Coordinated Research Infrastructures Building Enduring Life-science services
- CORBEL -

Deliverable D6.1
Review of identifier schemes, standards and interoperability maps and proposed harmonization strategy

WP6 – Data access, management and integration
Lead Beneficiary: EMBL-EBI
WP leader: Carole Goble (UNIMAN), Helen Parkinson (EMBL-EBI)
Contributing partner(s): EMBL-EBI, UNIMAN, UMCG, JacobsUni, Lygature, JacobsUni, UNIVDUN.

Contractual delivery date: 28 February 2017
Actual delivery date: 8 March 2017

Authors of this deliverable: Carole Goble, Helen Parkinson, Simon Jupp, Renzo Kottmann, Jan-Willem Boiten, Frank Oliver Glöckner, Eleanor Williams, Jason Swedlow, Morris Swertz, David van Enckevort, Nick Juty, Anna Gaulton

Contributors to this deliverable: Rafael C. Jimenez, Norman Morrison (ELIXIR-HUB), Alasdair Gray (Heriot-Watt University), Egon Willighagen, Chris Evelo, Friederike Ehrhart (Uni Maastricht), Ian Dunlop, Anna Leida Molder, Kristian Garza (UNIMAN), Marco Roos (LUMC)

Grant agreement no. 654248
Horizon 2020
H2020-INFRADEV-1-2014
Type of action: RIA

Content

Executive Summary

This work addresses identifier recommendations and harmonisation in the context of the BMS Research Infrastructures (RIs) participating in CORBEL. We summarise community activities in this space (Appendix 1), provide two checklists for: (i) RI identifier strategy and (ii) identifier harmonisation of common entities and services for identity use cases which span the participating BMS RIs. We use six case studies to guide the work. We also define priorities and document existing services that can be adopted, or extended in support of CORBEL.

Project objectives

Within the scope of the CORBEL Project, WP6, this report has contributed to the following objectives:

- Completing an up-to-date investigation into current standards and systems used (Objective 1).
- Visualising the landscape of current standards and systems to all collaborative partners (Objective 1).
- Delivering a harmonisation document based on the current state-of-the-art (objective 2).
- Delivering a set of best-practice documentation, based on the current state-of-the-art (Objective 1 & 2).

Detailed report on the deliverable

This deliverable addresses identifiers within the BMS RIs participating in CORBEL. It builds on previous work that documents requirements for design, provision and reuse of identifiers, and now documents available services and provides materials for CORBEL partners to design, implement and evaluate an identifier framework.

This is crystallised in six case studies reported here addressing:

- Case Study 1 - rare disease (BBMRI and ELIXIR),
- Case Study 2 - biobanking (BBMRI, ELIXIR)
- Case Study 3 - imaging data (Euro-BioImaging and ELIXIR),
- Case Study 4 - marine metazoan models (EMBRC and ELIXIR),
- Case Study 5 - ocean sampling (EMBRC and ELIXIR)
- Case Study 6 - genes, proteins and drugs (ELIXIR and ISBE).

By identifying diverse use cases we are able to thoroughly evaluate our checklist approach and iteratively improve the harmonisation strategy as CORBEL progresses. In focussing on CORBEL and its participating BMS RIs we address the different entities (e.g. genes, proteins, compounds, biological samples, individuals, cohorts) that need to be identified within and between infrastructures, these are diverse and range in granularity from identification of genetic variants in an individual to

identification of populations. Our approach is therefore to first examine entities in common across CORBEL as well as the necessary infrastructure to manage identifiers, then to apply our checklist and to iterate over the checklist as new use cases and services are identified and supported.
Identifier management for life sciences covers two chief areas:

1. **Identifier Policies and Schemes**
    - What to identify: entities, granularities, entity life cycles, identifier life cycles.
    - When to identify: the lifetime of identifiers, identifier creation and assignment, transitioning identifiers from internal to external, temporary to persistent.
    - How to identify it: identifier properties, standard schemes for identifiers, standardised metadata associated with identifiers, standard identifiers for entities.
    - Who oversees identifiers: governance, policies and identifier authorities.

2. **Identifier Infrastructure**
    - Identifier services for the execution of the policies and schemes: Creation, Conversion, Resolution, Mapping, Catalogues, Aggregation

We do not go into details of these policies, schemes and infrastructure as these are discussed elsewhere and in our previous work (Appendix 2 provides a list derived from the BioMedBridges project with suitable updates). Here we summarise the key points and organise these into checklists for use in CORBEL. In each use case we will examine the existing infrastructure framework and supporting policies for identifier management and will report on the harmonisation of these. Our basic checklist is supplied for CORBEL partners as part of this deliverable and this will be extended in a use case specific way via engagement with the partners in delivery of WP3 and WP4.

Our chief challenges in CORBEL are:

Entity-based authorities

- Ensure relevant entities are clearly identified by a designated authority(ies) using permanent and persistent identifiers
- Determine where no designated authority(ies) exist and identify/delegate one
- Ensure that all authorities are registered in suitable registries and the ELIXIR registry identifiers.org

Services

- Determine whether the existing services are fit for purpose
- Identify where gaps exist in service provision and recommend how to fill these
- Ensure that all identifier services within CORBEL are registered in the ELIXIR bio.tools service registry and other relevant registries (e.g. Euro-BioImaging)

Practices with RIs

- Support the FAIR principles with (Findable, Accessible, Interoperable and Reusable) identifiers best practice and communicate this, for example on the ELIXIR knowledge Hub and with members of each BMS RI.

Harmonisation on common entities between RIs

- Determine which entities overlap between use cases in CORBEL and are suitable for harmonisation via use cases
- Determine which services overlap, can be transferred or extended across RIs
- Determine the scope of our activities illustrated by use cases

We also report on relevant work in the wider international community. This deliverable draws upon work previously undertaken on identifiers in Life Sciences and their management services. Our approach is not to replicate this work but to use it drive our methodology and inform our roadmap and we are engaged with relevant efforts in this domain.

Framework

We have implemented a checklist approach to drive systematic documentation, gap analysis, recommendations and actions (Figure 1). Our roadmap is the application of these checklists to CORBEL case studies, and the assembly of an action list for CORBEL. The questions and issues comprising the checklists are summarised in Tables 1-9.
The work reported here operated in **three parallel threads** to prepare for the framework:

1. Review prior work and create comprehensive checklists for the RIs and Use Cases (Tables 1-9)
2. Review six concrete case studies using multiple datasets each linked to two RIs
3. Review of community activities

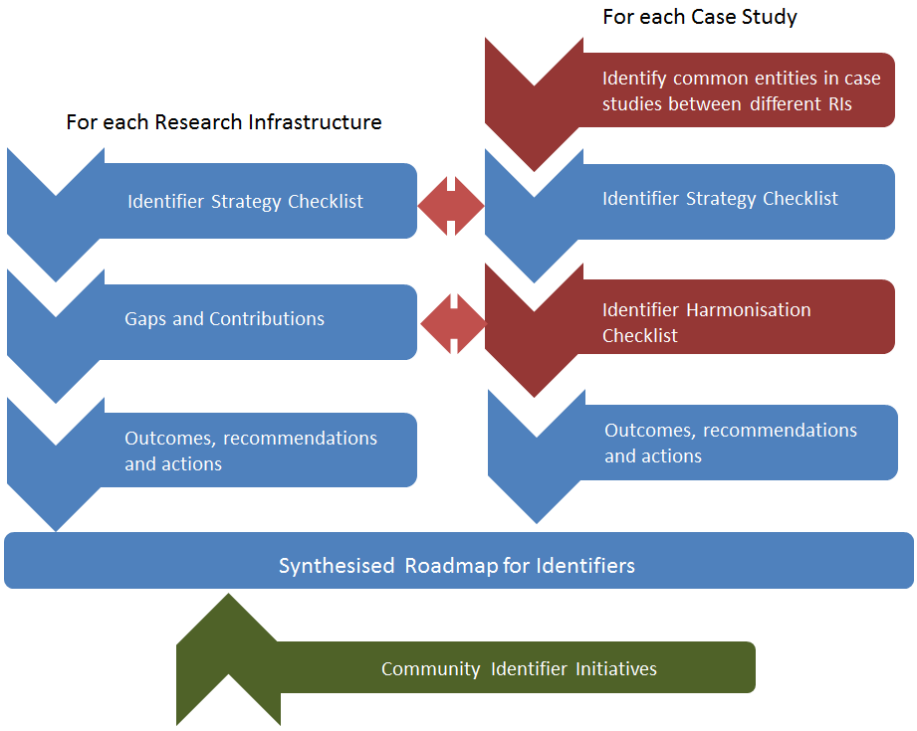

Figure 1. Identifier Strategy and Harmonisation Framework

Two checklists are intended to work together: the Identifier Strategy checklist and the Identifier Harmonisation Strategy checklist (see Tables 1-9).

Identifier Strategy Checklist

This is a systematic approach for:

- Each Research Infrastructure to examine (i) the entities they are identifying (ii) the identifier practices and services currently in use with those entities.
- Each case study crossing Research Infrastructures to examine (i) the entities they have in common (ii) the identifier practices and services currently in use with those entities (iii) to compare the practices and services.

The checklist is synthesis of recommendations and highlights proposed in prior work and informed by international identifier working groups:

- BioMedBridges deliverable Identifier Best Practice and Supporting Tools 2014
- Julie McMurry, et al (2016). Identifiers for the 21st century: How to design, provision, and reuse identifiers to maximize data utility and impact. http://doi.org/10.5281/zenodo.163459
- Martin Fenner, et al (2016). A Data Citation Roadmap for Scholarly Data Repositories doi: https://doi.org/10.1101/097196

Identifier Harmonisation Strategy Checklist

This is a systematic approach for reviewing the Identifier strategy used by different RIs for Common Data and identifying the harmonisation points that are necessary or desirable. The approach is founded on four principles:

1. *Bootstrapping on demand:* what is the least possible that be done to achieve harmonisation given limited resources. Harmonisation should be only in the face of a specific need and follow the "just in time" paradigm (also known as "pay as you go") whereby harmonisation actions are bootstrapped on a demand basis. This is also known as "just enough, just in time not just in case".
2. *Minimize disruption:* what is the least disruption we can cause to legacy systems. CORBEL is not operating in a "green field". All the resources are pre-existing and incur costs for maintenance and development. Some disruption is inevitable, particularly in the light of outcomes of the Identifier Strategy analysis, we therefore strive to address real problems and to minimise disruption.
3. *Common Data view:* we focus entirely on common data types and exclude harmonisation efforts on data types that are exclusive to one RI. A RI may choose to harmonise identifier mismatches within its own domain but this is outside the scope of CORBEL. We aim to produce a shortlist of common core data types that warrant community focus for harmonisation and warrant some disruption: for example Samples.
4. *Shared Service view:* we focus on services that are in common or that can be shared and new services that are needed. We aim to produce a shortlist of services that warrant harmonisation: for example, resolution services in Europe and the USA.

Identifier Strategy Checklist

General questions

General		
	Explanation and Examples	*Identifier issue*
Can you approximate how many entities need to be identified?		
What sustainability model is used in the assignment of identifiers, i.e. who pays for them to be assigned and maintained.	For example, UniProt assigns protein identifiers as part of delivery of their database. DOIs have a cost model associated with minting	
If you use identifier systems please provide the name, URL and example functionality.	For example, Protein Identifier Cross Reference Service http://www.ebi.ac.uk/Tools/picr/, mapping of protein identifiers between different identifier schemes	

Table 1. General identifier questions

What is being identified? What is being assigned an identifier?

Identifiers in different domains may refer to similar, but slightly varying records. For example: The description of a protein using its amino acid sequence or its three dimensional structure.

Entity type	Consistent identification of entities can have consequences in analyses. For example, in Rare Disease, the need to uniquely and unambiguously identify individuals is required to avoid the same individuals appearing in multiple resources represented differently and therefore skewing calculations of disease frequency.	
	Explanations and Examples	*Identifier issue*
Is it a database?	E.g. UniProt	Clarity of what is being identified
Is it the location of the database?	E.g. www.uniprot.org	
Is it an entity concept?	An entity concept that persists E.g. variant	Resolve to the same entity
Is it an ontology term?	E.g. abnormality of the eye	Find a unique stable identifier for ontology term

Is it an entity's database record?	An entity's record will likely evolve through many versions E.g. ENA genepage	Resolve to a version of a record
Is the entity record an entry in an experimental archive?	Time date stamped and stable datasets deposited by researchers. E.g PRIDE archive entry	Resolves to a record
Is it a knowledgebase entry?	Curated and annotated biological entities and their relationships representing current biomedical knowledge and analytical processes over other datasets. Records are updated, may diverge or merge over time, and may change considerably in their interpretation as new knowledge is acquired. E.g UniProt entry, Orphanet entry, Reactome entry, ChEMBL entry	Resolve to a version of a record without guarantee that the content of the entry is the same upon each resolution
Is it a physical object?	A physical resource that is described by a metadata record E.g. Biobank sample a strain from a microbial Biological resource Collection (mBRC)	Resolves to a metadata record RRID https://scicrunch.org/resource is an identifier assignment for research materials
How is metadata stored and bound to an entity?	E.g. a variant may be functionally classified as missense. Is how this metadata was determined and when it was assigned transparent?	Clarity of provenance and type of metadata
Is it metadata for an entity?	E.g. date of creation or update	What metadata is required by an entity
Does this entity have close related link records in different datasets?	The relationship would be displayed as a mapping E.g. A protein described as 3-D structure in one dataset and a sequence in another.	Identifier resolution landing page include mapping to other dataset identifier
Entity Granularities	CORBEL has high variability in the granularity of its identification needs, for example a physical biological sample or material, a collection of samples, or a genomic variant. Persistent identifiers for datasets must support multiple levels of granularity to support both the identification and citation of a specific version and/or individual dataset, as well the identification and citation of an unspecified version of a dataset and/or a collection of primary data.	
Is it an indivisible record?	There are no identifiable sub elements E.g. single nucleotide polymorphism (indivisible) or a chemical salt (divisible)	
Is it a collection of	Primary data is uniquely identified and cited as a collection of	

elements?	potentially many individual items which need their own unique identifiers to support later reuse and recombination into different sets while maintaining the ability to cite the constituent data elements. Lower-level identifiers need to be able to be grouped via a collection identifier and accessed as set elements from the overall collection landing page (Honor, Haselgrove, Frazier, & Kennedy, 2016) E.g. Collection of biosamples Neuroimaging individual subject scans using a given imaging modality are the lowest level at which objects will be identified, while the primary publication will cite a collection level unique identifier.	
A collection with hierarchically organised elements?	The collection has an identifier and its elements have identifiers that may be relative to the collection or independent of the collection. The hierarchy needs to be preserved The BioStudies database (McEntyre, Sarkans, & Brazma, 2015) provide storage for all the underlying data links and files for a publication	

Table 2. Identifier questions for each entity type

What are your data and identifier life cycles?

Creation	Also known as "minting"	
	Explanations and Examples	*Identifier issue*
From which stage of the data lifecycle are the entities you have identified?	E.g. for a clinical trial system the trial may be identified when approval is granted. For an analytical system intermediate data files may be generated but not referenced by any identifier until publication	
When is the identifier created?	Are identifiers created at the same time as the entity is created or versioned or is it created separately and then assigned by a separate process? Allowing minting (creation) of identifiers in advance of submission to a dataset allows the submitter to establish the correct double cross-linking among the correct entities (Smithsonian NMNH Biorepository mints identifiers in advance, Pangaea and ENA do not, so it is impossible to add a Pangaea identifier to ENA Sample data and vice versa add ENA Sample identifier to Pangaea data)	Establishing two-way linking Identifier reservations
How do you check whether an identifier already exists for the entity? Do you use an identifier type catalogue?	Where an entity is already well identified, you should reuse the existing canonical identifier. The types of available identifiers in a given domain, including metadata about the identifier type itself (preferred name, synonyms, definition, example)	

	E.g. Identifiers.org, Gene Ontology, NCBI, EDAM	
What is the designated authority for the identifier? Is the authority registered in a suitable registry?	How do you chose who issues the identifier? For example a major database. Is it ad hoc?	All authorities should be registered in the ELIXIR registry identifiers.org as well as other registries.
Change		
Does the entity object ever change or get updated or revised?		
Are entity objects ever split? How do you handle that?	How do you communicate and synchronise with users of the entity?	
Are entity objects ever merged? How do you handle that?	How do you communicate and synchronise with users of the entity?	
Versions		
Is the entity object versioned? How?	For example is a ".1" appended to the identifier	
Are the versions linked together?	For example do subsequent versions transparently link together	
Are the versions retained?		Required e.g. to replicate analyses
Is the identifier versioned when the entity object is versioned?		Clarity of versioning policy
Is a brand new identifier minted and assigned when the entity object is versioned?		Tracking as new identifiers appear
Is the identifier reassigned to the latest version?		Bad practice
Deprecation		
Is the entity object or its metadata permanent and persistent? What happens if the	Resolution of life-limited identifiers such as OSD id with an "Expected Expiration Date" i.e. the expected date from which on the resource will most probably not be available anymore.	

identifier is resolved for data that no longer exists?		
Is the entity transient (at first)?	Transient or intermediate results may be generated as part of an analysis pipeline E.g. Minids (Minimal Viable Identifiers)[1] are examples: Minid can either be active or "tombstoned"- that is represents data that are no longer available and a Minid may also be obsoleted by another Minid. Entities identified by minids are assigned non-minid identifiers if they are promoted to final product status	Transient data is typically internal to a RI. Not considered a candidate for RI identifier harmonisation.
Do you ever deprecate or reassign the identifier?	How will you propagate to other dataset providers referencing the identifier?	Best practice in identifier reuse discourages deprecation and reassignment

Table 3. Data and identifier life cycle questions

Entity object visibility outside the RI

Entity object visibility		
Is the entity entirely internal to the RI?	It could be that the entity is not expected to be available outside the organisation that generated it. But is the identifier known outside? If so, external datasets could be using an internal identifier.	Maybe only locally unique
Is the entity shared outside the RI at some point?	There could be a step for publishing outside the borders where a new identifier is assigned	The identifier will need to be globally unique and appropriately licensed

Table 4. Entity object visibility outside the RI questions

How do you deal with names and ontology terms mapped to names?

Identifier lookup / indexing		
What scheme do you use for names?	A name is a label that could be formal or informal and is often non-unique. It will be using in combination with an identifier. Some entities have naming authorities, such as the Human Gene Nomenclature Committee (HGNC) for human genes.	CDH16
Do you use a "concept or ontology to identifier" lookup service?	E.g. the Ontology Service Lookup Service (OLS), Open PHACTS IRS system, ConceptWiki.	

[1] http://minid/bd2k.org

Legacy naming systems		
How is the legacy naming systems incorporated with the identifier scheme? How will interoperability between naming systems be handled?		Related to mapping services
Ontology identifiers		
How do you find the identifier for a term from an ontology?	What is the identifier for liver in mouse?	
How do you map data to an ontology identifier?	How do I annotate metadata about a bio samples to ontology terms	
How do I find mappings between ontology identifiers		
How to I create a new identifier for a new ontology term		

Table 5. How to deal with names and ontology terms mapped to names

What are your identifier properties, policies and practices?

For each Entity type

Properties		
	Explanations and examples	*Identifier issue*
Is the identifier globally or locally unique?	E.g. does the identifier refer to a patient available in a local system, or is it some externally visible identifier which is intended for global use?	Understanding how the identifier is expected to be used
Does it use a well known global identifier scheme?	Schemes that are machine actionable, globally unique, widely and currently used by the community, long term persistence, cross-disciplinary. Guaranteed to be globally unique High level entities are typically supported by international standards e.g. ORCIDs for researchers or Digital Object Identifiers (DOIs) for publications. Handle.net (DOI, ARK, ePIC)	Established global solutions are well tested, widely adopted and have supporting infrastructure such as identity registration and resolvers.

	URI (PURL), URL CURIE (compact URIs) Proteomics database PRIDE adopted DOIs. PRIDE dataset identifier PXD000000 DOI:10.6019/PXD000000." Internationalized Resource Identifiers (IRIs) complement URIs. An IRI is a sequence of characters from the Universal Character Set (Unicode/ISO10646). A mapping from IRIs to URI means that IRIs can be used instead of URIs where appropriate to identify resources.	DataCite (DOIs) ePIC (ARKs)
Does it use a local dataset identifier scheme?	An identifier that is unique within the scope of a single database. Also known as "Accession Numbers". The user community more often uses dataset identifiers than global identifier schemes. ZDB-GENE-980526-388	Tools are needed to transition between global and local id schemes
Is the identifier permanent and persistent? Or do you expect it to be temporary and deprecated at some point?	A permanent identifier has a layer of indirection which provides a unique stable reference to data which may be located in multiple locations and to which we can bind specific descriptive metadata elements such as the author and creation date. A persistent identifier once minted will persist for eternity and attributes associated with it, such as its creator, cannot be changed (though the object identified might change)	Clarity on permanence of identifiers
Does the identifier have a checksum or check digits?	The last character in the ORCID iD is a checksum.	Understanding identifiers
Do you know of alternative URIs/Accession Numbers that other groups use for your identifiers?	Alternates are not recommended for use, knowing what which URIs are equivalent facilitates data integration	Different resolvers
Under what license are identifiers made available?	Can the identifiers be reused freely? A practical problem for compounds has shown to be ID mappings to closed data. DrugBank made their IDs available as CCZero, even if the data itself has a CC with NC clause. https://orcid.org/legal	Reach through on licences and clarity on who can use identifiers for which purposes
What identifier creation services do you use?	For example ORCID for researchers	Lack of knowledge on creation services

Where do you register your identifiers?	For example, identifiers.org	Different places to register identifiers
Identifier formats		
Are opaque or semantic identifiers used?	Is any meaning embedded in the identifier or is it simply alphanumeric e.g ORCID	Detecting when identifiers are also conveying meaning
What is the format? Is it unambiguous? regular? documented?	UniProt describes its formats here: www.uniprot.org/help/accession_numbers	Transfer of knowledge to users and best practice in identifier design
Is there a URI pattern? Are there multiple, equally-valid URI patterns coexisting?	INSDC.org has four such schemes as the entire dataset is fully represented by each of three authorities: NCBI, ENA, and DDBJ	Which to resolve to
Does the identifier have a unique prefix? Do you use ":" or "-" other than just for your delimiter? Is it registered in a prefix commons?	If your Local IDs already have a colon what is your preferred corresponding compact URI syntax? GO:0007049, the prefix 'GO' can be expanded to http://purl.obolibrary.org/obo/GO_ and prepended to the numeric fragment to yield http://purl.obolibrary.org/obo/GO_0007049. UZFIN UniProtKB MyDB_gene_6622 will turn into MyDB_gene:6622 or MyDB:gene_6622 or MyDB:gene:6622	https://github.com/prefixcommons/biocontext
Are there white spaces or non-ASCII characters or patterns that will cause confusion Is the expression case-sensitive?	E.g. May-15, 5e1234, bad-12 mean other things Bad.12 is confusing for versioning ab-1235 ≠ AB-1235 is waiting to cause trouble	"." should be reserved for versioning schemes
Do you use or can you create a CURIE for the identifier? What is your preferred CURIE?	CURIE's are compact URIs. They are <Prefix>:<Local ID> where the prefix is expandable to a URI pattern. A URI Pattern is fixed sequence of characters used to resolve a database's local IDs. UZFIN:ZDB-GENE-980526-388 URI pattern: http://zfin.org	URI patterns vary. See Appendix 1
Do you use identifier conversion tools/services, what are the selection	E.g. PICR for proteins	Many tools, selection criteria needed

criteria, where do you look for these?		

Table 6. Identifier properties, policies and practices

How do you lookup and resolve the identifier?

Identifier resolution	Given an identifier a resolution service returns a representation of the entity	
	Explanations and examples	*Identifier issue*
What resolution service do you use?	For example: InChI resolver, identifiers.org, n2t.net	Multiple resolution services
Does the resolution service return a landing page?		Landing pages are essential for citation and viewing
Does the resolution service return machine processable metadata?	E.g. JSON, XML, RDF, JSON-LD formats	Use in programmatic applications
What is the metadata associated with the identifier?	Key for provenance, versioning, citation, and to build the description for a landing page or for machine processable metadata.	
What happens when the identifier resolves to a deleted data object?	References to identifiers and their metadata may persistent beyond the lifetime of the data entity, especially when identifiers are referenced in third party datasets	Clean response to failed resolution.

Table 7. Identifier lookup and resolving issues

How do you support identifier metadata and data citation?

Data Citation	Martin Fenner, et al (2016). A Data Citation Roadmap for Scholarly Data Repositories doi: https://doi.org/10.1101/097196	
	Explanations and examples	*Identifier issue*
How do you support data resolution landing pages?	The persistent identifier expressed as URL must resolve to a landing page specific for that dataset. The persistent identifier must be embedded in the landing page in machine-readable format.	The repository must provide documentation and support for data citation.
Do you support citation metadata associated with the identifier?	The landing page should include metadata required for citation, and ideally also metadata helping with discovery, in human-readable and machine- readable format.	

	The machine-readable metadata Should use schema.org markup in JSON-LD format. Metadata should be made available via HTML meta tags to facilitate use by reference managers. Metadata may be made available for download in Bibtex or other standard bibliographic format	
Do you support content negotiation?	Content negotiation for schema.org/JSON-LD and other content types may be supported so that the persistent identifier expressed as URL resolves directly to machine-readable metadata. HTTP link headers may be supported to advertise content negotiation options	
Do you have recommendations for best practice for data citation	E.g. Proteomics database PRIDE provides best practice for *data citation* using DOIs	

Table 8. Metadata support identifier and data citation questions

What relationships are there between identifiers?

Identifier Referencing		
	Explanations and examples	*Identifier issue*
Do you identify entities that are also identified by others? Who are these others?		Standardised and shared common identifiers
Do you reference identifiers that are issued by other authorities? If so, in what cases? How often are the identifiers synchronized?		Identifier resolution and maintenance outside a dataset needs synchronisation and maintenance
Do you reference identifiers issued by other authorities? Where can your mappings be found?	What are the mappings used for prefix-to-URI patterns? What is the source of these mappings (e.g. manual or identifier service).	
Identifier mapping	Map identifiers on entries in one resource to another to assign equivalences, similarities, or otherwise map or link	

Are there relationships between your identifiers? Where are these described?	Relationships may be embedded with the report (e.g. Uniprot) or held by mapping services. They could even by in Excel spreadsheets	
Do you map identifiers to each other?		
Do you map directly?	Compare identifier values or maintain cross-reference linksets Gene, Protein, Metabolite Identifier mapping service BridgeDb	
Do you map indirectly?	Semantic-free cross references Ontology mapping service	
Are the maps between equivalent or similar entities?	The choice when two entries about a small molecule in different datasets are the same depends upon the application to which the data will be put. "Equivalence" rules may be based on the context and interpretation of the links[2].	
Are the maps between related entities?	Cross-references map entities of different types e.g. genes to proteins, or between stereoisomers of chemical compounds	
Are the maps calculated?		
Where are these mappings found and who, if anyone, maintains them?	For example, the Open PHACTS warehouse maintains "linksets". (http://www.openphacts.org)	
What mapping services do you use?	Identifier mapping service BridgeDb[3] (non-IRI mappings) Identifier Mapping Service from the Open PHACTS Project (IRI mappings)	

Table 9. Relationship between identifiers

Summary of Identifier services you use*

Service	Description of service
Identifier registration	Register an identifier with an authority so that it can be resolved
Identifier resolution	Resolve an identifier: given an identifier return a representation of the entity.

[2] Batchelor et al 2014, Doi:10.1007/978-3-319-11964-9_7
[3] http://www.bridgedb.org/

Identifier conversion	Convert one form of an identifier to another.
Identifier creation	Create a URI or other identifier given certain parameters
Identifier verification	Verify that the identifier is well-formed, valid and resolves
Concept to identifier lookup	Ontology or name indexing/lookup to return identifier
Identifier type catalogues	Catalogue the types of available identifiers in a given domain.
Identifier mapping	Map identifiers on entries in one resource to those in another, in order to assign equivalence to entries or otherwise link two resources.
Annotation aggregation tools	Compilations of existing identifiers together with higher-level structures related to an entity.

Table 10. Summary of Identifier Services

*Examples of Services and Tools are listed in Appendix 2

Identifier Harmonisation Strategy Checklist

This checklist (shown in Table 11) is less mature than the previous one. As we review the case studies and apply to WP4 Use Cases this will be iteratively improved.

What are the data entities in common (overlap)? Exclude transient data entities *For entities in common (overlap) harmonisation on properties and practices refers to* ● *Convergence on similarities* ● *Compatibility on differences that need to harmonised* ● *Complementaries on differences that do not need harmonisation* *For entities **not in** common (overlap) but used in the use case, harmonisation refers to the ability for both RIs to access and reuse those entities through their identifiers.*	
Can the identifiers be reused freely? *A practical problem for compounds has shown to be ID mappings to closed data.* *DrugBank made their IDs available as CCZero, even if the data itself has a CC with NC clause.*	
For each of the entities in common	
Common Entities	Are the entities canonically the same? *E.g. both RIs use UniProtKB for protein*
Reused entities	Do you identify entities in each others infrastructure? Do you reference identities issued by each other's authorities? How often are they synchronised? Do you use the native identifiers and services?

Entity matches/ mismatches where they are not canonical	Are they really the same entity at the same level? Concept, record, physical object? E.g. the identifier for a fully sequenced gene and the identifier for the same gene, described by a minimal set of descriptive alleles which may turn out to carry different versions of a previously hidden mutation. Do the granularities align? If not, how do they match/mismatch? Are the entities wholly or partially in common?
Access	Are they local to the RI or intended for global access?
Life cycle	Are there differences in practice for the entity ● Creation ● Changes ● Versioning ● Deprecation What are the differences in identifier practices?
Relationships	Are the entities already linked? How?
Equivalences	Are the entities equivalent? Or similar? Or related? Can you or do you use a identifier mapping service between them?
For each of the identifiers for entities in common	
Authorities and assignment	Do you share an identifier authority or use different ones? Are the authorities compatible?
Names	Do you share common names? Do you use the same name authority?
Identifier schemes	Do you use the same schemes? Are they compatible? Are they standardised? Can they be mapped? Do licenses limit mapping potential?
Identifier services	Do you share any services? Do you have services in common? Are they compatible? Can you use a service?

Table 11. Identifier Harmonisation Strategy Checklist

Harmonisation Strategies

We aim to harmonise at several levels: identifier format, the entity identifier, service and policy.

Identifier format level

Common adoption

Common adoption of standardised format adhering to the principles in Identifiers Format Checklist based on McMurry et al., (2016). For example, global resolution of locally-assigned accession numbers using CURIEs. This impacts legacy datasets and the datasets that reuse the legacy identifiers.

Conversion

Conversion between formats, notably the conversion between legacy formats and standardised formats.

Identifier entity level

Common adoption

Common adoption of canonical standardised identifiers for an entity from an agreed authority or set of authorities. This impacts legacy datasets and the datasets that reuse the legacy identifiers and requires agreement on the authority.

Mappings

Mappings may be point to point (identifier to identifier) or to a canonical preferred identifier. Mappings may be computed or enumerated into "linksets" or "mapping sets", indirect or direct and vary in the semantics of the mapping. In all cases they need to be maintained and retro-fitted into resolution practices. Examples of mapping services are identifier aggregation systems (e.g. myGene), identifier mapping services (e.g. BridgeDB, the Open PHACTS IMS), and local mappings.

- *Rosetta Stones* standardize on one identifier (by a common identifier authority) as the common connection between various data providers. The identifier must guarantee stability, sustainability and redundancy. The Genome Standards Consortium proposed a Genomic Rosetta Stone (Van Brabant et al., 2008), using a BioProject identifier and the NCBI LinkOut system[4]. Each data provider updates their own mappings and any data provider that wishes to join the system needs to register with NCBI LinkOut and upload a description of its mappings. They are given a provider identifier and can submit their own name and name abbreviation into the system. The system is similar to that of using a provider prefix to group identifiers from that provider, with the addition of the mapping to the Rosetta Stone (see Figure 2). Mappings are uploaded in a specialized XML format documented by NCBI, consisting of object ID (or query) and URL pairs. The GRS Resolver[5] provides a REST interface of the querying of mappings.
- *Identifier aggregation systems*, compile existing identifiers together with higher-level structures (see Appendix 2) related to an entity. *e.g.* the bioDBnet[6] includes tools to report

[4] https://www.ncbi.nlm.nih.gov/projects/linkout/
[5] https://github.com/wdesmet/grs-web
[6] https://biodbnet-abcc.ncifcrf.gov/

all available information for an identifier, interconvert identifier formats, and converts molecular sequence identifiers for one organism into the corresponding identifiers of a different organism. MyGene.info[7] provides REST web services to query/retrieve gene annotation data. The Monarch system[8] enables the aggregation, provenance, and currency of hundreds of external resources, while integrating them to ontologies for phenotypes, diseases, genotypes, and anatomy.

- *Identifier mapping services* (e.g. BridgeDb, the Open PHACTS IMS), *map* identifiers on entries in one resource to those in another, in order to assign equivalence to entries or otherwise link two resources. Methods are direct or indirect. Direct methods map by comparing identifier values, including those that are the database accession and/or which provide a cross-reference to another resource. For example, the Protein Identifier Mapping Service[9] resolves protein identifiers across multiple databases that correspond to the same protein. BridgeDb[10] is a framework for finding and mapping equivalent database identifiers: it is a framework, live services, and identifier mapping files for genes, proteins, and metabolites. *Indirect methods* do not rely on identifier values to achieve the same ends, for example, a mapping of equivalent concepts in two ontologies may be achieved through comparison of terms and synonyms that are associated with the concepts. Mappings with provenance, different types of mappings, derived and equivalent entities, exact and non exact synonyms. In some cases XREFs, which are pointers to related entities, are sufficient e.g. when the entities are of the same type. However, XREFS can be used to map entities of different types e.g. genes to proteins and this can present ambiguities for some applications.
- *local mappings*. Uniprot, for example, maintains mapping data (XREFS) between a Uniprot identifier and identifiers from several external registries.

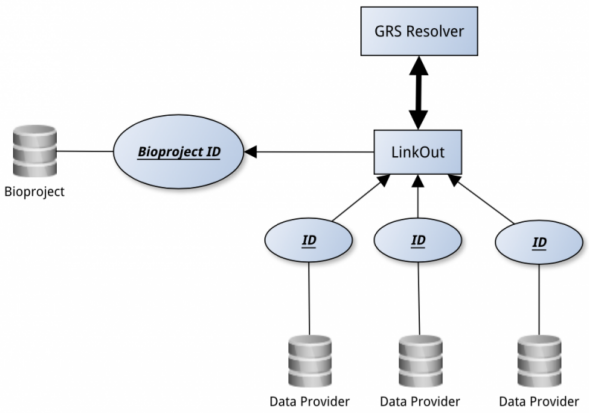

Figure 2. The Genomic Rosetta Stone maps identifiers to each other using the BioProject identifier as a common identifier. These mappings are stored in LinkOut, which can be queried for mappings from BioProject identifiers to data provider records. The GRS Resolver indexes LinkOut mappings and extends its functionality by providing reverse mappings.

[7] http://mygene.info/
[8] http://monarchinitiative.org/
[9] http://www.ebi.ac.uk/Tools/picr/
[10] http://www.bridgedb.org/

Service level

Common adoption

Common adoption of a shared service between RIs: for example common adoption of the identifiers.org registration and resolution service.

Common approaches

Common approaches are agreed by different services through shared and standardised specifications, APIs and schemas. For example, the harmonisation of n2t.net and identifiers.org for global resolution of locally-assigned accession numbers using CURIEs, based on a shared registry of defined resource prefixes and provider codes.

Example: Compact identifier resolution service harmonisation

The BioCADDIE/Force11 BD2K Data Citation Implementation Pilot sets out to work with publishers and data repositories on implementing the Data Citation Principles. Emerging from this work came an activity to align identifiers.org (ELIXIR) and n2t.net (BD2K) identifier resolvers around a shared syntax of Compact ID syntax: <resolverURI/<prefix>:<localAccession.

The International collaboration operated throughout the past year. FORCE11 brought together participants from the USA and Europe in a workshop on June 2, 2016 at Harvard University:

- USA: T Clark, J Grethe, I Fore (BioCADDIE/Force11) N Kunze & G Janeé (Californian Digital Library), Julie McMurry (Prefix Commons).
- Europe ELIXIR: N Juty & S Wimalaratne (EMBL EBI), R Jimenez (ELIXIR Hub), N Beard (ELIXIR-UK, Manchester)

This common approach for global resolution of locally-assigned accession numbers, based on a shared registry of defined resource prefixes and provider codes (see Figure 3), was presented by Juty and Clark at the Pidapalooza international meeting in November 2016.

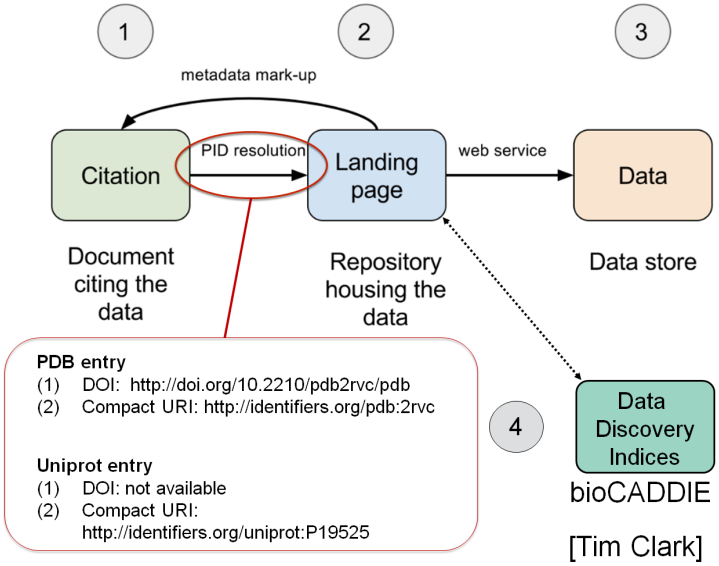

Figure 3. Common approach for global resolution of locally-assigned accession numbers.

A technical approach for common prefix registry has been agreed and specification document has been drafted and a pilot implementation developed at EBI and CDL.

The work is reported in: Sarala M. Wimalaratne, Nick Juty, John Kunze, Greg Janée, Julie A. McMurry, Niall Beard, Rafael Jimenez, Jeffrey Grethe, Henning Hermjakob and Tim Clark *Uniform Resolution of Compact Identifiers for Biomedical Data*, bioRxiv doi: http://dx.doi.org/10.1101/101279

Policies level

Policy level focuses on ensuring that identifier policies and conventions are compatible. In particular lifecycles (create, change, versioning, deprecation) and policies for local vs global identifiers need to be compatible.

Case studies

Here we extend our previous case studies from the BioMedBridges project. Previous case studies assessed usage of identifiers for different entities but did not address the requirements for identifier management services (see Table 10 for examples) or needs for harmonisation across BMS RIs.

We have selected the following case studies as these address new areas for integration, address new infrastructures and have pan resource implications.
Each was asked to

- Describe the study, why selected, relevance to the project, scientific use case
- Provide a list of identifiable entities relevant to the study
- Describe any limitations for the user or the resource developer
- Describe any existing services used
- Describe any gaps in identifer services
- Describe Outcomes, recommendations and actions as above

Case Study 1. Rare Disease Case Study Rett Syndrome

Rett Syndrome is a rare genetic neurological disease for which there are numerous resources, though these are not well integrated with specialist databases. For example rich genetic information, publications in the biomedical literature, BioBank Samples, variant-gene-disease associations, linked gene expression and pathway data as well as complex phenotypic descriptions and familial information held in closed research databases. The Rare Disease community typically generates small databases of < 1000 records, saved as XLS or CSV files but these have rich phenotypic and other clinical information. These data owners have rich domain knowledge but limited access to technical staff. For these users, simplified procedures to upload their data, display landing pages and track provenance as data changes must be supplied with a minimum of technical support. An ELIXIR sponsored implementation study examined resources for Rett syndrome and this has provided input to this case study. Rare disease was selected as it touches many of the BMS RI's activities, provides a small, highly curated and highly accessed dataset with high visibility. Additionally it builds on several of the entity specific case studies (gene, literature etc) delivered by the BioMedBridges project thereby leveraging previous work.

The details of this case study were extracted from the Rett Syndrome BYOD held in November 2016 and reviewed by the study organiser Marco Roos for accuracy. Several scientific use cases were provided by participants to this workshop, the most relevant of these to the process of identification of several entities and linking between them using identifier management services and expert knowledge. The workshops outcomes addressed areas out of scope for this deliverable, this case study will therefore focus on these use cases:

Which genes/variants/metabolites are known to be linked to Rett Syndrome?
Verifying entities identified in the workshop are FAIR (Findable, Accessible, Integrated, Reuseable)

A summary list of identifiers and example data with resources relevant to Rett Syndrome is provided in Table 12. Additionally there are three supporting identifier management services identified by rare disease experts and documented in their workshops. For example Ensembl provides identifier mapping for proteins. These are documented in Table 13. Additionally the participants in the Rett workshop performed their own identifier mapping and listed the identifiers they had found as well as several unidentified concepts (Table 14).

An excellent example of the domain complexity is provided when searching Rett Syndrome in DisGeNet (Figure 4). Multiple lexical variants of Rett Syndrome are provided from the Unified Medical Language System (UMLS), all cannot easily be queried together due to the complexity of the underlying data. The different identifiers are supplied to the user indicating that these are distinct concepts but the user must decide what to query based on the identifiers and labels alone. There is no indication of how these semantic concepts are related to each other to aid the user.

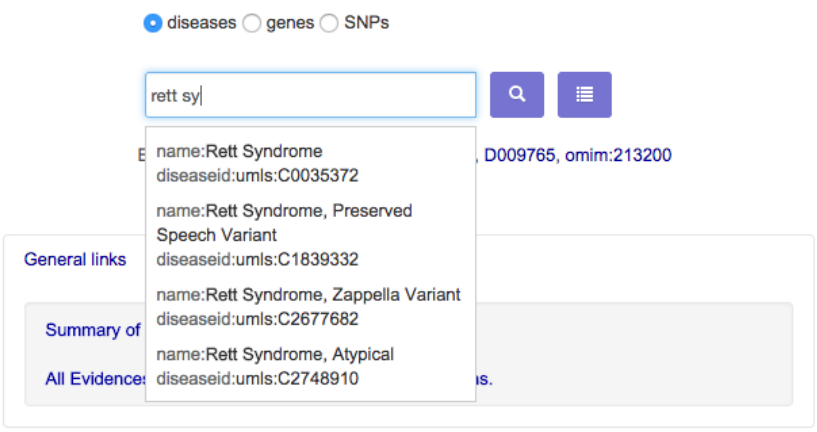

Figure 4. A screenshot from a DisGeNet disease search for Rett Syndrome indicating the complexity of existing semantic resources in searching different forms of Rett Syndrome

Entity	Resource and example identifier
Variant	dbSNP rs28934905

Gene	HGNC:6990
Protein	UniProt P51608
Gene Expression	ArrayExpress E-GEOD-6955
Gene Disease Associations	DisGeNet, N/A
Drug	ChEMBL, CHEMBL69073
Animal Models	Monarch https://monarchinitiative.org/disease/OMIM:312750#models
Pathway	WikiPathways MECP2 and Rett Syndrome pathway: WP3584
Patient Phenotype	Human Phenotype Ontology HP:0004395 (malnutrition)
Patient Registry	*http://www.rarediseasesnetwork.org/cms/rett/Get-Involved/Contact-Registry N/A*
Clinical Trial	*NCT00069550 Clinicaltrials.gov*

Table 12. Examples of Entities Relevant to Rett Syndrome Case Study, those in bold contain entities addressed in the workshop, others are present for completeness

Entity	**Service Provider**	**URL**
Protein	Ensembl, identifiers and mappings	http://identifiers.org/ensembl/{ID
Small molecules (Drug)	PubChem, identifiers and mappings	http://identifiers.org/pubchem.compound/{ID
N/A	Identifiers.org (resolution using templates for services above)	http://identifiers.org/

Table 13. Examples of services used in the Rett syndrome workshop selected by the participants for their entities of interest

Name as found in Rett database	**ID**	**ID name (if different)**
Phelan Mc Dermid syndrome	http://identifiers.org/OMIM/606232	
Lesch-Nyhan syndrome	http://identifiers.org/OMIM/300322	
stereotypical hand wringing	http://identifiers.org/hpo/HP:0012171	
abiotrophy	http://identifiers.org/hpo/HP:0007369	Atrophy/Degeneration affecting the cerebrum
epilepsy	http://identifiers.org/hpo/HP:0001250	seizures
seizures by fever	http://identifiers.org/hpo/HP:0002373	febrile seizures
tremor	http://identifiers.org/hpo/HP:0001337	
non-epileptic phenomena		

intractable epilepsy		
epilepsy becoming resistant to therapy		
dyspraxia	http://identifiers.org/hpo/HP:0011442	Abnormality of central motor function
ataxia	http://identifiers.org/hpo/HP:0002066	gait ataxia
trunctural ataxia		
scoliosis	http://identifiers.org/hpo/HP:0002650	
cervical scoliosis	http://identifiers.org/hpo/HP:0002947	cervical kyphosis
kyphosis	http://identifiers.org/hpo/HP:0002808	
hyperlordosis	http://identifiers.org/hpo/HP:0003307	
asymmetry in muscle tonus		
dystonic	http://identifiers.org/hpo/HP:0001332	dystonia
floppy infant	http://identifiers.org/hpo/HP:0001290	Generalized hypotonia
diplegic gait		
tiptoe walking	http://identifiers.org/hpo/HP:0030051	tip toe gait
crying spells		
microcephaly	http://identifiers.org/hpo/HP:0000252	
microencephaly		
autoplexia		
frozen rigidity	http://identifiers.org/hpo/HP:0002063	rigidity
equinus feet/feet deformation	http://identifiers.org/hpo/HP:0040069	Abnormality of lower limb bone
unvoluntary movements	http://identifiers.org/hpo/HP:0004305	involuntary movements
normal handuse		
swallowing problem	http://identifiers.org/hpo/HP:0002015	dysphagia
limited handuse		
atactic gait	http://identifiers.org/hpo/HP:0002066	gait ataxia
knee walking		
severe torsion scoliosis passively redressable		
cavus foot	http://identifiers.org/hpo/HP:0001761	pes cavus
prone to agitation	http://identifiers.org/hpo/HP:0000713	Agitation
friendly interactive		
lethargic	http://identifiers.org/hpo/HP:0001254	lethargy
autistiform behaviour	http://identifiers.org/hpo/HP:0000729	autistic behaviour
sleep disorder	http://identifiers.org/hpo/HP:0002360	sleep disturbance
hyperactivity	http://identifiers.org/hpo/HP:0000752	
preserved speech		

friendly and quiet behaviour		
night crying	http://identifiers.org/hpo/HP:0030215	inappropriate crying
mood changes	http://identifiers.org/hpo/HP:0001575	
somnolent	http://identifiers.org/hpo/HP:0001262	somnolence
bad character	http://identifiers.org/hpo/HP:0006919	Abnormal aggressive, impulsive or violent behavior
obesity	http://identifiers.org/hpo/HP:0001513	
low BMI	http://identifiers.org/hpo/HP:0004325	decreased body weight
malnutrition	http://identifiers.org/hpo/HP:0004395	
obstipation	http://identifiers.org/hpo/HP:0002019	constipation
bloating	http://identifiers.org/hpo/HP:0003270	Abdominal distention
diabetes mellitus type 1	http://identifiers.org/hpo/HP:0100651	
growth deficit	http://identifiers.org/hpo/HP:0001510	Growth delay
cholelithiasis	http://identifiers.org/hpo/HP:0001081	
regurgitation	http://identifiers.org/hpo/HP:0002013	Vomiting
osteoporosis	http://identifiers.org/hpo/HP:0000939	
breathing irregularities	http://identifiers.org/hpo/HP:0002793	Abnormal pattern of respiration
feeble breather		
forceful breather		
forcefull breathing with probable vacant spells		
Valsalva type breathing		
plays with breath		
hypoventilation	http://identifiers.org/hpo/HP:0002791	
apneustic breather	http://identifiers.org/hpo/HP:0002882	Sudden episodic apnea
cyanosis	http://identifiers.org/hpo/HP:0000961	

Table 14. Examples of Identifiers and Labels for Patient Phenotypes Detected in the Rett Syndrome Workshop.

Rett Syndrome Case Study Outcomes and Recommendations

1. Several unidentified entities were found specifically in the area of reporting patient phenotypes (listed in Table 14), for example, '*feeble breather*' could not be mapped to a concept in the Human Phenotype Ontology. A subsequent search in the Ontology Lookup Service and BioPortal provides no exact match for this term and the expert therefore has to generate a new term, provide an identifier for this term and make this term accessible or choose a less precise term to describe the phenotype.

Recommendation: Provide a user-friendly application to generate and clearly identify a new term, support review/moderation by an expert and identify it or provide a system to record non-exact mappings to existing terms.

Action: Test the Webulous ontology development and URIgen ontology identifier generation system with CORBEL users to assess if it is performant for this use case (D6.3)

2. Multiple semantic forms of Rett syndrome present in one or more resources with differences hard to detect by a non-expert

Recommendation: Provide mappings between terms with provenance and expose these as a service so that users may benefit from ontology mappings created by other projects. For example, the Monarch project has apparently mapped these and the search behaves more intuitively than DisGeNet.

Action: Test the OxO ontology cross reference system under development with CORBEL users (D6.3)

3. Identifiers.org was selected as the resolution service and the required resources, Ensembl and PubChem were already present in identifiers.org enabling work to proceed

R**ecommendation: Extend the representation of CORBEL partner resources in identifiers.org based on audit of BMS RIs using the checklist (this deliverable)**
Action: Identify novel resources, determine if resolutions services exist and add if not

4. Several of the ontology terms had different labels to the ones used in the Rett database.
Recommendation. Provide a means of suggesting synonyms to the ontology
Action: Develop a template and process for adding new synonyms to ontologies

5. The workshop participants selected several data sources and services of these based on recommendations of informatics experts present at the workshop. This is not always possible for all users and communication should be improved to allow those without access to experts to select a service
Recommendation: Deliver improved information on the ELIXIR knowledge hub for services and resources based on quality and provide CORBEL specific information for cross BMS RI users
Recommendation: Ensure the services identified by the workshop are present in the bio.tools life sciences services registry
Action: All services are present in the bio.tools registry

6. Patient identifiers were discussed in the workshop as the participants used a specialist Rett syndrome database. These are presumed to be local to participants and not widely shared.
Recommendation: Patient identifiers for consented and managed access data are not in scope for CORBEL sharing unless these resolve to some global service
Action: Provide recommendations on identifier assignment and expected resolution for consented and/or managed access databases and determine CORBEL scope. For example, the European Genome-phenome Archive (EGA) contains such data but doesn't make these available at the sample/individual level.

7. A major mapping exercise between gene and variant was needed for this project and a mapping database was constructed specifically for it. This means that the data sources need to be updated as new genes, variants, protein ids and genome builds become available through dbSNP and Ensembl who produced the mappings. This is a considerable task at a per disease granularity and was released through the BridgeDb mapping database for future use.

Recommendation: Define best practice in design and updating of mapping services

Action: Deliver to the ELIXIR Knowledge hub for community use

Case Study 2. BioBanking

The Estonian Genome Center[11] is the national BBMRI node in Estonia[KM1] , integrating data from donors with regular updates from the national healthcare system. The nationally highly developed technological infrastructure, where each citizen is given a unique ID number provides a good foundation for data integration. Biobank data is collected from questionnaires, measurements and blood samples, stored as mainly DNA, plasma and buffy coat in standard liquid nitrogen storage, using MAPI[12] storage and identification system. Participant identification, as well as identification of incoming data from the national eHealth system is centrally encrypted and biobank data is stored using two anonymized identifiers (both locally minted); one internal for record keeping and one external for release of data to researchers. Tables 14 and 15 summarises the identifiers need and protocols for biobanking.

Sample data is stored in a similar way, with the internal sample ID linked to the sample tube barcode. Upon request, phenotype data is released in the as standard xls, csv or tsv tables and genomic data is available Oxford genotype file *format*[13] (gen/sample), Impute2 format BAM or VCF formats. Healthcare data is integrated upon request (no technical infrastructure link in place) once per year after QC validation is completed in the medical system. Routinely obtained imaging in healthcare is available from a national archive and can be integrated upon request, however no such research has yet been undertaken. A schematic diagram is provided in Figure indicating how biobanking connects with the research platform.

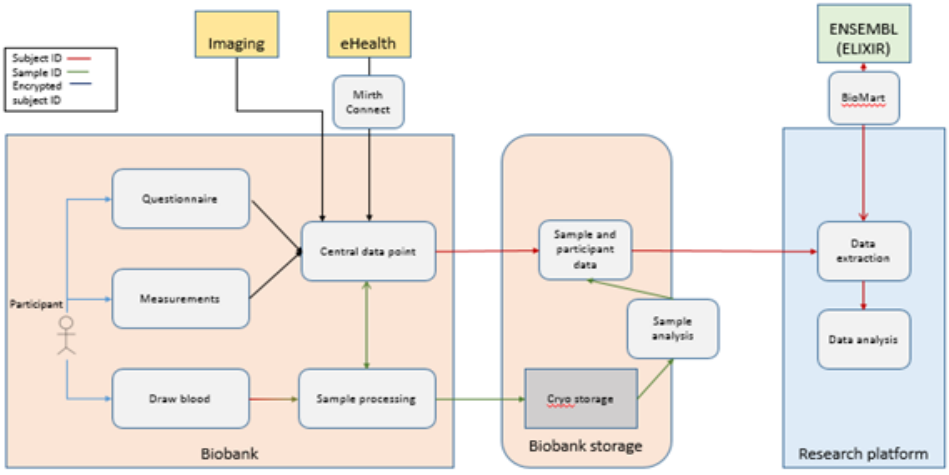

Figure 5. Biobanking processes and infrastructure

[11] http://www.geenivaramu.ee/en/access-biobank

[12] https://www.cryobiosystem-imv.com/en/biobanking/equipments/mapi/13-mapi.html

[13] http://www.shapeit.fr/pages/m02_formats/gensample.html

Identifier property	Need	Comment
Minting	Yes	Local
Resolution	Yes	LIMS
Version	Provenance tracking of samples	Timestamp
Collection	Hierarchy between samples and participants maintained	Maintained by LIMS-system
Mapping	GWAS studies	BioMart
	Healthcare data (HL7)	Mirth Connect
Physical objects	Sample-data link maintained	Double identification system of two different identifiers; physical barcode on tube and colour/position code identifying sample position
Sensitive data	Patient privacy protected, re-identification possible	Two-step identifier encryption at central location with restricted access

Table 14. Biobanking case study identifier needs

Data identified	Standard used	Example item	Example code
Data gathering QC	ISO 9001	-	-
Medication	ATC	Lidocain	N01BB02
Diagnosis	ICD-10	Hypertension	I10
Protocols	Experimental Factor Ontology	Illumina HiSeq 2000	EFO:0004203
Healthcare data	HL7	-	-
Gene annotation	ENSEMBL	BRCA2	ENSG00000139618

Table 15. Biobanking case study identifier protocols

BioBanking Case Study Outcome

1. Time and effort of integration of new healthcare data

Incorporating a new set of healthcare data can take several months, due to the complexity of mapping eHealth records to biobank data records and regular changes to the HL7 standard.

Action: **If the issue is the same across several infrastructures integrating eHealth data, investigate possibility of providing (partial) templated link sets for mapping, or tools for easy-of-use.**

2. Mapping of biobank analytical results to external systems, e.g. for gene annotation
Currently, any gene, protein, gene expression or other annotation is being done by each researcher, and no guidelines exist for the best possible tools and recommended namespaces.
Action: Provide a list of (few) recommended namespaces for each identifier domain, including but not limited to genes, variations (SNPs and larger mutations), proteins, cell lines, drugs and molecular interactions.
3. Biobank data discovery and linkage
Currently, the EGCUT is listed in the BBMRI directory as a single collection together with collection metadata and sample numbers, but any information about samples or participant metadata must be obtained by each individual researcher upon contact with the biobank. The case is the same for all biobanks with the current version of the BBMRI directory[14]. Inclusion of participant and sample metadata in the BBMRI directory is included in the roadmap for upcoming versions, allowing researchers to browse the entire collection. As the use of biobanks increase, many countries have several and the mobility of participants increases, it becomes increasingly likely for cohorts to partially overlap, with separate participant IDs for the same individual. Not likely to be an immediate issue, preparing for a solution now at the early integration stages would reduce the amount of work integrating IDs at a later stage.
Action: Start scoping the possibility of a global (voluntary) participant biobank ID, possibly given at the time of signing consent.

Case Study 3. Euro-BioImaging Image Data Resource

The Image Data Resource (IDR)[15], developed by Euro-BioImaging partners at the University of Dundee and EMBL-EBI, is an added value platform that combines data from multiple independent imaging experiments and from many different imaging modalities, integrates them into a single resource, and makes the data available for re-analysis in a convenient, scalable form. IDR provides, for the first time, a resource that supports browsing, search, visualisation and computational processing within and across datasets acquired from a wide variety of imaging domains including high-content screening, super-resolution microscopy, time-lapse imaging and digital pathology. IDR is built using the OMERO data-management platform (Allan et al, 2012), a widely used system for managing imaging data both in department imaging facilities and in public image data repositories such the *Journal of Cell Biology* DataViewer (Hill, 2008), the ASCB Cell Image Library (Orloff et al, 2013) and the archive for raw electron microscopy image data EMPIAR (Iudin, 2016).

Integrating disparate, distinct datasets requires common vocabularies for annotating experimental, imaging and phenotypic metadata. If used comprehensively and correctly, common vocabularies for gene names, reagents (e.g., small molecule drugs), phenotypes and measurements can provide the basis for querying across datasets collected in different experiments, using different imaging modalities and at different imaging facilities. For a resource like IDR, commonly agreed identifiers and descriptors are critical for enabling links to be made between independent studies and for linking to external information in domain specific databases. Table 16 lists the main types of identifiers used in the Image Data Resource.

[14] http://old.bbmri-eric.eu/bbmri-eric-directory-2.0
[15] http://idr-demo.openmicroscopy.org

Identifier Type	Examples	Example Identifiers	Example of Use in IDR
Study identifier	IDR accession, DOI	**IDR accession**: idr0027, **Data DOI:** http://dx.doi.org/10.17867/10000102	
Study descriptors	Type of high content screen, imaging method, protocol types, PubMed ID	**high content screen of cells treated with library of siRNAs**: http://www.ebi.ac.uk/efo/EFO_0007551, **Spinning disk confocal microscopy**: http://purl.obolibrary.org/obo/FBbi_00000253	https://github.com/IDR/idr-metadata/blob/master/idr0020-barr-chtog/idr0020-study.txt
OMERO entity identifiers	Screen, plate, well, project, dataset, image, annotation	**Screen identifier**: http://idr-demo.openmicroscopy.org/webclient/?show=screen-3, **Image identifier:** http://idr-demo.openmicroscopy.org/webclient/?show=image-1885618	
Sample descriptors	Organism, cell line, organism part	**Homo sapiens:** http://purl.obolibrary.org/obo/NCBITaxon_9606 **HeLa:** http://www.ebi.ac.uk/efo/EFO_0001185	https://github.com/IDR/idr-metadata/blob/master/idr0002-heriche-condensation/screenA/idr0002-screenA-library.txt
Reagent identifiers	Gene, siRNA, chemical compound	**Gene** ENSG00000145919:http://www.ensembl.org/id/ENSG00000145919 **siRNA** M-008868-01 **Compound** Cyclopiazonic acid:https://pubchem.ncbi.nlm.nih.gov/compound/54682463	http://idr-demo.openmicroscopy.org/webclient/?show=gene-ENSG00000145919 http://idr-demo.openmicroscopy.org/webclient/?show=compound-CYCLOPIAZONIC%20ACID
Phenotypes	Cellular phenotypes	**round cell phenotype**: http://www.ebi.ac.uk/cmpo/CMPO_0000118	http://idr-demo.openmicroscopy.org/webclient/?show=phenotype-CMPO_0000118

Table 16. The types of identifiers used within the Image Data Repository.

Studies are identified by DataCite minted DOIs[16] in addition to internal accession numbers (e.g. idr0027).

[16] http://dx.doi.org/10.17867/10000101

BioPortal[17], the Ontology Look Up Service[18] and curator knowledge allowed us to identify that the Experimental Factor Ontology (EFO)[19], NCBI Taxonomy (NCBITaxon)[20] and the Biological Imaging Methods ontology (Fbbi)[21] gave coverage for most of the sample attributes, experimental methods, variables and protocols we needed. EFO already contained the term "high content analysis of cells" this was extended to specify "high content screen"[22] and then to describe types of screen e.g. "high content screen of cells treated with a library of siRNAs", (synonym "RNAi screen")[23] to accurately describe studies in IDR. Additional protocol types were also added. EBI's JIRA ticketing system provided a means to add these new terms. The Biological Imaging Methods ontology was chosen for imaging method terms because it gives good coverage, is used by other resources such as the CELL Image Library, PhenoImageShare, and Virtual Fly Brain and also covers sample preparation and visualization methods which may be added to study annotations in future. However, our evaluation of Fbbi revealed that it does not yet cover concepts that are crucial for IDR/Euro-BioImaging, e.g., super-resolution microscopy. This ontology is not actively maintained so adding new terms is not straightforward and guidance on choosing alternatives is sought.

Reviews performed in the BioMedBridges project[24] demonstrated that consistent annotation of cellular image data sets is required for their interoperability but that no existing ontology comprehensively covered the phenotypes observed in cellular microscopy images. A major output of BioMedBridges was the Cellular Microscopy Phenotype Ontology (CMPO)[25] to fill this gap. This species neutral ontology was built around phenotypes observed in an initial set of high content screens and histopathology datasets (Jupp et al., 2016) and is still being actively developed at the European Bioinformatics Institute. CMPO has therefore been used as the source of phenotypic annotations in the IDR. Terms such as protein localization phenotypes, which follow a standard ontology pattern, have been added to the ontology using the Webulous[26] term submission system. More complex terms e.g. "abnormal microtubule cytoskeleton morphology during mitotic interphase" have been added after discussions with expert ontologists. Some imaging studies have recorded phenotypes that are beyond the scope of CMPO e.g. tissue level gene expression patterns in plants and changes in tumor components such as blood vessels, and identification of suitable ontologies to hold such phenotypic descriptions is needed.

With around 170 ontology terms now being referenced in IDR it will become increasingly important to be able to detect any updates to ontologies and to start recording the ontology version used in annotation. An API that provides alerts and resolution of updates is essential for a resource like the IDR.

[17] http://bioportal.bioontology.org/
[18] http://www.ebi.ac.uk/ols/index
[19] http://www.ebi.ac.uk/efo/
[20] http://purl.obolibrary.org/obo/ncbitaxon
[21] http://purl.obolibrary.org/obo/fbbi
[22] http://www.ebi.ac.uk/efo/EFO_0007550)
[23] http://www.ebi.ac.uk/efo/EFO_0007551
[24] http://www.biomedbridges.eu/
[25] http://www.ebi.ac.uk/cmpo
[26] https://www.ebi.ac.uk/efo/webulous/

Identifiers for genes targeted by siRNAs and gene knock outs, and chemical compound reagents in high content screens are also used to link datasets. In the case of genes the identifiers come from a variety of sources (e.g. Ensembl, NCBI Entrez, RefSeq variants, FlyBase identifiers, Drosophila CG identifiers, SGD, PomBase) and gene annotation builds which means that there is limited linking between studies based on both gene identifiers and gene symbols. For example Table 17 illustrates the case of the gene BOD1 that has been described using Ensembl and NCBI Entrez Gene Identifiers in different studies and different gene symbols have also been used. While it is desirable to keep a record of the identifiers used in each study as it relates to the published data, it would be useful to have a simple way to map between identifiers for search purposes. In small-scale analyses within the IDR, conversion tables created using Ensembl's BioMart facility were used to map genes to a common identifier but a service that provided this resolution would be helpful.

Gene Identifier	Gene Symbol	High Content Screen Accession
ENSG00000145919	BOD1	idr0009-A
ENSG00000145919	FAM44B	idr0013-A
91272	BOD1	idr0006-A
91272	FAM44B	idr0012-A

Table 17: The gene identifier and gene symbol used for the BOD1 gene (formerly known as FAM44B) in different high content screens submitted to the IDR.

In the case of the two large compound screens in IDR, the compound name is used as the link between the reagents because only internal identifiers were provided and no PubChem nor ChEMBL identifiers were submitted with the datasets. Compound names are not the ideal linking identifiers since there may be many synonyms for the same compound including commercial and IUPAC names, but without expert knowledge of the chemical compound domain it is difficult to convert names to ChEMBL, PubChem or InChIKey Identifiers. A service that provided resolvable identifiers would be another useful tool for the IDR.

IDR Case Study Outcomes and Recommendations

1. Requirement for a resolver chemical identifier service for non-experts in the context of IDR data where internal compound identifiers were supplied

Recommendation: Services from projects such as Open PHACTS address these requirements though as many services are available these may not always be easy to find for non-chemistry experts. These are listed in the ELIXIR tools registry as of Feb 2017 and we will discuss with Euro-BioImaging partners how these can be made more accessible.

Action: Work with CORBEL partners to test the services recommended vs. the IDR datasets

2. Requirement for a gene name/identifier conversion service

Recommendation: There are several of these available listed in the bio.tools registry but the search is rather non-specific and there are no means to sort by metrics other than when the service was added/updated/name.

Action: Work through the list of possible services to determine which of these meets the use case and how these can more easily be identified by imaging users when searching bio.tools.

3. Lack of ontology terms for imaging technology and unmaintained existing ontology
Recommendation: A local ontology can be developed using Webulous as part of Task 6.2. However, there is the question of whether the existing ontology can be used as a development framework. It is hard to make a definitive recommendation without discussion with the ontology owners for the non maintained ontology. In this case a local development that follows OBO foundry rules in referencing another ontology's content is suggested.
Action: Develop a local solution while exploring if this can be added to the existing ontology content
4. Description and identification of complex phenotypes outside the current ontology scope
Recommendation: Deans et al. (2016) provide three recommendations on the representation of phenotypes but these do not provide a simple solution to this complex problem. A new development 'PhenoPackets'[27] which is developing standard representations of phenotype-genotype information can be explored to address the complex representational needs. This is a nascent effort and is therefore not an off the shelf solution.
Action: Follow up with examples of complex phenotypes and test the phenopackets representation
5. Ontology change detection with respect to existing annotations
Recommendation: Specify as a use case for Task 6.2
Action: Test existing tools to vs. use cases and extend if necessary

Case Study 4. Marine Metazoan Development Models

CORBEL WP 4 Use Case 4 involves EMBRC who are currently working to develop model databases for marine metazoan development that integrate transcriptomic and morphological data. The initial database will cover 4 marine models: ascidian (Phallusia, Ciona), amphioxus (Branchiostoma), sea urchin (Paracentrotus) and jellyfish (Clytia). A CORBEL workshop held in February 2016 identified the need for a ontology for each model organism to standardise the annotation of the experimental and morphological data. The development of this ontology is being supported by WP6 partners from EMBL-EBI using tools delivered as part of Task 6.2.

The initial phase in any ontology development project is to explore if any suitable ontologies already exists or are under development. The OBO foundry[28] provides a registry describing over 160 biomedical ontologies in the public domain. A number of third party registries have emerged to support searching and browsing these ontologies. Table 18 summarises the major ontology repositories that should be used to locate available ontologies in order to assess the coverage in existing ontologies.

Ontology Repository Name	Repository URL
Ontology Lookup Service	http://www.ebi.ac.uk/ols
BioPortal	https://bioportal.bioontology.org/

[27] https://github.com/phenopackets
[28] http://www.obofoundry.org

OntoBee	www.ontobee.org/

Table 18. Ontology repositories available to EMBRC. OLS is supplied by the CORBEL project. Repositories exhibit overlap of ontologies but offer differing services and tools.

Although some metazoan ontologies exist in the OBO library (e.g. the Cephlopod Ontology), there is no specific ontology describing anatomical and development stages for the model organisms of interest. This is sufficient justification to begin work on a new domain ontology for these organisms. Representatives for each organism have begun the ontology development process by collecting relevant terminology in spreadsheets that will form the basis of the new ontologies. The next step is to select an identifier policy for new terms and incorporating these into a an ontology file published in either the OBO or OWL ontology format. These new identifiers will be used to annotate the data from this use case.

The OBO library is the de-facto standard for publishing new ontologies in the life sciences domain. The OBO foundry is an open community of ontology developers that strive to provide a coherent set of orthogonal ontologies that are developed according a core set of shared principles. One of these principles is the provision of identifiers for an ontology that are stale and persist through a globally unique Uniform Resource Identifiers (URI) and an accompanying and convenient short form identifiers known as a compact URI (CURIE)[29].

The OBO foundry have a process for accepting new ontologies described here http://www.obofoundry.org/faq/how-do-i-register-my-ontology.html. As part of this process OBO will provision an identifier namespace for your ontology (e.g. GO for the Gene Ontology). The policy for OBO identifiers is described here http://www.obofoundry.org/docs/Policy_for_OBO_namespace_and_associated_PURL_requests.html. OBO identifiers are 7 digit numerical identifiers that are prefixed with the ontology namespace to generate a CURIE. For example, a project might request and obtain the prefix "CLYTIA". The ontology would then use ids of the form CLYTIA:0000001, CLYTIA:0000002, where the identifiers part is incremented by 1 for each new term. As these identifiers are compact URIs they can be expanded to their full URI to provide a global identifier that can be resolved via a URL on the web. All sanctioned OBO identifiers are registered with the OBO Permanent URL (PURL) server that resolves using a common URL pattern. For example CLYTIA:0000001 would expand and be resolvable from http://purl.obolibrary.org/obo/CLYTIA_00000001.

The management of identifiers for ontology terms is typically handled by ontology authoring software such as OBO edit or Protege. Both tools provide the ability configure an identifier creation strategy that supports an incremental identifier pattern. Ontologies developed in the OBO format, that only generate the CURIE identifier, will have their identifiers converted to full URIs via an OBO to OWL translation, supported by many tools such as ROBOT and the OWL API. In collaborative ontology development projects where multiple editors add new terms to an ontology it is often necessary to assign identifier ranges for individual ontology authors. This allows authors to add new terms in isolation without clashing with other editors who may be modifying the ontology independently. There are some limitations to this approach as it requires a level of coordination

[29] http://obofoundry.org/principles/fp-003-uris.html

between the authors and often restricts the authoring of the ontology to a suite of tools that are aware of the identifier policy. To address the issues of concurrent editing and identifier provision a number of tools have emerged that manage the minting of new identifiers. Web Protege is a collaborative, web-based version of the popular Protege editing software, and can be configured to provide new identifiers as a service. EMBL-EBI also developed the URIGen server[30] that is a small lightweight application for managing and creating new ontology URIs. URIGen provides an REST API so that identifier generation can be done independently from any particular authoring software and a URIGen plugin to Protege is also provided.

Marine Metazoan Development Models Outcomes

1. Lack of an available ontology for this domain

Recommendation: Deliver a new ontology

Action: Transform existing spreadsheets to a formal representation sharable by an ontology repository

2. Need for identifiers for new ontology terms

Recommendation: follow OBO foundry principles in term identification

Action: apply for a namespace and use tooling e.g. URIGEN to design identifiers

Case Study 5. Ocean Sampling Day - Generating Cross-Domain Data and Entities

The Ocean Sampling Day (OSD) is a simultaneous microbial sampling and sequencing campaign of the world's oceans. It took place on the summer solstice (June 21st) between 2014 and 2016. These cumulative samples, related in time, space and environmental parameters, provide insights into fundamental rules describing microbial diversity and function and contribute to the blue economy through the identification of novel, ocean-derived biotechnologies.

The aim of the OSD Consortium is to generate the largest standardized microbial data set including a rich set of environmental data in a single day that can serve as a reference data set for generations of experiments to follow in the coming decade.

In order to achieve this goal, the OSD Consortium a) centralized the logistics and data management incl. sample handling and sequencing and initial quality control b) aimed to publish OSD data open access to all as soon as the data is ready for scientific analysis and c) followed established data standards and d) followed the recommendation of BioMedBridges[31] to "work with established authorities, e.g. major databases, on assignment of new identifiers, especially where they are expected to eventually host your dataset" e) cross-link the distributed OSD data across domains and infrastructures i.e. for example data items in ENA have links to the appropriate data in Pangaea and vice versa.

OSD is relevant to CORBEL, because OSD created cross-domain, distributed data sets at different granularities and at different times working with several infrastructures across the world including among others ELIXIR, EMBRC and MIRRI. This use case describes the usage of identifiers (Table 19) and services (Table 20) from a data- generation and management perspective.

[30] http://ebispot.github.io/urigen/

[31] https://zenodo.org/record/13924#.WLVbzBBb7XR

Establishing Entities and Identification

At the beginning the OSD consortium needed to gather information about who is interested in participating and where they plan to perform the sampling. Therefore, the project specific OSD Registry Web-App[32] was minting and using the *OSD Id* to uniquely identify geographic sites in connection with data on participants.

At the day of sampling the participants collected volumes of water (i.e. *OSD Sample*) from each OSD site and pumped it through a set of *filters*. The filters had to be sent to a single laboratory for DNA extraction and subsequent sequencing and therefore were labeled with the simple hierarchical identifier scheme of *OSD Id* followed by filter number starting at one. These filters are treated as technical replicates and always one such filter per sample was sent to the Smithsonian NMNH Biorepository for bio-archiving labeled with an identifier provided by the repository. The other filters were extracted, sequenced and several different kinds of quality controlled *sequence data sets* were produced. Additionally, the participants measured different environmental parameters. These data were sent to the OSD Registry using an online web form, which handles all *OSD Registry submissions* with an own local and internal submission id used for managing consistency internally.

The sample and environmental data was subject to intensive manual curation including enrichment of the data. For example each sample was in addition annotated with terms and identifiers from the Marine Gazetteer, Longhurst Regions and IHO Sea Areas. All of this data was derived from Marine Regions[33] (hosted by Marine Flanders Institute, member of EMBRC) to harmonize the naming of the sampling localities. The data was also annotated with terms and identifiers from the Environmental Ontology[34].

After several rounds of manual curation, the first version of environmental data alongside first sets of sequence data were submitted to ENA. This submission required the generation of several Entities as required by ENA: *ENA Study/Project*, *ENA Component Project*, *ENA Sample*, *ENA Experiment, and ENA Run*.

EBI Metgenomics then picks up the submitted data for further metagenomic analysis, creates a new own *EBI Metagenomics Project* and copies the *ENA Sample* data re-using the *ENA Sample* Identifier (see e.g.[35]).

Several weeks later, the *OSD Sample* data was also submitted to Pangaea (now part of de.NBI BioData) who created an *Environmental Dataset* and assigned a DOI (Table 19).

Entity	**Example Identifier**	**Corresponding Resolvable URI**	**Availability Identifiers.org**
OSD Site	OSD1	n/a	-
OSD Sample	-	n/a	-
Filter	OSD1_1	n/a	-
Smithsonian	AB0KM13	-	-

[32] https://mb3is.megx.net/osd-registry/list

[33] http://marineregions.org/

[34] https://www.ebi.ac.uk/ols/ontologies/envo

[35] https://www.ebi.ac.uk/metagenomics/projects/ERP009703/samples/ERS667567

NMNH Biorepository Filter			
Sequence Datasets	http://mb3is.megx.net/osd-files?path=/2014/datasets/workable/metagenomes/merged	http://mb3is.megx.net/osd-files?path=/2014/datasets/workable/metagenomes/merged	-
OSD Registry Submission	333	n/a	-
Marine Gazetteer	3315	http://marineregions.org/gazetteer.php?p=details&id=3315	-
Longhurst Regions	NECS	n/a	-
IHO Sea Areas	1912	http://marineregions.org/gazetteer.php?p=details&id=1912	-
Environmental Ontology	marine biome (ENVO:447)	-	-
ENA Study/Project	PRJEB5129	http://www.ebi.ac.uk/ena/data/view/PRJEB5129	http://identifiers.org/ena.embl/PRJEB5129
ENA Component Project	PRJEB8682	http://www.ebi.ac.uk/ena/data/view/PRJEB8682	http://identifiers.org/ena.embl/PRJEB8682
ENA Sample	SAMEA3275549 and ERS667567	http://www.ebi.ac.uk/ena/data/view/SAMEA3275549	http://identifiers.org/ena.embl/SAMEA3275549
ENA Experiment	ERX714221	http://www.ebi.ac.uk/ena/data/view/ERX714221	http://identifiers.org/ena.embl/ERX714221
ENA Run	ERR770958	http://www.ebi.ac.uk/ena/data/view/ERR770958	http://identifiers.org/ena.embl/ERR770958
EBI Metagenomics Project	ERP009703	https://www.ebi.ac.uk/metagenomics/projects/ERP009703	http://identifiers.org/ebi metagenomics.proj/ERP009703
EBI Metagenomics Sample	ERS667567	https://www.ebi.ac.uk/metagenomics/projects/ERP009703/samples/ERS667567	http://identifiers.org/ebi metagenomics.samp/ERS667567
Environmental Dataset	10.1594/PANGAEA.854419	https://doi.pangaea.de/10.1594/PANGAEA.854419	-

Table 19. DOI only identifies all sample data as one single entity.

Entities	Service Provider	Service URL	bio.tools
OSD Site , OSD Sample , Filter , Sequence Datasets, OSD Registry Submission	Micro B3 IS	https://mb3is.megx.net/osd-registry/list	-
Filter	Smithsonian NMNH Biorepository	https://naturalhistory.si.edu/rc/biorepository/index.html	-
Marine Gazetteer, Longhurst Regions, IHO Sea Areas	Marine Regions	http://marineregions.org	-
ENA Study/Project, ENA Component Project, ENA Sample, ENA Experiment, ENA Run	EBI Archive	http://www.ebi.ac.uk/ena/data/view/PRJEB5129	https://bio.tools/tool/ENA/version/1
EBI Metagenomics Project, EBI Metagenomics Sample	EBI metagenomics	https://www.ebi.ac.uk/metagenomics/projects/ERP009703	https://bio.tools/tool/ebi_metagenomics/version/1
Environmental Dataset	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.854419	-

Table 20. Examples of services used by the Ocean Sampling Day Consortium for archiving and analysis of OSD data relevant to this use case.

Ocean Sampling Day Case Study Outcomes

At the time of writing only the OSD 2014 data has been submitted for long term archiving to ENA and Pangaea, respectively. This has several reasons: a) some inconsistencies and corrections of the environmental and other data are usually found during analysis b) the submission process tends to be complex and involved and c) updating of data is cumbersome, can lead to to new identifiers for new resources e.g. corrections to environmental dataset at Pangea would lead to receiving a new DOI instead of a new identifiable version of the same resource) d) updates can only be made by selected persons.

This combination of reasons made it impossible to solely rely on BMS infrastructures in order to achieve the goal of giving public access to all necessary OSD data openly to all as soon as the data is ready for scientific analysis and attach a minimum of documentation maintained by the OSD Consortium. Therefore, the OSD Consortium has set-up a GitHub project[36] to document up-to-date information on the data, their availability and structure and hosts *sequence data sets* among others[37]. Only at the last two resources the public can find the most recent data and documentation e.g. already including *sequence data sets* of OSD 2015. Currently, the OSD consortium failed to

[36] https://github.com/MicroB3-IS/osd-analysis/wiki
[37] http://mb3is.megx.net/osd-files

establish cross-links between the distributed OSD data across domains and infrastructures. The only reliable way to navigate and map OSD data is by following the OSD Data Guides[38].

Recommendations

1. OSD identifier is not persistent

The web app and services of the MicroB3 Information System[39] were founded by the EU Micro B3 project that ended in 2015. Therefore, the existence and maintenance of this resource is not secured.

Recommendation: Consider enhancing existing service(s) or establish a new service which resolves medium-term identifiers such as OSD id with an "Expected Expiration Date" i.e. the expected date from which on the resource will most probably not be available anymore.

2. Inconsistent use of definitions and terminology

E.g. At ENA the terms/concepts "Study" and "Project" seem to be used in an inconsistent manner.

The documentation at http://www.ebi.ac.uk/ena/submit/data-formats states that Study and Project can be used interchangeably. However, it is more than confusing that on one page http://www.ebi.ac.uk/ena/data/view/PRJEB8682 there are two different links one for a "Project XML" (http://www.ebi.ac.uk/ena/data/view/PRJEB8682&display=xml) and one for a "Study XML" (http://www.ebi.ac.uk/ena/data/view/ERP009703&display=xml) which differ in structure and content.

This keeps it unclear and confuses what is actually identified, two different Entities or one?

Recommendation: Encourage and check that Entities are used consistently at least within a resource.

3. Receiving identifiers after data submission hinders first-off double linking

Neither Pangaea nor ENA allow receiving identifiers in advance of the submission processes. Only the Smithsonian NMNH Biorepository is minting identifiers in advance. This makes it nearly impossible to establish double linking between different resources. I.e. add a Pangaea identifier to ENA Sample data and vice versa add ENA Sample identifier to Pangaea data.

Recommendation: Allow minting of identifiers in advance of submission, that allows the submitter to establish the correct double cross-linking among the correct entities. The submitter knows the data the best.

Action: Investigate if the preregistration of samples in the EBI's BioSamples database and sharing of accessions in future would solve this problem.

4. Environmental Data Content Drift

The current entries at EBI and Pangaea are not up-to-date. The most recent data is only found on project specific non-infrastructure resources. This is a content drift issue where identifiers references different versions of same data sets.

Recommendation: Encourage establishment of data submission brokers who take the responsibility to offer users, in need of complex distributed submissions, a single entry point and take over the responsibility of managing distribute submissions to resources in different domains. On such example is the molecular data broker of the German Federation for Biological Data (GFBio).

Action: Determine what is required to be a broker to ENA or Pangea

[38] https://github.com/MicroB3-IS/osd-analysis/wiki

[39] https://mb3is.megx.net/osd-registry/list

Case Study 6: Gene, Protein, and Drug Data - Open PHACTS

Precompetitive sharing of knowledge was identified is important to move the international drug discovery forward. The Innovative Medicines Initiative (IMI), a collaboration between the European Commission and the pharmaceutical industry set out a project to develop a precompetitive infrastructure to support drug discovery. The Open PHACTS consortium was tasked to implement a semantic web-based platform to support integration of pharmacology resources. Practically, it was asked to define industry-relevant scientific questions and develop solutions to support answering those questions (Williams et al 2012, Azzaoui et al 2013).

The innovation lies in the fact that the research questions have been selected such that they could not be answered by single data sources, and that proper data integration was essential. There originates the need for identifier mapping. However, another design decision is to use only data sources available in the Resource Description Framework (RDF). This requires an identifier scheme based on Internationalized Resource Identifiers (IRIs).

The required data sources (e.g. UniProt, ChEMBL, WikiPathways, DisGeNET) describe various different concepts and entity types. The design of the system also includes a name to identifier framework (the Identifier Resolution Service, IRS). Because the data sets are taken as is, it requires mapping identifiers between identifiers, which is implemented with an identifier mapping service (IMS), based on the BridgeDb platform (Van Iersel et al. 2010), extending an older version with IRI support. Problems that needed to be overcome regarding this data linking, including differences in representation of the pharmacology knowledge. For example, the pathway information (WikiPathway, Kutmon et al 2016) and disease information (DisGeNET) may refer to genes, while ChEMBL and UniProt are strictly about proteins. The concept of Scientific Lenses was set up to accurately describe mappings (Batchelor et al 2014), while the mapping data itself may come from even further data sources, like Ensembl.

A second example of identifier mapping complications are that different resources describe drug information in different ways. Some data sources focus on a chemical representation as found in the pharmaceutical formulation (e.g. salt form), while other resources focus on the so-called parent compound. Furthermore, for data analysis further chemical similarity concepts need to be taken into account, like stereochemistry and charge states. Each different representation typically has different respective entity identifiers, even when names for the entities are the same. A chemical registration service (CRS) was developed to create mappings between all those identifiers (Karapetyan et al. 2015), resulting in link sets using the aforementioned scientific lenses.

Internationalized Resource Identifiers (IRIs) generalize URIs and have replaced the latter in the RDF standards (Table 21). An IRI is a sequence of characters from the Universal Character Set (Unicode/ISO10646). A mapping from IRIs to URI means that IRIs can be used instead of URIs where appropriate to identify resources.

This use case was selected because it involves IRI-based identifiers, covers a wide range of biological and chemical entities, it developed Open Science approaches which can be easily reused, and covers an integration problem central to a lot of the life sciences research and industry in Europe.

Entity	Example Identifier	Availability Identifiers.org
Gene	https://www.ncbi.nlm.nih.gov/gene/282478	http://identifiers.org/ncbigene/282478
Protein	http://purl.uniprot.org/uniprot/Q9Y5Y9	http://identifiers.org/uniprot/Q9Y5Y9
Target	http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL5451	http://identifiers.org/chembl.target/CHEMBL5451
Compound	http://www.chemspider.com/187440	http://identifiers.org/chemspider/187440
Pathway	http://rdf.wikipathways.org/Pathway/WP1019	http://identifiers.org/wikipathways/WP1019
Assay	http://openinnovation.lilly.com/bioassay#29	-
Disease	http://linkedlifedata.com/resource/umls/id/C0004238	-
Patent	http://rdf.ebi.ac.uk/resource/surechembl/patent/EP-1339685-A2	-
Tissue	http://www.nextprot.org/db/term/TS-0171	-

Table 21. Identifiable Entities relevant to the study

Limitations for the users and resource developers

The IRS was identified as a critical and non-trivial component of the infrastructure. Resolving an identifier from a name or label shown to be a complicating aspect in the usability. Partially this is overcome by using semantic typing. Furthermore, the results also depend strongly on correct IRI mappings, which is addressed by the adoption of the scientific lenses, but users and resource developers need guidance around the correct lens to use.

Existing services used

The infrastructure depends strongly on external data sources, both for the knowledge but also for identifier mappings. For genes and proteins such mappings have been adopted from Ensembl, while for small compounds a new platform, the CRS, was introduced. For other entities ontologies have also been used (e.g. the CALOHA ontology for human anatomy). Existing data sources include Ensembl, UniProt/NextProt, ChEMBL, WikiPathways, DrugBank, DisGeNET, and others. See https://dev.openphacts.org/docs/2.1 to retrieve a full list with provenance. BridgeDb and ConceptWiki were services reused and extended to provide various components of the platform (IMS and IRS). Some resources needed conversion into RDF, such as ChEMBL (Willighagen et al. 2013) and WikiPathways (Waagmeester et al. 2016).

Describe any gaps in identifer services

Concepts developed in Open PHACTS, like the scientific lenses (Batchelor et al. 2014), FAIR data set descriptors for identifier mapping data, and structure normalization using the CRS are not generally applied yet. Moreover, databases generally do not precisely describe the concepts for which they

define identifiers in their database(s). This requires further adoption of ontologies, and particularly ontologies that capture the specific concept of that entity. That is, an ontology that describes an entity as a specific amino acid sequence of a protein, rather than the biological concept of that protein (which ignores amino acid variations).

Outcomes and recommendations

Outcome: Identifier Mapping is central to data integration but cannot be done properly without ontologies and suitable provenance models. Furthermore, the meaning of mappings can be ontologically defined using scientific lenses. The latter requires biologists and chemists to accurately specify what specific concepts are used as core entities in databases and to have a shared understanding of these.

Recommendation: set up an open science, community IMS and provide means to support uptime and maintenance of such a community service. This infrastructure should consist of both the mapping service (IMS instance) and open and FAIR availability of link sets.

Action: set up a public IMS service for the European life sciences community, for example at http://ims.bridgedb.org/

Outcome: Identifier Mapping can be extended to IRIs for use in linking semantic web databases and provided as an independent module in the integration platform. That suggests that a community-driven identifier mapping service can benefit many projects and should be considered a general need to enable a European linked data network.

Recommendation: provide guidance about how link sets and scientific lenses can and should be applied to data sources, and show how more traditional ID mapping use cases can be automatically derived from these.

Action: Generate best practice documentation on the ELIXIR Knowledge Hub

Outcome: Each newly introduced data set can be expected to require at least one new link set specifying identifier mappings. Because different data sets use different entity concepts, unifying identifiers like the InChI are only of limited use. The creation of link sets is a significant amount of the work of the data integration. Chemical structure curation, normalization, identification, and classification is essential and should involve EUOpenScreen, EUToxRisk, and OpenRiskNet.

Recommendation: Establish an open science, community-driven chemistry registration service, accepting SMILES and SD files, based on established open source tools (e.g. the Chemistry Development Kit) to report structural errors (e.g. missing stereochemistry), normalize structures, classify structures (e.g. using the ChEBI ontology), and generate identifier mappings and link sets for BridgeDb/IMS, capturing stereochemical, charge state, and tautomer relations (scientific lenses).

Action: set up a task force to get together stakeholders and open modules for the various steps

CORBEL Roadmap

Thus far we have bootstrapped our identifier review through selected use cases. We have generated a framework and checklists with which we will continue to audit the partners, their RIs and Use Cases. The roadmap also focuses on greater knowledge dissemination of best practices and services

available. Checklists and other guidelines will be added to the ELIXIR Interoperability Platform Knowledge Hub along with pointers to other identifier initiatives. The roadmap also proposes engagements with international initiatives many of which have already been initiated.

1. Review, Revise and Re-apply checklists and case studies

Our work on checklists and selected case studies ran in parallel. We next need to systematically reapply the checklists to the case studies presented here to iteratively improve both the lists and the case study reports, and sharpen tasks.

Success Metric: Documented case studies, documented checklists

2. WP4 Use Cases Review

Our work to date identified concrete recommendations and actions based on our selected case studies. As WP4 has now completed its call for proposals and is short listing these we will engage the participants and will use the checklist to audit the WP4 Use Case needs.

Success Metric: Workshops attended and WP4 Use Cases documented

3. Communications

Several case studies struggled with where to find services and knowledge of how to do things.

- Ensure the services are present in the bio.tools life sciences services registry
- Make recommendations to bio.tools and other relevant repositories for markup and search that is useful for identifier services
- Deliver information on the ELIXIR Knowledge Hub for services and resources based on quality and provide CORBEL specific information for cross BMS RI users
- Deliver information on the ELIXIR Knowledge Hub on best practices for identifier formats, assignment, design and updating of mapping services.
- Provide recommendations on patient identifier assignment and expected resolution for consented and/or managed access databases. Patient identifiers are only in CORBEL's scope if they resolve to some global service.
- Provide guidelines and a list of (a few) recommended namespaces for each identifier domain, including but not limited to genes, variations (SNPs and larger mutations), proteins, cell lines, drugs and molecular interactions.

This highlights a lack of information on which services are available and why these were selected by a project. The checklist, future versions and best practices to be shared on the ELIXIR Knowledge Hub and will be published for community use. A number of guides have already been identified and more will emerge from roadmap tasks 1 and 2. A guide structure has been defined and the infrastructure for the Knowledge Hub set up with ELIXIR Hub. A series of training materials are needed to make the process of choosing an identifier management service clearer. This could be handled as FAQ - e.g. what do I do when my entity is not identified, e.g. What are the criteria for choosing a Gene name/id mapping service? registered on the ELIXIR TeSS Training Portal. The CORBEL partner resources services registered in bio.tools and the advocating of improvements to bio.tools search and curation functionality, and the content kept up to date.

Success Metrics:

- **Content on the ELIXIR Knowledge Hub for the checklists. A checklist publication and guides on cross infrastructure needs beyond ELIXIR.**

- **Training materials on identifier management services**
- **Services registered in bio.tools and other registries where appropriate.**

4. Service adoption

In several case studies common services had been adopted or desirable services were unavailable in one RI and available in another.

- Extend the representation of CORBEL partner resources in identifiers.org based on audit of BMS RIs using the checklists
- Identify novel resources, determine if resolutions services exist and add to identifiers.org if not
- Identify lists of services - for example Open PHACTS chemistry identifier services and gene name/identifier conversion services for Image Data Repository - and work with CORBEL partners to test the services recommended for the new datasets.

Success Metric:

- **Resources registered in identifiers.org**
- **Lists of services registered in bio.tools and reviewed for Use Cases.**

5. Identifier Practices on key datasets and datatypes

A shortlist of key data types will be the focus of our harmonisation efforts. Our preliminary audit of our case studies has indicated a shortlist of key data types and datasets to be the initial focus of our harmonisation efforts:

Dataset specific identifier practices for ENA and Pangaea

- Granularity/consistency: e.g. ENA terms/concepts “Study” and “Project” seem to be used inconsistently
- ENA and Pangaea should allow minting (creation) of identifiers in advance of submission to allow the submitter to establish the correct double cross-linking among the correct entities. The preregistration of samples in the EBI’s BioSamples database and sharing of accessions in the future is another recommendation.
- Content drift: identifiers reference different versions of same data sets (e.g. current entries at EBI and Pangaea are not up-to-date). A data submission broker would take responsibility to offer users, in need of complex distributed submissions, a single entry point and take over the responsibility of managing distribute submissions to resources in different domains (e.g. molecular data broker of the German Federation for Biological Data (GFBio)).

General identifier practices

- Resolve limited-life identifiers such as OSD id with an “Expected Expiration Date” i.e. the expected date from which on the resource availability cannot be guaranteed.

Success Metric:

- **Publication of key data types, datasets and RI using these**
- **Publication of recommendations of improved identifier practices in datasets.**

6. Identifier Services for key datasets and datatypes

The case studies proposed several services:

Identifier mappings and link set services. Several of the case studies had issues with identifier mappings, which is not surprising as this is the prime mechanism for dealing with identifier harmonisation and interoperability. In particular linksets as first class objects with their own metadata and services is called for.

- Define metadata for linksets (provenance, ontologies)
- Proposed set up an open science, community Identifier Mapping Service and provide means to support uptime and maintenance of such a community service. This infrastructure should consist of both the mapping service (IMS instance) and open and FAIR availability of link sets at e.g. http://ims.bridgedb.org/

Chemistry services: Propose an open science, community-driven chemistry registration service, accepting SMILES and SD files, based on established open source tools (e.g. the Chemistry Development Kit) to report structural errors (e.g. missing stereochemistry), normalize structures, classify structures (e.g. using the ChEBI ontology), and generate identifier mappings and link sets for BridgeDb/IMS, capturing stereochemical, charge state, and tautomer relations (scientific lenses)

Biobank services: A possible of a global (voluntary) participant biobank ID, possibly given at the time of signing consent.

We will review with the WP4 Use Cases and make available a prioritised revised list of services (adhering to our four principles) and a plan for their execution.

Success Metric:

- **Publication of key identifier services**
- **Plan for prioritised services.**

7. Community Initiative engagement

There are several initiatives that we need to be sure we engage in effectively. We propose to continue our engagement with the following: Bioschemas.org, BioCADDIE/DCIP, PrefixCommons, Resource Identification Initiative (RRID). We will regularly review our engagement strategy.

Success Metrics:

- **Meetings attended, contributions made**
- **Results disseminated to CORBEL partners through Knowledge Hub.**

8. Ontology concerns - Task 6.2, outside the remit of Task 6.1

There are several case studies which cited identification and identifier management services related to ontology terms as a challenge. These are mostly not identifier issues, but arise as they relate to the consistent labelling of entities and their properties. These are issues to address in task 6.2.

- Adding terms on-demand to established ontologies
- Recording non-exact mappings to existing terms in an ontology
- Ontology change detection with respect to existing annotations
- Need for identifiers for new ontology terms

The process of determining if a term is already identified, is a synonym of an existing term, choosing terms where there are multiples clearly represents a challenge for the users. In Task 6.2 we have started to work with communities in this space to improve the available toolkit and have similar requirements from industrial users. All ontological use cases will therefore be addressed in deliverable 6.3 (Month 40) and a prototype ontology mapping service is under development. Further an ontology usage checklist developed for CORBEL partners which complements the identifiers

checklists here. Specifically, we will address best practice in identification of existing ontology terms, creation of new terms, mapping terms and supporting services. These services are typically not deployed per RI but the toolkit under development for D6.3 will support this use and is already deployed locally by some organisations.

Success Metric:

- **Ontology use cases satisfied in D6.3 and ontology usage checklist**

Plan and Milestones

We have 39 PM available for the task distributed over partners. The project runs 1st Sept 2015 - 31st August 2019 (Table 23). The workplan for the rest of the project is shown in Table 22 and future milestones in Table 23.

Schedule

- MS6.1 Review of identifier schemes and standards, identifier interoperability maps,and proposed harmonisation strategy, (D6.1)
- MS6.4 Delivery of sustainable cross-infrastructure identifiers service(s) deployed for core services in the pilots M24 (August 2017)
- MS6.7 Access to identifier related services available from the ELIXIR service registry M36 (D6.2) - August 2018

	Mar-May 2017 M18-M20	Jun-Aug 2017 M21-M23	Sep-Nov 2017 M24-M26	Dec-Feb 2017-2018 M27-M29	Mar-May 2018 M30-M32	Jun-Aug 2018 M33-M35	Sep-Nov 2018 M36-M38	Dec-Feb 2018-2019 M39-M41	Mar-May 2019 M42-M44	Jun-Aug 2019 M45-47
1 Review and reapply checklists										
Documented case studies										
Documented checklists										
2 WP4 Use Cases Review										
WP4 Use Cases documented										
3 Know-how communications										
Services registered in bio.tools										
Knowledge Hub guides										
Training material on TeSS										
Checklist publication										
4 Service adoption										
Resources registered in identifiers.org										
Services reviewed for Use Cases										
5 Identifier practices on datasets/types										
Publication of key datasets/types										
Publication of recommendations										
6 Identifier services on datasets/types										
Publication of key identifier services										
Plan formulated for prioritised services										
Start up the plan										
7 Community initiatives engagement										
Bioschemas.org										
BioCADDIE/Force11 DCIP identifier res.										
Prefix Commons										

Table 22. Gantt chart of draft workplan

M21 June 2017	Review and revision of checklists and case studies Published on ELIXIR Knowledge Hub
M23	Services registered in bio.tools

Sep 2017	First tranche of Knowledge Hub Guides
M29 Feb 2018	WP4 Use Cases fully documented (workshops leading up to this)
M30 May 2018	Publication of key identifier services and plan for prioritised services
M36, Sept 2018	Access to identifier related services available from the ELIXIR service registry

Table 23. Key Milestones for the next reporting period.

Acknowledgements

We wish to acknowledge the contributions of the following people:
Julie McMurry (Monarch Consortium, USA), Anita Bandrowski (UCSD, USA), Elena Bravo (Istituto Superiore di Sanità, Rome, Italy), Tim Clark (FORCE11, Harvard Medical School, USA), Herbert Van de Sompel (Los Alamos National Laboratory, USA), Dorota Skowronska-Krawczyk (UCSD, USA), Randi Vita (La Jolla Institute for Allergy & Immunology, USA).

References

- Allan C, Burel JM, Moore J, Blackburn C, Linkert M, Loynton S, Macdonald D, Moore WJ, Neves C, Patterson A, Porter M, Tarkowska A, Loranger B, Avondo J, Lagerstedt I, Lianas L, Leo S, Hands K, Hay RT, Patwardhan A, Best C, Kleywegt GJ, Zanetti G, Swedlow JR. OMERO: flexible, model-driven data management for experimental biology. Nat Methods. 2012 Feb 28;9(3):245-53. Doi: 10.1038/nmeth.1896. PubMed PMID: 22373911; PubMed Central PMCID: PMC3437820.
- Azzaoui K, Jacoby E, Senger S, Rodríguez EC, Loza M, Zdrazil B, Pinto M, Williams AJ, de la Torre V, Mestres J, Pastor M, Taboureau O, Rarey M, Chichester C, Pettifer S, Blomberg N, Harland L, Williams-Jones B, Ecker GF. Scientific competency questions as the basis for semantically enriched open pharmacological space development. Drug Discov Today. 2013 Sep;18(17-18):843-52. Doi: 10.1016/j.drudis.2013.05.008. Review. PubMed PMID: 23702085.
- Colin Batchelor, Christian Y. A. Brenninkmeijer, Christine Chichester, Mark Davies, Daniela Digles, Ian Dunlop, Chris T. Evelo, Anna Gaulton, Carole Goble, Alasdair J. G. Gray, Paul Groth, Lee Harland, Karen Karapetyan, Antonis Loizou, John P. Overington, Steve Pettifer, Jon Steele, Robert Stevens, Valery Tkachenko, Andra Waagmeester, Antony Williams, Egon L. Willighagen, *Scientific Lenses to Support Multiple Views over Linked Chemistry Data,* (2014), Lecture Notes in Computer Science Vol 8796, Proc 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I, Springer International Publishing, pp:98--11, Doi:10.1007/978-3-319-11964-9_7
- Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, Cooper LD, Courtot M, Csösz S, Cui H, Dahdul W, Das S, Dececchi TA, Dettai A, Diogo R, Druzinsky RE, Dumontier M, Franz NM, Friedrich F, Gkoutos GV, Haendel

M, Harmon LJ, Hayamizu TF, He Y, Hines HM, Ibrahim N, Jackson LM, Jaiswal P, James-Zorn C, Köhler S, Lecointre G, Lapp H, Lawrence CJ, Le Novère N, Lundberg JG, Macklin J, Mast AR, Midford PE, Mikó I, Mungall CJ, Oellrich A, Osumi-Sutherland D, Parkinson H, Ramírez MJ, Richter S, Robinson PN, Ruttenberg A, Schulz KS, Segerdell E, Seltmann KC, Sharkey MJ, Smith AD, Smith B, Specht CD, Squires RB, Thacker RW, Thessen A, Fernandez-Triana J, Vihinen M, Vize PD, Vogt L, Wall CE, Walls RL, Westerfeld M, Wharton RA, Wirkner CS, Woolley JB, Yoder MJ, Zorn AM, Mabee P. Finding our way through phenotypes. PLoS Biol. 2015 Jan 6;13(1):e1002033. doi: 10.1371/journal.pbio.1002033. PubMed PMID: 25562316; PubMed Central PMCID: PMC4285398.

- Fenner, M. et al (2016). A Data Citation Roadmap for Scholarly Data Repositories doi: https://doi.org/10.1101/097196
- Hill E. Announcing the *JCB DataViewer*, a browser-based application for viewing original image files. *The Journal of Cell Biology*. 2008;183(6):969-970. doi:10.1083/jcb.200811132.
- Iudin A, Korir PK, Salavert-Torres J, Kleywegt GJ, Patwardhan A. EMPIAR: a public archive for raw electron microscopy image data. Nat Methods. 2016 May;13(5):387-8. doi: 10.1038/nmeth.3806. PubMed PMID: 27067018.
- Jupp S, Malone J, Burdett T, Heriche JK, Williams E, Ellenberg J, Parkinson H, Rustici G. The cellular microscopy phenotype ontology. J Biomed Semantics. 2016 May 18;7:28. doi: 10.1186/s13326-016-0074-0. PubMed PMID: 27195102; PubMed Central PMCID: PMC4870745.
- Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012 Jan;40(Database issue):D580-6. doi: 10.1093/nar/gkr1097. PubMed PMID: 22140103; PubMed Central PMCID: PMC3245029.
- Karapetyan K, Batchelor C, Sharpe D, Tkachenko V, Williams AJ. The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets. J Cheminform. 2015 Jun 19;7:30. Doi: 10.1186/s13321-015-0072-8. PubMed PMID: 26155308; PubMed Central PMCID: PMC4494041.
- Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research*. 2016;44(Database issue):D488-D494. doi:10.1093/nar/gkv1024.
- McMurry, J. et al (2014) BioMedBridges deliverable Identifier Best Practice and Supporting Tools
- McMurry, J. , & 40 additional authors (see file). (2016). Identifiers for the 21st century: How to design, provision, and reuse identifiers to maximize data utility and impact. http://doi.org/10.5281/zenodo.163459
- Mohammad F, Flight RM, Harrison BJ, Petruska JC, Rouchka EC. AbsIDconvert: an absolute approach for converting genetic identifiers at different granularities. BMC Bioinformatics. 2012 Sep 12;13:229. doi: 10.1186/1471-2105-13-229. PubMed PMID: 22967011; PubMed Central PMCID: PMC3554462.
- Ohno-Machado, L., Alter, G., Fore, I., Martone, M., Sansone, S.-A., Xu, H. (2015): bioCADDIE white paper - Data Discovery Index. Figshare
- Orloff DN, Iwasa JH, Martone ME, Ellisman MH, Kane CM. The cell: an image library-CCDB: a curated repository of microscopy data. Nucleic Acids Res. 2013 Jan;41(Database issue):D1241-50. doi: 10.1093/nar/gks1257. PubMed PMID: 23203874; PubMed Central PMCID: PMC3531121.

- Van Brabant B, Gray T, Verslyppe B, Kyrpides N, Dietrich K, Glöckner FO, Cole J, Farris R, Schriml LM, De Vos P, De Baets B, Field D, Dawyndt P; Genomic Standards Consortium.. Laying the foundation for a Genomic Rosetta Stone: creating information hubs through the use of consensus identifiers. OMICS. 2008 Jun;12(2):123-7. doi: 10.1089/omi.2008.0020. PubMed PMID: 18479205.
- van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010 Jan 4;11:5. Doi: 10.1186/1471-2105-11-5. PubMed PMID: 20047655; PubMed Central PMCID: PMC2824678.
- Waagmeester A, Kutmon M, Riutta A, Miller R, Willighagen EL, Evelo CT, Pico AR. Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. PLoS Comput Biol. 2016 Jun 23;12(6):e1004989. doi: 10.1371/journal.pcbi.1004989. PubMed PMID: 27336457; PubMed Central PMCID: PMC4918977.
- Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today. 2012 Nov;17(21-22):1188-98. doi: 10.1016/j.drudis.2012.05.016. Review. PubMed PMID: 22683805.
- Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, Hastings J, Chen B, Wild DJ. The ChEMBL database as linked open data. J Cheminform. 2013 May 8;5(1):23. doi: 10.1186/1758-2946-5-23. PubMed PMID: 23657106; PubMed Central PMCID: PMC3700754.

Delivery and schedule

The delivery is delayed: Yes, see Grant Agreement Amendment AMD-654248-22
Reason: Change of personnel (twice) in critical roles.

Related documents

CORBEL WP6 DoW

Appendices

Appendix 1. Summary of Community Activities

Initiatives

We actively participate in the following selected international activities

bioCADDIE / FORCE11

The NIH BD2K bioCADDIE project are developing a data discovery index (DDI) prototype to index data that are stored elsewhere. The DDI will play an important role in promoting data integration through the adoption of content standards and alignment to common data elements and high-level schema. The DataMed prototype(v1.5)[40] aims allow users to search for and find data across different repositories in one space.

ELIXIR members are partners of bioCADDIE and members of the ELIXIR Interoperability Platform. DATS, the data model for tagging datasets, includes metadata associated with identifiers (Ohno-Machado et al, 2015).

FORCE11 is a grassroots Not For Profit US-based association that aims to bring about a change in modern scholarly communications through the effective use of information technology. It runs a popular annual international conference, 31 community working groups and projects funded by grant awards. It produced the influential Joint Declaration of Data Citation Principles (JDDCP)and a sister Software Citation Principles, and is the home of the Resource Identification Initiative[41].

First, bioCADDIE/FORCE11 are funded under the USA NIH BD2K programme to Pilot to implement the Joint Declaration of Data Citation Principles (JDDCP) with Data Repositories and Publishers. A roadmap for data repositories (Fenner et al, 2016) and publishers have been published and we aim to promote this roadmap across CORBEL. Second, as a result of our cooperations with this project, a Joint project was brokered by FORCE11 between the CDL and EBI to develop a common approach for global resolution of locally-assigned accession numbers, based on a shared registry of defined resource prefixes and provider codes. We reported this in the Harmonisation section and we attended the key workshop in 2 June 2016.

bioCADDIE/FORCE11 Data Citiation Implementation Pilot is lead by Dr Tim Clark, who collaborated on this deliverable. We are pleased to acknowledge his contribution.

Resource Identification Initiative (RRID)

The Resource Identification Initiative (RRID)[42] is designed to help researchers sufficiently cite the key resources used to produce the scientific findings reported in the biomedical literature. A diverse

[40] https://datamed.org
[41] https://www.force11.org/group/resource-identification-initiative
[42] https://scicrunch.org/resources

group of collaborators are leading the project, including the Neuroscience Information Framework[43] and the Oregon Health & Science University Library, with the support of the National Institutes of Health and the International Neuroinformatics Coordinating Facility[44].

The Resource Identification Portal, supports NIH's new guidelines for Rigor and Transparency in biomedical publications. Authors are instructed to authenticate key biological resources: Antibodies, Model Organisms, and Tools (software, databases, services), by finding or generating stable unique identifiers.

PrefixCommons

PrefixCommons is a cross-cutting framework to aggregate, document, and harmonize identifier prefixes from multiple sources, most notably Identifiers.org, the OBOFoundry, Bioportal, and prefix.cc. Prefix consistency and lack of global collisions are helpful to aid human understanding; however, neither of these features is required in order to make the existing situation better than it is.

PrefixCommons have begun to develop tools for composing and validating sets of prefixes as used in different contexts. It also provides a framework for iteratively approaching a more unified standard for minting new identifier strategies.

Prefix Commons is lead by Dr Julie McMurry who collaborated on this deliverable. We are pleased to acknowledge her contribution.

Bioschemas.org

The FAIR principles highlight Findability, and its sister “citability”, as the first steps to accessible data. Bioschemas is a flagship project of the ELIXIR Interoperability Platform to tackle the F of FAIR data.

Schema.org provides a way to add semantic markup to web pages by describing ‘types’ of information, which then have ‘properties’. This structured information makes it easier to discover, index, cite, collate and analyse distributed data by general search engines and specialist harvesters. DataCite has recently added support[45] for schema.org in JSON-LD format to DOI content negotiation, and embedded in search results on DataCite Search. Schema.org metadata can be converted into DataCite metadata and used with the DataCite Metadata Store, DataCite’s DOI registration and management service.

Bioschemas aims to improve data interoperability in life sciences by encouraging people in life science to use schema.org markup, so that their websites and services contain consistently structured information. In 2017 Bioschemas aims to produce a collection of specifications that provide guidelines to facilitate a more consistent adoption of schema.org markup within the life sciences for datasets, and hence for the metadata for the identifiers of those datasets, and to support identifier indexing and resolution.

ELIXIR has funded a Bioschemas pilot activity focusing on data repositories, datasets, samples, plant phenotypes, and protein annotations. BBMRI are partners.

[43] https://neuinfo.org//
[44] https://www.incf.org/
[45] https://doi.org/10.5438/0000-00CC

Other Relevant Initiatives

Name	Activity details
Research Data Alliance https://www.rd-alliance.org Persistent Identifier Interest Group https://www.rd-alliance.org/groups/pid-interest-group.html	RDA PID IG Aims to synchronize identifier-related efforts, address important and emerging PID-related topics and coordinate activities, including appropriate RDA Working Groups, to practically solve PID-related issues from the engaged communities. *Unclear if this group is still active*
RDA/WDS Publishing Data Services WG and the The Scholarly Link Exchange Workgroup.	The RDA/WDS WG focused on a one-for-all cross-referencing service for the links between data and publications. Its follow-on, the Scholarly Link Exchange Working group, aims to enable a comprehensive global view of the links between scholarly literature and data. The working group will leverage existing work and international initiatives to work towards a global information commons by establishing: ● Pathfinder services and enabling infrastructure ● An interoperability framework with guidelines and standards (see also www.scholix.org) ● A significant consensus ● Support for communities of practice and implementation
Signposting the Scholarly Web, http://signposting.org	PIDs need to be used to achieve their intended persistence. Signposting conveys the PID in the HTTP Link response header of all resources identified by the PID, including the landing page and content resources such as "the PDF" and "the dataset". This allows tools, such as citation managers, to auto-discover and use the PID. Similarly, the HTTP Link header can be used to allow tools to auto-discover, from the landing page, which resources are part of a PID-identified object. These and other uses of the HTTP Link header to achieve a coarse yet meaningful level of interoperability in the scholarly communication system.
The DONA Foundation	Administers and maintains the stable operation of the Global Handle Registry (GHR) along with multiple parties around the globe known as the Multi-Primary Administrators (MPAs). Responsibility for the GHR, previously held solely by CNRI in Reston Virginia USA, was transferred to the DONA Foundation in May, 2014. Since then, five MPAs have been authorized and credentialed by DONA to provide global handle services based on their credential. New organizations are currently in the process of being considered for authorization as future MPAs by the DONA Board of Directors.
Data Identifiers in Publications	
CoBRA, Citation of BioResources in journal Articles.	Part of the BRIF (Bioresource Research Impact factor) initiative, a guideline for reporting bioresource use in research articles.The guideline aims to improve the quality of bioresource reporting and will allow their traceability in scientific publications, thus increasing the recognition of bioresources' value and relevance to research. Bravo E, Calzolari A, De Castro P, *et al.* Developing a guideline to standardize the citation of bioresources in journal articles (CoBRA). *BMC Medicine* 2015;13:33(doi: 10.1186/s129 16-015-0266-y)

THOR https://project-thor.eu/	A 30 month project funded by the European Commission under the Horizon 2020 programme to establish seamless integration between articles, data, and researchers across the research lifecycle. Partners include DataCite
openAIRE https://www.openaire.eu/	OpenAIRE supports the implementation of the EC and ERC Open Access policies and links the aggregated research publications to the accompanying research and project information, datasets and author information.
Citation	
CRediT http://docs.casrai.org/CRediT	CRediT (Contributor Roles Taxonomy) is a CASRAI activity that brings together a diverse set of stakeholders with a common interest in better understanding and communicating the different kinds of contributor roles in research outputs. CRediT is an open standards activity aligned with the OpenStand principles to which CASRAI is a signatory.
The Global Research Identifier Database (GRID)	A free (CC-BY), manually curated database of organizations associated with research.

Table 26. Other relevant initiatives

Outreach Activities Log

Auditing of Biobank	16th Feb 2017 Service Pilot	Consultation
Estonian Genome Center	16th Feb 2017	Consultation
Aria Workshop (Instruct/iNEXT/CORBEL Open Call)	3rd Feb 2017	CORBEL partners engagement
ELIXIR Interoperability Platform All Hands Face to Face	30thJan - 1st Feb 2017, Amsterdam, the Netherlands	CORBEL partners engagement
UK IVF clinic initiative	13th Jan & 24th Jan 2017 UK IVF clinic initiative around technical specifications for image/data sharing for research and standardisation purposes. Grant proposal ongoing.	Consultation
ELIXIR WP7, Plants use case	25th Nov, 30th Nov 2016 Initial discussion around Identifier needs	Consultation
SSBA (IAPR swedish branch)	21-22th Nov 2016 SSBA (IAPR swedish branch) - workshop on challenges in image computerised analysis.	Outreach
1st Information exchange with BRIF Initiative	11th Nov 2016 http://www.hsern.eu/index.php/news/show/cobra-citation-of-bioresources-in-journal-	Engagement

	articles 3 meetings (Identifiers WG and external stakeholders), focused around CORBEL 6.1 (this deliverable). Engagement with Finnish National Library/IETF/URNBIS WG around use of URN namespace and its global resolvability: https://tools.ietf.org/html/rfc3151 https://datatracker.ietf.org/wg/urnbis/documents/	
PIDappalooza conference http://pidapalooza.org	8th Nov 2016 Persistent identifiers festival, Nov 2016. Nick Juty (EBI) and Tim Clark (Harvard Medical School/Force11) presented talk	Outreach Presentation
NETTAB Hackathon and EMBnet/NETTAB workshop	24-25th Oct 2016 (https://www.elixir-europe.org/events/elixir-bioinformatics-hackathon-nettab-2016 and https://www.elixir-europe.org/events/charme-embnet-and-nettab-2016-workshop)	Outreach Presentation
CORBEL annual meeting	18-19th Oct 2016	CORBEL partners engagement
MIUF	12th Oct 2016 http://www.ecrin.org/event/corbel-meeting-survey-outcomes	CORBEL partners engagement
ELIXIR Rare Disease BYOD Workshop, Rome	29-30th Sept 2016 https://docs.google.com/document/d/1p_QRoK_5ra43QY5DHVhkTxu-UiOmoi5-M-JvKdJ6y5k/edit?ts=57ecd444#	CORBEL partners engagement
MIABIS/OBIB	27th Sept 2016 Integration Sample definition, 1st meeting	CORBEL partners engagement
New Scientist Live, ExCel, London	24th Sept 2016	Outreach
Software faster	20th Sept 2016 https://www.elixir-europe.org/events/software-faster-months-minutes-elixir-training-course	Training
Europe Biobank Week	13-16th Sept 2016 http://europebiobankweek.eu/	Outreach
NSF Workshop Data and Software Citation http://www.software4data.com	6-7 June 2016, Harvard Medical School, Boston, USA Invited talk: "FAIR Software (and Data) Citation: Europe, Research Object Systems,	Outreach Presentation

	Networks and Off the Shelf Infrastructure"	
bioCADDIE/Force11 DCIP Identifiers Workshop	2nd June 2016, Harvard Medical School, Boston, USA	Collaboration
ELIXIR-NIH BD2K Bioschemas Workshop	10th March 2016, Barcelona, Spain Lightning talk: Identifier Mapping	Outreach Presentation
ELIXIR All Hands 2016	8 – 10th March 2016, Barcelona, Spain	CORBEL partners engagement
ELIXIR Software Working Group	14th December 2015, Hinxton, UK	CORBEL partners engagement
ELIXIR EXCELERATE WP8 Rare Disease Kickoff	23rd November 2015	Consultation
ELIXIR EXCELERATE kick-off	8-9th December 2015, Cambridge, UK, ran breakout session	CORBEL partners engagement
ELIXIR EXCELERATE WP8 Rare Disease Kickoff	23rd November 2015	CORBEL partners engagement
CORBEL Kick off	18-19th November 2015 , Hinxton, UK	CORBEL partners engagement

Table 27. Outreach activities in the reporting period March 2016 - February 2017

7 meetings with groups from Force11 initiative	https://www.force11.org/group/dcip/eg2identifiers https://www.force11.org/group/resource-identification-initiative
1 meeting with Collaborative Drug Discovery (CDD)	https://www.collaborativedrug.com/
3 meetings with ELIXIR Linked Data and Ontologies task Force	https://docs.google.com/document/d/1TMRpoa7ahYYwv4qAUMU5uUTawBihn3YpIL3vDlBHj9Y/

Table 28. Meetings attended in the reporting period March 2016 - February 2017

Outcomes

Gathering use cases, interoperability bugs and user requests in backlog of work items on ELIXIR Intranet, and setting up task structure for initiating work. Gathering two community groups; identifiers directed in ELIXIR Identifiers Working Group as expert Advice Panel and community group for continued survey of identifiers and interoperability needs and as User Panel. Stub for information, intended for continued update, setup on ELIXIR Knowledge Hub in collaboration with ELIXIR Web Team for backend.

Appendix 2. Identifier services

This list is provided here for convenience and was extracted from previous work in the BioMedBridges project with some updates.
Identifiers are created, maintained, and applied using software summarised below:

- **Identifier creation services** *create* a URI or other identifier given certain parameters. For example, URIGen is an online tool for managing the creation of URIs for collaborative ontology development projects. InChI generator converted SMILES or chemical structure drawings to InChIStrings and InChIKeys, but has since be decommissioned in favour of Chemspider.
- **Identifier converter services** *convert* one form of an identifier to another. For example NCBI provide a converter[46] that takes an identifier for an article that is in PMC (PubMed Central) and returns other identifiers associated with the article; including the PMID, PMCID, Manuscript ID or DOI. Mohammad et al. 2012 compiled list of converter tools for genomic identifiers[47] which should be revisited and updated.
- **Identifier resolution services** *resolve* an identifier, i.e. given an identifier, they will return a representation of the entity. For example, the InChI resolver took an Input InChIString or InChIKey and looked up the associated chemical structure (service since deprecated in favour of Chemspider). Resolvers are typically Web applications that run in your browser. The simplest resolver therefore is simply the browser itself which resolves a URL identifier to an informative Web page.
- **Identifier mapping services** (**identity mapping** or **identity binding** services) *map* identifiers on entries in one resource to those in another, in order to assign equivalence to entries or otherwise link two resources. Methods are direct or indirect; indirect method are often represented as semantic-free cross references or XREFs. Direct methods map by comparing identifier values, including those that are the database accession and/or which provide a cross-reference to another resource. For example, the Protein Identifier Mapping Service[48] resolves protein identifiers across multiple databases that correspond to the same protein. *Indirect methods* do not rely on identifier values to achieve the same ends, for example, a mapping of equivalent concepts in two ontologies may be achieved through comparison of terms and synonyms that are associated with the concepts. Mappings with provenance, different types of mappings, derived and equivalent entities, exact and non exact synonyms. In some cases XREFs, which are pointers to related entities, are sufficient e.g. when the entities are of the same type. However, XREFS can be used to map entities of different types e.g. genes to proteins and this can present ambiguities for some applications.
- **Catalogues of identifier types** describe the types of available identifiers in a given domain. For instance, Identifiers.org[49], the Gene Ontology[50], NCBI and EDAM[51], describe metadata about the identifier type itself,e.g. preferred name, synonyms, definition, example IDs.
- **Categorising resources according to entity types**. Resources such as NIF and BioSharing organise information according to the type of biological entity (Eg. gene, protein[52] etc), using terms from several ontologies. However, there is no broad consensus for an ontology of biological entities to use across different resources and d domains; harmonisation would therefore be beneficial. Strongly typed biological entities would also be beneficial for catalogues of identifier types (above).

[46] http://www.ncbi.nlm.nih.gov/pmc/pmctopmid/
[47] http://www.biomedcentral.com/1471-2105/13/229/table/T2
[48] http://www.ebi.ac.uk/Tools/picr/
[49] www.identifiers.org/
[50] http://wiki.geneontology.org/index.php/Identifiers
[51] http://edamontology.org/
[52] http://www.biosharing.org/biodbcore/?q=&selected_facets=domains_exact:protein

- **Annotation aggregation tools** also provide access to combinations of the services listed above. They compile existing identifiers together with higher-level structures related to an entity. *e.g.*, the Bioconductor[53] platform has extensive facilities for mapping between microarray probe, gene, pathway, gene ontology, and other annotations. Bioconductor can easily access NCBI, Biomart, UCSC, and other sources[54] *e.g.* the bioDBnet[55] **db2db** tool handles all the conversions from one database identifier to another.includes tools to report all available information for an identifier, interconvert identifier formats, and converts molecular sequence identifiers for one organism into the corresponding identifiers of a different organism. MyGene.info[56] provides REST web services to query/retrieve gene annotation data. The Monarch system[57] enables the aggregation, provenance, and currency of hundreds of external resources, while integrating them to ontologies for phenotypes, diseases, genotypes, and anatomy.
- **Identifier platforms,** such as the domain-agnostic Handler system[58] provide some combination of the above services. For the biomedical domain, Identifiers.org[59] has the broadest adoption; Identifiers.org is built on the MIRIAM registry and provides direct access to data at one specified (but modifiable) resolving location. There is a fundamental difference between identifiers which point to experimental datasets (GenBank/ENA/DDBJ, PRIDE, etc) and identifiers which point to a current understanding of a biological concept (Ensembl Gene IDs, UniProt IDs). The life cycles and management of these identifiers are very different. In these cases eg (INSDC, wwPDB), alternative resolving locations may also be accessed via Identifiers.org (so long as they are recorded); however this current approach could be revisited in the future.

Identifier management services for life sciences entities are typically community generated, examples of these are provided in Table 29.

Example generic services relevant to CORBEL (Table 30) there are many of these and we do not attempt to provide complete context as our previous work addressed this (Van Iersel et al., 2010).

Service Name	Service Type	Community
www.identifiers.org[60] (Juty et al., 2012)	Life science identifier resolution	Generic life sciences
BridgeDb[61]	Identifier mapping	Gene, protein, metabolite
UniChem https://www.ebi.ac.uk/unichem/	Identifier mapping	Chemistry

[53] http://www.bioconductor.org/
[54] http://www.bioconductor.org/help/workflows/annotation/
[55] https://biodbnet-abcc.ncifcrf.gov/
[56] http://mygene.info/
[57] http://monarchinitiative.org/
[58] http://en.wikipedia.org/wiki/Handle_System
[59] http://identifiers.org
[60] Identifiers.org and MIRIAM Registry: community resources to provide persistent identification
[61] http://bridgedb.org

RRID https://scicrunch.org/resources	Identifier assignment for research materials	Life sciences
Ontology Lookup service	Identifier resolution and mapping service (mapping service under construction)	Life sciences ontologies
Protein Identifier Mapping Service http://www.ebi.ac.uk/Tools/picr/	resolves protein identifiers across multiple databases that correspond to the same protein	Life sciences
Converter tools for genomic identifiers http://www.biomedcentral.com/1471-2105/13/229/table/T2	convert genetic identifiers at different granularities	Life Sciences
NCBI mapper https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/	Maps identifier for an article that in PMC; returns other identifiers associated with the article	Generic Life Sciences
Biodbnet db2db https://biodbnet-abcc.ncifcrf.gov/db/db2db.php	db2db allows for conversions of identifiers from one database to other database identifiers or annotation	Generic Life Sciences

Table 29. Life Sciences identifier management services

Service Name and URL	**Service Type**
EZID, http://ezid.cdlib.org/ The EZID service is currently the main way to add to N2T's database of names, descriptions, and forwarding addresses. Resolver replication for N2T is underway.	Assignment of DOI/ARK identifiers
The Name-to-Thing (N2T) Resolver http://n2t.net/	URL, ARK, DOI, URN, Handle, LSID, etc
ORCID https://orcid.org/	Identifier assignment for researchers
DataCite https://www.datacite.org/	Assignment and resolution of DOI identifiers, mapping between identifiers https://project-thor.readme.io/docs/examples-of-linking-across-identifiers (Handle identifiers)
ePIC http://www.pidconsortium.eu	Identifier assignment, registration and resolution, used by EUDAT

	(Handle identifiers)
CrossRef	Assignment and resolution of DOI identifiers (Handle identifiers)
The Yale Persistent Linking Service (YPLS)	Creation of permanent links or handles that can be resolved to target URLs via Handle.Net Registry (HNR) by CNRI under the DONA Foundation.
Open Funder Registry (FundRef) codes https://www.crossref.org/services/funder-registry/	A freely-downloadable file, lists funders and their unique identifiers.

Table 30. Examples of generic identifier management services relevant to CORBEL