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

Deliverable D6.2

Access to Sustainable cross infrastructure identifier service(s) through ELIXIR service registry

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

Contractual delivery date: 31 August 2018
Actual delivery date: 28 September 2018

Authors of this deliverable: Nick Juty, Carole Goble, Helen Parkinson, Simon Jupp

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

Content

Executive Summary

The CORBEL project operates across 13 individual BMS RIs, each of which have developed over time towards service provision targeted to support their individual user communities. Hence, there are potentially differences in data handling and processing resulting from community specific preferences and customs (for example the deposition of data into preferred repositories), which will differ across infrastructures. These potential disparities will need to be addressed to harmonise data and services across RIs, and is fundamental to enable downstream and wide scale interoperability.

The work described here addresses the specific objective: "Improve interoperability with European e-infrastructures and leverage existing investments these capacities within the biomedical and life science domain", as detailed in the DoW[1]. This work involved input from and collaboration with 13 RIs, with 36 submitted project proposals (to date), and engaged numerous service providers. We detail below the work, how it was conducted, as well as providing the feedback from the cross-infrastructure partners. We also distilled from these observations a set of recommendations to improve such endeavours in the future.

To achieve "Access to Sustainable cross infrastructure identifier service(s) through ELIXIR service registry", we have identified and promoted the following Interoperability Registry Services, with particular focus placed on components 'a', 'b' and 'd':

a) An identifier registry, with resolution capability (https://identifiers.org/)
b) An Ontology registry, with cross-referencing capability (Ontology Lookup Service, OLS https://www.ebi.ac.uk/ols/index)
c) A Standards registry (FAIRsharing: https://fairsharing.org/)
d) A tools registry (https://bio.tools/)

Project proposals accepted through scientific CORBEL WP4 review were subsequently processed to identify and register project components to the most appropriate registry. Further details and feedback were solicited, from both scientists and service providers, for a prioritised subset of projects deemed most likely to benefit from additional interoperability evaluation. Over the course of this work, we have identified gaps in knowledge that warrant further exploration. Combined with our collective experiences (across WP partners, scientists and service providers), this feedback was used to generate a list of recommendations for future such operations.

Project objectives

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

- Providing an up-to-date investigation into current standards and systems used across partner infrastructures and service providers (Objective 1).

[1] https://docs.google.com/document/d/1D0J_d-scpdzLGDHH-Gic2drMTHec8o06GyxE4Hrwr_g/edit - heading=h.of89p1o34scs

- Generating a set of recommendations to improve interoperability across European e-infrastructures to leverage existing investments and improving capacities within the biomedical and life science domain (Objective 8).

Detailed report on the deliverable

This deliverable targets the identification and amelioration of potential interoperability concerns with respect to projects instigated through CORBEL calls. These calls involve a coalition of partners and services across the other BMS RIs that participate in CORBEL. By focusing on projects, it was envisaged that we would identify different component entities (e.g. ranging from compounds to animal models) that need to be identified within and between infrastructures, which may currently be handled in disparate ways.

The report is organised as follows:

i) **Overview of the Open Call process**.
Description of the call for proposals (with WP4); services available for proposals, partners involved; triage of proposals to exclude those not (currently) appropriate for further analysis with respect to interoperability.

ii) **Assessment of projects with respect to interoperability**.
Identification of project components which could be addressed immediately, as well as noting which could be needed in the future (gap analysis); ensuring project components (databases, controlled vocabularies and ontologies, tools and standards) are registered in an appropriate ELIXIR service registry

iii) **Showcase projects**
Highlighting outputs from this work that would otherwise not have been possible

iv) **Collated feedback from CORBEL partners**
Summary of the key concerns and experiences from both scientists and service providers who have contributed to CORBEL Open Call projects

v) **Recommendations on process improvements**
An evolving list of recommendations that is based upon the experiences of the authors of this report

CORBEL Open Call for project proposals

Led by CORBEL WP4 ('Bioscience Research Use Cases'), an Open Call was launched which led to 30 eligible project proposals being submitted. As a result of this initial call[2] (closing August 2016), 21 proposals were accepted across 4 'Access Tracks' (Table 1).

The distinct access tracks were designed to drive the development of cross-RI scientific connections, through a process based on cross-RI prioritisation. Hence, access tracks cover all aspects of a bioscience translational pipeline: from novel model organisms to genotype-phenotype predictions. The ultimate objective is to work towards a framework enabling transnational open user access across infrastructures.

[2] http://www.corbel-project.eu/1st-open-call.html

Access Track	No. projects accepted
1: Genotype-to-phenotype analysis	3
2: Predictive systems pharmacology for safer drugs and chemical products	10
3: Structure-function analysis of large protein complexes	3
4: Marine Metazoan Developmental Models	2
5: Cross-Access Tracks	3
total	21

Table 1. CORBEL 1st Open Call accepted projects. A more detailed list describing an overview of each project that was selected for further analysis is presented in Appendix A1.1.

Over 90 different technologies and services have been made accessible through these calls, though initially there were significantly fewer. These cover domains such as biological and medical science, and providing access to specific technologies, databases, biological samples and all other tools and resources. The list of providers and their services is available here[3]. An overview of services is given in Table 2. Services range from consultation and advice from experts, through to providing access to biological samples, data analysis tools, and specialist equipment and facilities (e.g. electron microscopy, high resolution imaging).

Category	CORBEL hosting institute (services)
Advanced Imaging Technologies	Advanced Light Microscopy Facility (ALMF), Heidelberg, Germany (Automated image processing, Correlative light electron microscopy, High-throughput microscopy, Multi-modal advanced light microscopy, Super resolution microscopy, Functional imaging)
	European Population Imaging Infrastructure (EPI2), Rotterdam, The Netherlands (Imaging biomarkers, Medical image storage)
	Cell Microscopy Core (CMC), University Medical Centre Utrecht, Utrecht, The Netherlands (Electron microscopy)
	EMBL-Barcelona, Barcelona, Spain (Mesoscopic imaging)
Biobanking and Biomolecular	

[3] http://www.corbel-project.eu/open-call/technologies-services.html

Resources	BBMRI-ERIC, virtual access (BBMRI-ERIC Directory and BBMRI-ERIC Negotiator, Access to biological samples and associated data, Access to additional services of biobanks) Biomedical Research Foundation of the Academy of Athens (BRFAA) Athens, Greece (Biological sample archiving, Genomics and transcriptomics services e.g. sequencing, proteomics and metabolomics services, induced stem cell services)
Clinical Research	ECRIN-ERIC, Paris, France (Project planning advice)
Translational Research	EATRIS Coordination & Support Office, Amsterdam, The Netherlands (Expert advice and support for biomarker validation)
Curated Databases	EMBL-EBI, Hinxton, UK (Curated Databases - ChEMBL, EMPIAR/PDBe, Ensembl Metazoa)
Marine Model Organisms	EMBRC-Fr, OOV: Villefranche-sur-Mer and SBR: Roscoff, France (sea urchin database, amphioxus database, advanced imagine for marine model organisms) Scottish Oceans Institute (SOI), University of St Andrews (marine ecosystem access, microinjection, seal facility, aquarium, jellyfish continuous culture)
Mouse Mutant Phenotyping	German Mouse Clinic (GMC), Helmholtz Zentrum München , Munich, Germany (standardised phenotyping)
Screening	Screening Unit and Medicinal Chemistry group, Leibniz-Forschungsinstitut für Molekulare Pharmakologie, FMP, Berlin, Germany (high throughput screening)
Structural Biology	Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine (CIRMMP), Sesto Fiorentino, Italy (NMR facilities, E.coli and human cell labelling, Circular Dichroism) Instruct Image Processing Center (CSIC), Madrid, Spain (Bioinformatic and computational structural biology tools, EM)
Systems Biology	Molecular Cell Physiology, Vrije University, Amsterdam, The

	Netherlands (Dynamic network modelling and model analysis)
	Division of Theoretical Systems Biology, German Cancer Research Center (DKFZ), Heidelberg, Germany (Support for the development of data-based mathematical models of cellular processes)
	Bioinformatics platform of the Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine, Berlin, Germany (Computational analysis of molecular datasets)

Table 2. Summary of services available for projects responding to CORBEL Open Calls.

In an attempt to ascertain in advance the interoperability needs for individual proposed projects, the project submission form included an optional section asking specific interoperability related questions (Text Box 1).

- Data resources used (any public resource/database, APIs or service that you use to either download or submit data)
- List any existing tools or software that you use to process or analyse the data
- Datasets that you have to merge/integrate. How is this done, what are the challenges?
- Are you aware of any ontologies used to describe the data? (e.g. Gene Ontology)
- Where do you see the challenges from an interoperability point of view?

Text box 1. Questions included in open call forms, targeted to identify interoperability issues.

Subsequent CORBEL Open Calls

The projects accepted were initiated at different times (e.g. negotiations with service providers for facility access), had different durations (simple versus complex projects), and consequently concluded at different times. Following a 2nd Open Call, this has now become a continuous, rolling process. A further list of accepted projects, to date and by track, is given in Table 3.

Access Track	No. projects accepted
1: Genotype-to-phenotype analysis	0
2: Predictive systems pharmacology for safer drugs and chemical products	2
3: Structure-function analysis of large protein complexes	2
4: Marine Metazoan Developmental Models	0
5: Cross-Access Tracks	1 (2&4)
total	5

Table 3. 2nd CORBEL Open Call accepted projects, 1st application period. A more detailed list describing an overview of each project that was selected for further analysis is presented in Appendix A1.1.

Interoperability Assessment of Open Call Projects

The project proposal forms contained optional questions from WP6 on potential interoperability components within projects, but unfortunately were not answered by many applicants. Hence there has been limited opportunity to engage with many of those relevant teams, particularly for the first Open Call since the period between approval and project initiation was often very short, and because of a 'lag' period in WP6 in establishing a process and team to evaluate projects in that respect.

Regardless of whether the interoperability questions were addressed, each of the 26 project proposals was reviewed by at least one member of the ELIXIR interoperability platform[4], in particular focusing on the possible means to improve interoperability. For example, while questions may be unanswered, the application may mention specific databases or vocabularies used in the works, which would need to be listed in the appropriate registry. To achieve this, we have identified and promoted the following Interoperability Registry Services:

1. A Standards registry (https://fairsharing.org/)
2. An Ontology registry, with cross-referencing capability (Ontology Lookup Service, OLS, https://www.ebi.ac.uk/ols/index)
3. An Identifier registry, with resolution capability (https://identifiers.org/)
4. A tools registry (https://bio.tools/)

Of the initial 30 eligible projects submitted, 21 were accepted in the 1st Open Call. Of these, 5 completed during the initiation of WP6 collaboration (leaving 16 active). In addition 5 were submitted over subsequent calls (21 active). All active projects were re-reviewed (Appendix A1.1) to prioritise those demonstrating clear potential or opportunity to improve interoperability aspects. Through this review, 15 projects were subsequently selected for further feedback on experiences and issues faced (indicated in green in Appendix A1.1). Specific questions were targeted to both the service providers, to ascertain the standards and formats that they used in their work, and to the researchers, to determine their scientific process (data deposition practices for example). Excerpts from this further survey are provided in Appendix A2.1. It was deemed worthwhile to follow up with both providers and researchers since, while applicants own the data, and it is important that it becomes available to others where possible; knowing where data is hosted and how to access it (license permitting of course), is something that the providers may not know. The time investment in completing such questionnaires was a concern, particularly for providers who may be involved in multiple proposals. For this reason, we asked for a collective response from service providers, covering all projects in which they were involved. Feedback from this exercise is provided in 'Collated Feedback'.

[4] https://www.elixir-europe.org/platforms/interoperability

Note: At this time we are unable to share the individual project proposals that were submitted, or the reviews and scores assigned. Some limited information of awardees and projects is available from the website: http://www.corbel-project.eu/1st-open-call.html. We do, in this report, provide the persistent project identifier (Proj ID), which will be necessary to link to those proposals if they subsequently become available for public view.

Interoperability Services Registries

Key to the objective to achieve "Access to Sustainable cross infrastructure identifier service(s) through ELIXIR service registry" is the need to have registered databases, vocabularies, standards/formats and tools/software in the most appropriate registry. To that end, we summarise the status of registration with respect to all projects, where information was provided.

Databases and repositories used in projects

Identifiers.org is the preferred registry for listing databases and repositories used in projects. This registry is focused on life science databases, and the overwhelming majority of such databases embedded within specific projects are all registered in Identifiers.org, including databases hosted by service providers themselves, such as ChEMBL. The most commonly utilised repositories within projects include: UniProt, PDB, PubMed, Pubchem, dbSNP, NCBI Gene (often referenced incorrectly as Entrez Gene), GEO, BioStudies and KEGG.

The following database was used within a project, but has not been submitted to Identifiers.org:

Broad Institute (http://exac.broadinstitute.org). (used in Proj 2242)
This is an aggregated 3rd party resource, providing access to multiple data provider records (i.e. multiple databases), for none of which it is the primary provider. The projects using this resource should instead use the appropriate original data provider.

The following databases are unlikely to be suitable, but need to be more extensively analysed:

Proj. 2277: SalivaTec http://salivatec.viseu.ucp.pt/salivatec-db/main.php (incorrect URL in feedback form).
Proj. 2277: OralCard: http://bioinformatics.ua.pt/OralCard/

The following databases, which were not already listed in Identifiers.org, have been submitted, and are pending release:

Proj. 2281: Genomics of Drug Sensitivity in Cancer (https://www.cancerrxgene.org/)
Proj. 2234: Bioportal for Cancer Genomics (www.cbioportal.org)

Vocabularies and ontologies used in projects

The **Ontology Lookup service (OLS)** is the preferred registry for the listing of ontologies used in projects. Gene Ontology was the most used vocabulary across projects (specifically stated in projects 2277, 2281, 2234, 2354) and likely used in other (unconfirmed), particularly from tracks 1, 2, 4 and 5. It is likely that other ontologies were also used, but the lack of comprehensive feedback on this

aspect from all partners means this is as yet unconfirmed (see also 'Collated feedback'). Extensive ontology support has been provided for CORBEL WP4 for the Marine Metazoan Developmental Models (use case 4.4) where standard terminology was required to harmonise data for three animals (jellyfish, sea urchin and amphioxus) (see 'Showcase projects').

Standards and formats used in projects

The **FAIRsharing registry** is the preferred registry for the listing of standards and formats used within the various CORBEL projects. The interoperability questions posed in the original forms did not target this aspect, as it was not the focus for this interoperability exercise. i.e. it is a registry more suited for the selection of suitable project resources, in advance of actual implementation. However, a cursory inspection of the resources used within projects (e.g. 'Databases and repositories used in projects'), indicates that the majority of those databases are also listed in Identifiers.org (in a different context). See also 'Recommendation 5'.

Tools and software used in projects

The **bio.tools registry** is the preferred registry for the listing of tools and software used within these various CORBEL projects. As expected, a variety of tools and software were already registered in bio.tools, which include Cytoscape, IMOD, Chimera, Matlab, and ImageJ.

Unfortunately the information provided in the original proposal submissions was often insufficient to determine the tool or software that was being used, for instance providing only a 'name' with no reference or link, or else not specifying a 'build' or 'version'. Examples include:

Proj. 2281: RIGER
Proj. 2301: Amira
Proj. 2325: CellProfiler and Ilastik
Proj. 2234: specific instances of 'Affymetrix' software

Multiple projects mentioned the use of 'R' packages, and of ad hoc Python scripts, the details of which were not specified (see also 'Gap Analysis').

Gap analysis

Two major gaps were identified in the existing process, which could significantly impact interoperability potential for projects such as these, particularly post-completion, and would likely impact both 'Findability' and 'Reuse' (from the FAIR Principles).

Scripts and unspecified computational software and packages

Computational software, ad hoc scripts and 'packages' (extending software capability) were rarely described adequately. Downloadable software could be added to bio.tools listings, but this registry does lack comprehensive 'package' listings. Rather than add individual packages, it may be beneficial to make a more thorough evaluation of how such packages should be referenced, including

version/build information. Some external repositories such as GitHub may be suitable for this purpose, if not bio.tools. Other emerging initiatives include: standardised descriptions for command line scripts through the Common Workflow Language (http://www.commmonwl.org), Research Objects for packaging files (http://researchobject.org/) and standardised reporting of software for citation and findability using CodeMeta files (https://codemeta.github.io/) or the Citation File Format (https://citation-file-format.github.io/).

Further investigation is required.

Image data

Image acquisition, processing, analysis and storage comprised a significant component of the total services used in these calls. In the completion of this report, it has become apparent that there are many software tools and 'packages' used by this user community which are not commonly encountered in other domains. Further, there seem to be no community recommended practices for image sharing and storage, with multiple institutes hosting their own data, while users are provided with data through personal storage devices. Some work to address this has begun with efforts such as http://idr.openmicroscopy.org/about/'. Also of note is the concern expressed ('Service Provider Feedback', EuBI ALMF) that legacy image formats need to be supported long term to ensure accessibility to the data.

The UK's BBSRC Strategic Review of Bioimaging (https://bbsrc.ukri.org/documents/1805-bbsrc-strategic-review-of-bioimaging-pdf/) provides useful pointers for further investigation.

Showcase projects

The following example project(s) shows the deep engagement that has been enabled over the course of this exercise and demonstrates the clear benefits that would otherwise not have been realised.

Marine invertebrate ontologies

In an effort to standardise metadata in the Marine Invertebrate Models Database (http://marimba.obs-vlfr.fr) being developed by partners in CORBEL WP4, a number of new ontologies were needed to describe morphological features for three animals (jellyfish, sea urchin and amphioxus). WP6 provided support for developing these ontologies and will provide the necessary services for the ongoing hosting of the ontologies through the EMBL-EBI Ontology Lookup Service (https://www.ebi.ac.uk/ols). An ontology development pipeline was created that allowed domain experts to construct the ontologies using spreadsheet formats, with which they are more familiar. This pipeline serves as a prototype or template for how future ontologies could be developed for other domains.

The ontologies have been developed in line with community standards established by the Open Biological Ontology (OBO) foundry, and where each ontology will be submitted for inclusion in the OBO library (http://obofoundry.org). Canonical URI identifier provide a means for standardised

access to term information, as well as facilitating mapping (other ontologies) where necessary thorough additional OLS tools.

The ontologies developed for each organism capture terminology with stable identifiers for anatomy, cell types and developmental stages. The anatomical terms are organised in a partonomy and cover the entire life-cycle stage of the organism from embryogenesis to adulthood. Terms for developmental stages are provided and additional relationships have been added that capture how the stages typically progress and how the anatomical structures emerge in early development. These terms are being used to provide consistent annotation of gene expression data and images in the MARIMBA database and the ontology will provide new opportunities for querying, categorising and visualising the data.

The ontologies were developed using spreadsheet templates so that the domain experts could provide their expert knowledge using familiar tools. Two templates were provided for each ontology, one for capturing information about the development stages, and one for the anatomy terms. The experts were asked to provide labels, synonyms and definitions for each term along with details of how the terms were related to each other. The data collected in the spreadsheets was then put into a pipeline developed in WP6 that utilised the ROBOT (http://robot.obolibrary.org) software to convert the spreadsheet data into an ontology format (versions of the ontology were generated in both the OBO and OWL format). The generated ontologies were reviewed by the experts through looking at them in the Ontology Lookup Service, and the process was iterated to improve and extend the ontologies. Table 4 summarises the identifier namespace for each ontology and details on how to access the source files and the ontology through the OLS registry.

Ontology name	Organism	Source repository	Registry url (Not yet public)
CHEM	*Clytia hemisphaerica*	https://github.com/simonjupp/pliv_ontology	https://www.ebi.ac.uk/ols/ontologies/chem
PLIV	*Paracentrotus lividus* (urchin)	https://github.com/simonjupp/chem_ontology	https://www.ebi.ac.uk/ols/ontologies/pliv
AMPH	*Branchiostoma lanceolatum* (amphioxus)	https://github.com/simonjupp/amph_ontology	https://www.ebi.ac.uk/ols/ontologies/amph

Table 4. Ontologies developed to support Marine Invertebrate Models Database, in partnership between WP4 and WP6.

Collated feedback

15 projects were selected for further feedback (highlighted in green, Appendix A1.1), targeting both the service provider and the researchers who proposed the project. Excerpts from this further survey are provided in Appendix A2.1. The feedback is presented below in 2 categories, scientific feedback (on individual projects), and service provider feedback (in most cases, spanning multiple projects).

Scientific feedback

Feedback solicited from project lead investigators through email, listed by project.

Project 2301

	name	Identifiers.org	FAIRsharing	bio.tools	OLS
Formats	n/a	n/a	n/a	n/a	n/a
Tools	IMOD ImageJ	n/a	n/a	yes yes	n/a
Databases	EMDataBank* (http://www.emdatabank.org/) EMPIAR* (https://www.ebi.ac.uk/pdbe/emdb/empiar/)	yes yes	no yes	n/a	n/a
Ontologies	n/a	n/a	n/a	n/a	n/a

*Data will be considered for submission to these databases following publication

Note: Main challenge stated as data transfer; e.g. volumes acquired with FIB-SEM microscope are ~5 GB so hard to manage, transfer and store.

Project 2354

No databases used to date. No further information provided.

Project 2358

Too early in their project process - will provide feedback at a later date.

Project 2242

No ontologies used, no databases used to date. Also noted: there is no desire to submit data at a later stage unless compelled to do so. Survey response indicates there will be use of standard analysis packages when results are generated, but none specified.

Project 4719

	name	Identifiers.org	FAIRsharing	bio.tools	OLS
Formats	n/a	n/a	n/a	n/a	n/a
Tools	R software python scripts	n/a	n/a	yes n/a	n/a
Databases	KEGG gene KEGG pathway Uniprot CAS cancerrxgene	yes yes yes yes yes	yes yes yes yes no	n/a	n/a
Ontologies	GO cellular component GO molecular function GO biological process	yes yes yes	yes yes yes	n/a	yes yes yes

Service Provider feedback

Feedback solicited from research infrastructure provider associated with specific project, through email. Results are listed by service.

EuBI ALMF (Projects 2311, 2354, 2359, 2363)

Do not use ontologies or deposit data in any database. Image data is returned to users on a memory stick or disk.

Specific feedback:

1. An image data 'DropBox' would facilitate sharing
2. Concern expressed on usability of legacy image data formats in the future, and suggested support is considered for this.

ISBE-VU (Projects 2354, 2305, VIP Lisbon)

	name	Identifiers.org	FAIRsharing	bio.tools	OLS
Formats	SBML SBGN	n/a	yes yes	n/a	n/a
Tools	COPASI CellDesigner	n/a	n/a	yes yes	n/a
Databases	Not directly used but encourage users to deposit in BioModels or JWS Online, using FAIRDOM Hub (http://fairdomhub.org) for exchange of files with users				n/a
Ontologies	not specified				unknown

EMBRC (Projects 2325, 2357, VIP41, 5364)

Besides providing access to marine organisms, the provider is developing an imaging public database for 3 metazoan marine model organisms (jellyfish, amphioxus, sea urchin) called MARIMBA (MARine Invertebrate Models dataBAse). In that respect:

	name	Identifiers.org	FAIRsharing	bio.tools	OLS
Formats	FASTA	n/a	yes	n/a	n/a

	GFF3		yes		
Tools	COPASI CellDesigner CHADO GMOD JBrowse	n/a	n/a	yes yes yes yes	n/a
Databases	investigating use of Image Data Resource, http://idr.openmicroscopy.org/	n/a	n/a	n/a	n/a
Ontologies	SO	yes	n/a	n/a	yes*

*A further marine organism specific ontology in development in collaboration with OLS

ChEMBL (Projects 2219, 2305 and 2219)

	name	Identifiers.org	FAIRsharing	bio.tools	OLS
Formats	n/a	n/a	n/a	n/a	n/a
Tools	R RDkit Jupyter notebook	n/a	n/a	yes no no	n/a
Databases	ChEMBL ChEBI UnProt PubChem OpenTargets	yes yes yes yes under investigation	yes yes yes yes yes	n/a	n/a

Ontologies	GO	yes	yes	n/a	yes
	EFO	yes	yes		yes

Recommendations

1. *Interoperability questions could be mandatory.*
   The interoperability questions posed in the original proposal forms (Textbox 1) were optional, and indeed there were only answered in a handful of cases. Even then, the answers provided seemed to have misinterpreted the intent of the question. Therefore, the questions themselves should be linked to a fuller explanation of what is being asked. Hence, such questions should be mandatory, particularly with an increasing focus on research outputs, and their level of FAIRness. Improving interoperability potential will also further the ability to more seamlessly integrate to the EOSC, which is under active development. Making mandatory these questions is a subject that should be addressed at the most appropriate venue (e.g. CORBEL AGM through WP4&6).

2. *Interoperability issues should be reviewed early.*
   Whilst interoperability issues *per se* should not be used as a criterion in the evaluation of scientific merit, projects approved following such review should subsequently be explored to optimise such concerns. This process should be lightweight, and undertaken in collaboration with service providers, proposing scientists, and appropriate interoperability platform [5]personnel. It is imperative that processes are put in place to address the data that are being continuously generated through such programmes, making them as FAIR as possible, as early as possible. It should also be noted that not all projects, as made apparent in this deliverable, are suitable to be FAIR-ified, for example where the project requires only access to a key technology or facility.

3. *A data sharing culture needs to be evoked.*
   It was noted from feedback that some researchers were unwilling to deposit their data (output) into public repositories, unless compelled to do so by project funders (CORBEL). We would recommend that data deposition of project outputs should be the default behaviour, and non-compliance should be permitted only under exceptional circumstances and with good reason (for example sensitive data). This recommendation may require policy change, as well as a cultural shift in the behaviours of researchers. "Open by default" is a pillar of the EC's EOSC Declaration[6] .

4. *Interoperability registries should be amongst the first stops when specifying project details.*
   Projects using references to biological components (cells, chemicals, proteins, genes) should preferentially use URIs or compact identifiers linked to meta-resolvers such as

[5] https://www.elixir-europe.org/platforms/interoperability
[6] https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

identifiers.org. Links to tools and software should be verified to be listed in bio.tools, while ontologies used should be available through OLS. Where these are absent, requests should be made to the appropriate registry to have them listed. Software packages and code should be available through repositories such as GitHub, sourceforge or R package where appropriate, and their use should be accompanied by specific metadata on version and build, etc.

5. *Interoperability registries need to be better interlinked to share information.*
   There should be a better and more formal linking between the interoperability registries. This would allow more seamless access to information, enabling the linking between the repositories used within projects (listed in Identifiers.org), to the formats and standards used within those repositories (Identifiers.org <-> FAIRsharing), and to their selected vocabularies (FAIRsharing <-> OLS).

Acknowledgements

We wish to acknowledge the contributions of the following people:
Friederike Schmidt-Tremmel (CORBEL, UK)

References

- N/A

Delivery and schedule

The delivery is delayed: No

Related documents

CORBEL WP6 DoW[7]

Appendices

[7] https://docs.google.com/document/d/1D0J_d-scpdzLGDHH-Gic2drMTHec8o06GyxE4Hrwr_g/edit

Appendix A1.1

Project identifier	Title	Proposal keywords	Services used	Service category (technology /science)	Interoperability questions answered in proposal?	Interoperability scope	comments
2219	A metabolic dialogue in the microbiota-gut-brain interphase	NMR and mass spectrometry forcompound identification,	protein targets with ChEMBL, Structural Biology at Instruct Centre- CERM, German Mouse Clinic (GMC), and MIRRI for mouse strains access	technology mostly	none	uses unichem to map to Chembl	Chembl registered in identifiers.org and FAIRsharing
2242	Influence of the phenotype and genotype of cytochromes	geno- & phenotype effects on hepatic cytochromes in drug metabolising activity	brfaa, metabolomics services through CERM, NMR from instruct Florence	technology mostly	some	uses NCBI SNP database and Broad Institute identifiers (http://exac.broadinstitute.org)	NCBI SNP database (dbSNP) is registered in identifiers.org and FAIRsharing. 'RS' identifiers issued by Broad are directly equivalent to dbSNP.
2277	Molecular Insight into Autism Spectrum Disorder (ASD)	clinical samples, saliva, genetic profiling, 'omics data integration for markers	biobanks processing (BRFAA), modelling VU Amsterdam(ISBE), BIMSB	tech and science	yes	Databases: STRING, Panther, Uniprot accessions, Entrez gene; Vocabularies: Gene Ontology; Tools: Cytoscape	All databases and tools are registered in the appropriate registry. Further investigation is required to asses the use of SalivaTec (dedicated salivary biomarker discovery, http://salivatec.weebly.com

							) and OralCard (http://bioinformatics.ua.pt/OralCard). These seem primarily to reuse existing identifiers for proteins, and targeted to associating those identifiers to publications. Further investigation required.
2281	Targeting EVI-1 in Acute Myeloid Leukemia	screening compounds, CRISPR knockout, in vitro effects, clinical trial design	Screening and medicinal chemistry at Leibniz-Institute for Molecular Pharmacology (FMP), German Mouse Clinic (GMC), candidate validation in vivo at INFRAFRONTIERS/EMMA	technology and some databases and tools	some	databases: GEO, CCLE and GDSC ; tools and software: R project, GenePattern, GSEA, RIGER; ontologies: Gene Ontology	registered: GEO, Gene Ontology; CCLE requires authenticated access so is not listable; GDSC (https://www.cancerrxgene.org/) has been submitted for registration at identifiers.org. Besides some standards software (R packages), the other tools are unknown from the details provided. This should be explored further with the project lead.
2294	Toxin-antitoxin system in lactic acid bacteria	toxin/antitoxin systems, microbiology, fermentation/clinical apps, genomic sequencing/annotation, imaging for physiological effects	Super Resolution Light Nanoscopy (sharpe lab crg) BCN, BRFAA (bacterial/plasmid sequencing), CRG sequencing, BRFAA imaging	technology	none	low or unlikely	
2298	novel drugs cancer	lead compounds to be characterised to treat triple	Structural Biology at Instruct Centre-CERM/CIRMMP, Leibniz-Institute for	technology	none	none	

		negative brest cancer genotype through wnt signaling inhibition	Molecular Pharmacology (FMP), NMR/isotope labeling at CERM, screening / profiling at FMP				
2301	Analysis of neuronal subcellular architecture by FIB-SEM	require powerful microscopy to disentangle subcellular architecture in neurons	microscopy (CMC) at University Medical Centre Utrecht, Image Processing at CNB-CSIC/Instruct	technology	tools and software	IMOD, Amira, Chimera	IMOD and Chimera registered in bio.tools. AMIRA graphical software package needs to be investigated to see if appropriate to include in bio.tools.
2305	systems toxicology approach for dosimetry of endocrine disrupting compounds	develop computational models as training for MSc students	Molecular Cell Physiology (VU Amsterdam), Chemogenomics (ChEMBL), screening and medicinal chemistry Leibniz-Institute for Molecular Pharmacology (FMP)	technology	none	none	Potential for follow up with project lead on whether model created have been deposited in registered model repositories such as Biomodels database or SEEK.
2311	SERF inhibition suppress proteotoxicity in age-related disease	expt on mice mutants analysis for plaques, behavioural changes, ..	Mouse phenotyping at German Mouse Clinic (GMC), Advanced Light Microscopy Facility at EMBL, behaviour of mutant mice SERF k/o (GMC), dissection/brain analysis imaging (ALM)	technology	none	unlikely	
2325	mechanics of tissue morphogenesi	sea urchin embryo tissue morphogenetic	CNRS Marine Station, Marine Biology Facility	technology	software ImageJ, Matlab, Imaris,	little besides software	ImageJ, Matlab and Imaris are all registered in bio.tools. CellProfiler and

	s sea urchin	changes through microscopy	(EMBL), Marine Laboratory at Stazione Zoologica, Advanced Light Microscopy Facility (EMBL)		CellProfiler, Ilastik		Ilastik pending investigation for addition.
2334	exploring microsatellite instability in colorectal carcinomas	mouse model, compound screening, cancer tissue expression profiling agilent array probes, crispr to generate mouse model and test	Screening and medicinal chemistry (FMP), Mouse mutant phenotyping (GMC), Bioinformatics (BIMSB/MDC)	techonolgy mostly	yes	Gene Expression Omnibus (GEO) repository, Bioportal for Cancer Genomics (www.cbioportal.org), ensembl; tools - EnrichR (http://amp.pharm.mssm.edu/Enrichr) Affymetrix® Transcriptome Analysis Console (TAC) Software Affymetrix® Expression Console (EC) Software; databases - drugbank (https://www.drugbank.ca), PharmGKB (https://www.pharmgkb.org/) and genecards (http://www.genecards.org; GO and Kegg pathway database (http://www.genome.jp/kegg/pathway.html)	databases: GEO, PharmaGKB, DrugBank, GeneCards and KEGG pathway all registered, vocabularies: Gene Ontology registered; tools: EnrichR registered. Affymetrix software needs to be explored for possible addition to bio.tools (to determine if existing Afymetrix software registered is compatible)
2335	**COMPLETED**				no	interacted closely with EMBL-EBI team	databases: BioStudies (registered); vocabularies: gene ontology (registered)
2350	Screening for porin inducers as novel antimicrobials	porins for transmembrane entry of antibiotics which is issue for outermembrane	screening at FMP, analysis/genomics Chembl	technology	yes	uniprot, pubchem; R bioconductor; Assay classification according to in-house standards. Classification according to test systems used (e.g.,	databases: PubChem, Uniprot (both registered), R Bioconductor (registered in bio.tools). In house standards used should be investigated for

		. ie. sensitise bacteria. involves screening for natural porin inducers				Alpha Screen, redox reaction, enzyme classes) - check in house standards ; challenges - Making as much screening data publicly available as possible while restricting access to other parts of our data that might be used to generate IP	interoperability.
2354	Modelling ROS management ... in models of Parkinson disease	dynamic moodel through isbe, structure data from elixir, visualisation of mitochondrial proteins EU BI	Molecular Cell Physiology Vrije University Amsterdam (model fitting), Chemogenomics (ChEMBL), resonance (ALM)	technology	yes some	ncbi gene, uniprot; GO at ncbi	databases: NCBI gene, Uniprot are both registered; gene ontology (independent of NCBI) is also registered.
2357	Involvement of lin-28 and let-7 in neural development of amphioxus	collect fish, culture embryos, inject test genes, collect freeze embryo for analysis, imaging	CNRS Marine Observatory of Banyuls-sur-mer, France, Advanced Light Microscopy Facility at EMBL, Fluorescence Microscopy (CNRS),	technology	none	unlikely	
2358	Morphology and structure of chondrocytes in shark and ray tissues	shark cartilage as model for skeletal tissue studies	EMBRC: Skeletal specimens will be collected at the Centre for Marine Sciences (CCMAR), analysis SLN@BCN, Super Resolution Light Nanoscopy (BCN) Stochastic optical reconstruction microscopy (STORM) -	technology	limited	ImageJ/FIJI, MATLAB image processing, segmentation of volumetric data (e.g. using software such as Amira and DRISHTI for working with CT scans).	Software: ImageJ and Matlab are already registered in bio.tools. AMIRA already pending further investigation, FIJI also to be investigated.

			SLN/BCN,Laser scanning confocal Microscopy (LSCM), SLN/BCN ,Multiphoton microscopy systems - SLN/BCN, Light sheet fluorescence microscopy systems - SLN/BCN				
2359	High Content Screening of Compounds Inducing Oligodendrocyte Differentiation	myelin repair in MS, screen for chemical compounds promoting remyelination	screening at FMP, imaging at ALM	technology	little	tools- Cell insight; Arrayscan; Volocity; Zen, Fidji, Image J;	tools: ImageJ already registered. Other tools to be investigated with further input from investigator.
2363	GDE4: function of a curious intracellular LPA producing enzyme	GDE suppresses malignancy, and seems to generate bioactive lipids eg LPA. crispr to generate gde k/o. determine location of GDE, phenotype the mutants, charaxterise GDE signaling	Mouse mutant phenotyping at German Mouse Clinic (GMC), Advanced Light Microscopy Facility at EMBL, Screening and medicinal chemistry (FMP)	technology	none		
2375	Dynamics and structural characterisation of Bovine viral diarrhea virus host	live imagine of virus entry and interactions, and endocytosis	image processing (CNB-CSIC/Instruct), ALM	technology	none	unlikely	

	interactions during attachment and entry						
2376	Functional 3D Analysis of immune synapse contacts	substructural contacts and changes through immune cell communication - microscopy	Cell Microscopy (CMC), Advanced Light Microscopy Facility at EMBL, Super Resolution Node BCN, Image Processing at CNB-CSIC/Instruct	technology	little	tools/s/w - Imaris, Fiji, Arivis, Huygens, Paraview	Imaris is already registered in bio.tools. The remaining tools and software need to be explored with the project lead.
4719	High-throughput compound screening approach ..	AML, cell lines, compound screening	EU OPENSCREEN, provides compound screening, needs libraries (FMP), images (ALM@EMBL)	technology	none	identification of cell lines, image metadata/storage	
5364	Chloroplast plasticity in planktonic symbioses	photosynthesis, cell lines, symbiosis, 3D EM analysis	EU BI, organism collection (CNRS Marine), analysis (ALM@EMBL)	technology	only some software for image analysis	cell line identification, image metadata	
5377	Scaling up the analysis of phagocytosis in macrophages ...	phagocytosis, microglia, compound screen, imaging miscrosopy	provides zebrafish cell line, needs light sheet imaging (EU BI), compound screening (EU OPENSCREEN)	technology	databases, github and pubmed	github for analysis code, pubmed for literature, primarily image data issues	code deposited in github, databases: pubmed registered
5467	Neuron and microglia in sickness and in health...	neuronal/microglial interaction, phagocytosis, imaging	EU BI, needs microscope facilities barcelona, deposit in biostudies, share with ALM	technology	none	biostudies, imaging	BioStudies registered
5530	3D structural characterization of	x-ray tomography visualisation,	EU BI, needs miscroscopy (FIB focused ion beam	technology	relating to image	imaging	

	erythrocyte infection…	parasitic infection,	and light miscroscopy) - fixed samples sent for analysis				

Appendix A2.1

Excerpts from follow up questionnaire targeted to Service providers and researchers:

service providers, we wish to determine the standards that are used in their service provision. For example, controlled vocabularies and ontologies used in data, public databases (used or referenced), availability of public APIs and general availability of the data, identifiers assigned to the data (as routine, not specifically for this project), and whether any components have been registered in ELIXIR (see 'd' below). Please **note**: You may receive multiple requests if you are involved with multiple projects. So feel free to submit 1 response covering all projects you participated in (and note the project ids please).

applicants, the proposal forms contained a section on interoperability relating to datasets, tools and resources that would be required for the project, as well as envisaged challenges to achieving interoperability and reuse (copied below). Information on interoperability challenges and resources are an important component of the work in WP6 (Data access, management and integration).
Our aim is to gather knowledge of

1. Any databases used for referencing entities (e.g. proteins, samples, organisms, strains, etc.) i.e. input information.
2. Databases where the output (e.g. data files, sequence data, image data) has been deposited i.e. the outcomes
3. Any ontologies used in the project, i.e. to describe the data
4. Any tools or software that were used in the processing/work i.e. in generating the output.