project who support BMS RI data management, analyses, deposition and distribution

Key components

- Authentication and authorisation infrastructure (AAI)
- Secure data streaming
- Metadata standardisation and synchronisation
- Policy components linking to country level/Infrastructure level best practice

elixir

Global Alliance
for Genomics & Health

Improve interoperability with European e-infrastructures and leverage existing investments in these capacities within the biomedical and life science domain

CORBEL driver projects

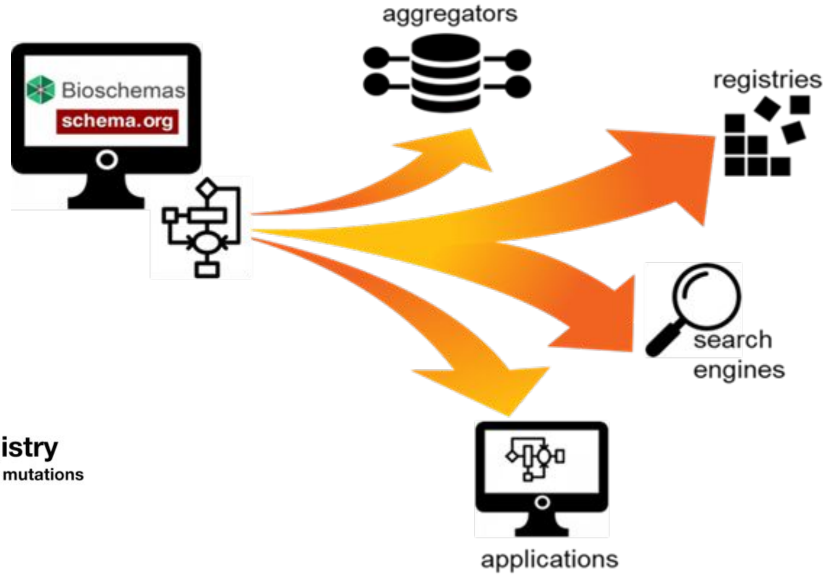
- Delegated access to digitalized biobank samples
- GoNL - Federated AAI + secure streaming
- BBMRI-NL - Bioschemas and Beacons
- BBMRI-ERIC/RD-connect - harmonization and ‘matchmaking’ service

BBMRI
Biobanking and
Biomolecular
Resources Research
Infrastructure

BIOSCHEMAS
METADATA MARK-UP AND PATIENT REGISTRIES

molgenis

Hosts many databases, amongst which many patient and mutation registries.



deb register

the international database of dystrophic epidermolysis bullosa patients and *COL7A1* mutations

International Microvillus Inclusion Disease (MVID) Patient Registry

International registry of patients with microvillus inclusion disease and database of associated gene mutations

AIP Mutation Database

Added structured machine readable metadata descriptions to multiple patient registries using BioSchemas.

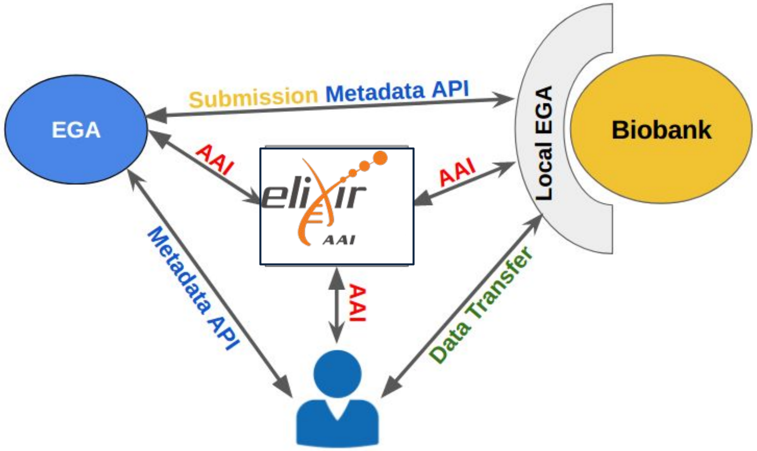
CHD7 Database

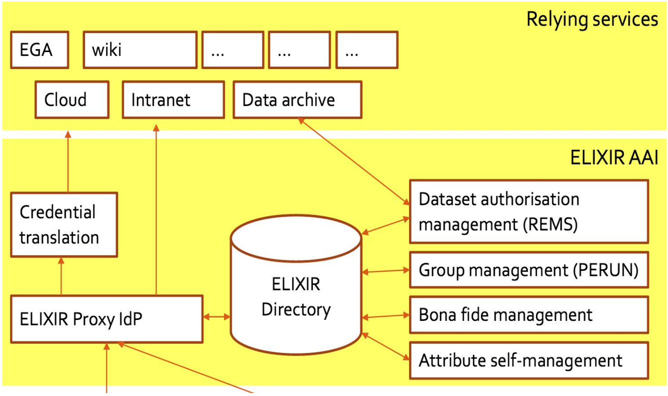
Increase the findability of multiple patient and mutation registries, making it easier to find and reuse critical information for rare disease patient care and research.

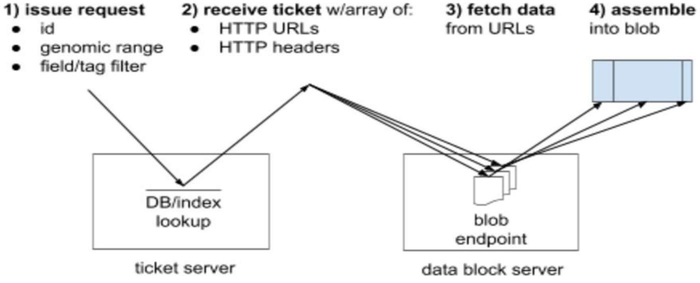
https://f1000research.com/posters/7-1228

TASK 6.3 SECURE ACCESS TO SENSITIVE DATA

- User authentication and authorisation
  - Align with ELIXIR AAI
- Dataset authorisation
  - REMS: Electronic tool for the management of access rights to controlled access research data
  - Beacon: automated process to all data in "registered" data access layer
- Secure data delivery
  - Align with GA4GH
  - htsget: remote streaming protocol







htsget (http://samtools.github.io/hts-specs/htsget.html)

TASK 6.3 DATA MATCHMAKING & HARMONIZATION TOOLKIT FOR POOLED ANALYSIS OF SENSITIVE DATA

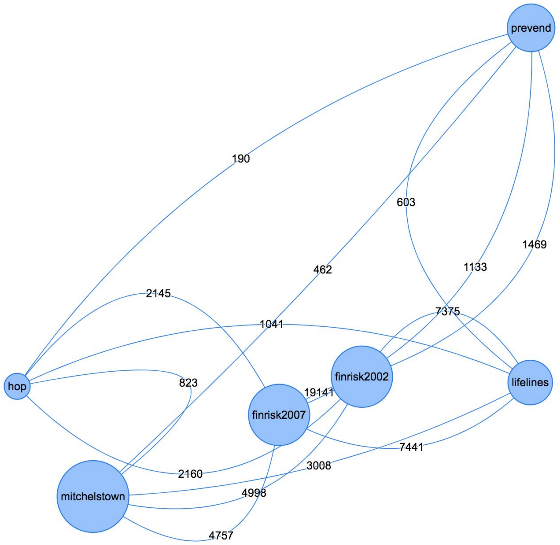
Heterogeneous data sharing across 20+ child cohorts - the LifeCycle Project

Find data collections matching your needs

new

BiobankUniverse

Find 'similar' collections based on data item metadata only, using ontology based lexical and semantic annotation and matching (6.2)

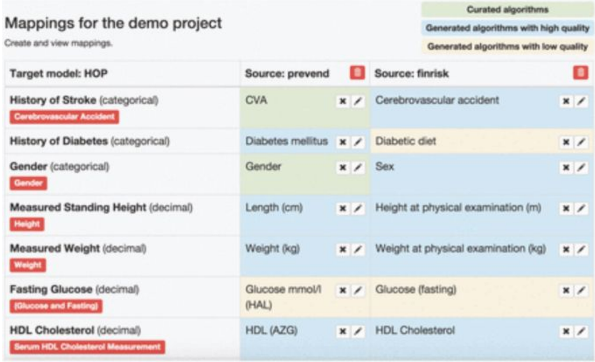


Match data items to research parameters and create algorithms

FAIRifier

BiobankConnect

Shortlist attributes matching research needs and auto-generate ETL algorithms; provide sharing of harmonization rules



Mappings for the demo project

Create and view mappings.

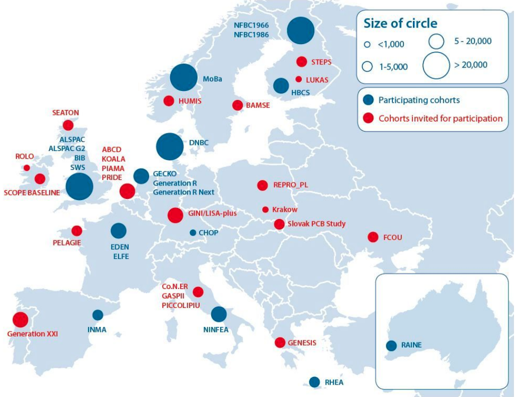
Curated algorithms
Generated algorithms with high quality
Generated algorithms with low quality

Target model: HOP	Source: prevend	Source: finrisk
History of Stroke (categorical) Cerebrovascular Accident	CVA	Cerebrovascular accident
History of Diabetes (categorical)	Diabetes mellitus	Diabetic diet
Gender (categorical) Gender	Gender	Sex
Measured Standing Height (decimal) Height	Length (cm)	Height at physical examination (m)
Measured Weight (decimal) Weight	Weight (kg)	Weight at physical examination (kg)
Fasting Glucose (decimal) (Glucose and Fasting)	Glucose mmol/l (HAL)	Glucose (fasting)
HDL Cholesterol (decimal) Serum HDL Cholesterol Measurement	HDL (AZG)	HDL Cholesterol

Apply harmonization algorithms and then automated (meta) analysis

Secure Digital Research Environment

Apply ETL algorithms on the data in secure analysis environment following DAC, e.g. analysis of 250.000 children (LifeCycle)



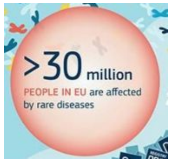
All available as open source http://molgenis.githubio. BiobankUniverse: manuscript in prep; beta@ http://biobankuniverse.org. BiobankConnect:PMID:27153686 ; demo @ http://biobankconnect.org .

TASK 6.3 SECURE STANDARDS & INFRASTRUCTURE

Virtual HPC cloned across sites

Deliver reproducibility

WHAT



Rare disease projects

GDPR compliance use cases

Secure Federated analysis solution

WHY

Openstack API

Ansible installer
Plug and play workflow

Embassy cloud deployment – proximity to EGA

HOW

Federated analysis enabled

Close to data archives at EBI, UMCG, NIKHEF

19K+165k individual's data

Broaden access to tools and compute

IMPACT

"We developed a Bioinformaticians Sandbox - a fully virtual HPC cluster for bioinformaticians which can be automatically and reproducibly cloned on different sites, ensuring reproducibility when analyses are done on multiple sites. This was motivated by a large multi-center study to have a controlled data access site with analysis capability that is GDPR-Compliant"

Morris Swertz

CONTINUITY FOR EOSC LIFE

Identifier best practice and checklists

Semantic infrastructure and ontologies

Secure data access technology

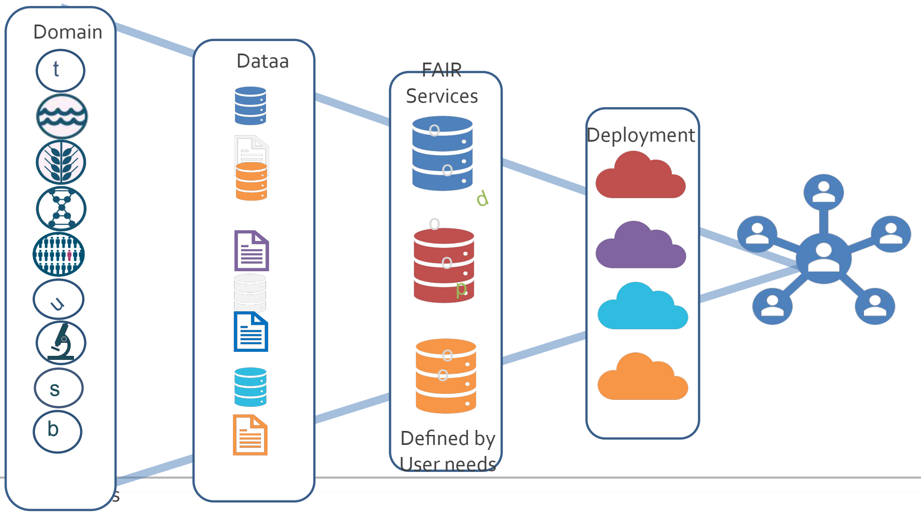
Pan infrastructure deployment / enhancement

Standards driving FAIR tool suite

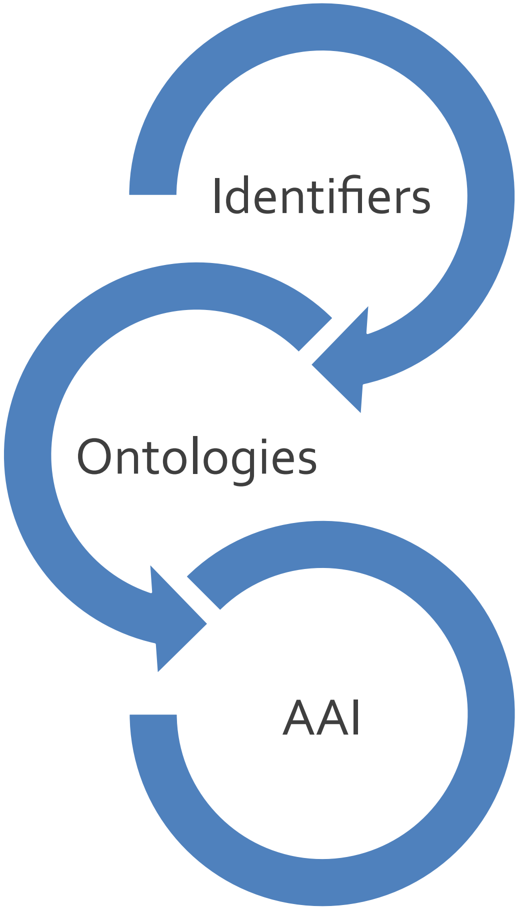
FAIR Tool suite in the cloud

Enabling AI on controlled access data

Deployment via Community Funding Calls



CONCLUSIONS AND FOLLOW-ONS



Within and cross BMS RI – open projects, standards and implementations, and better understanding.

ELIXIR, BBMRI, EuroBioimaging, EMBRC, ISBE

Broader utility than implementations within CORBEL – now seeing use in many projects

Many components now being taken forward to EOSC Life where they will form tech stack for the next phase of projects

Getting infrastructure out and across BMS RI a challenge.

Benefited from cooperation with ELIXIR Excelerate