# EXCELERATE Deliverable D5.1

| | | |
|---|---|---|
| **Project Title:** | ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences | |
| **Project Acronym:** | ELIXIR-EXCELERATE | |
| **Grant agreement no.:** | 676559 | |
| | H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1 | |
| **Deliverable title:** | Interoperability Implementation Regulations | |
| **WP No.** | 5 | |
| **Lead Beneficiary:** | 32 - UL | |
| **WP Title** | The ELIXIR Interoperability Backbone | |
| **Contractual delivery date:** | 28 February 2018 | |
| **Actual delivery date:** | 31 August 2018 | |
| **WP leader:** | Carole Goble; Chris Evelo | 4 - UNIMAN; 6 - NBIC |
| **Partner(s) contributing to this deliverable:** | 4 - UNIMAN; 6 - NBIC; 1 - EMBL (EBI and ELIXIR Hub); 2 - UOXF; 27 - INRA | |

**Authors and Contributors:**
Carole Goble (UNIMAN); Chris Evelo (NBIC); Helen Parkinson (EMBL-EBI); Nick Juty (UNIMAN); Susanna-Assunta Sansone (UOXF); Peter McQuilton (UOXF); Rob Hooft (NBIC); Sarala Wimalaratne (EMBL-EBI); Sirarat Sarntivijai (EMBL-EBI); Rafael C Jimenez (ELIXIR Hub); Jon Ison (ELIXIR-DK), Simon Jupp (EMBL-EBI); Cyril Pommier (INRA); Célia Miguel (INRA); Michael Crusoe (CWL); Marco Roos (Leiden); Sirarat Sarntivijai (ELIXIR Hub); Jerry Lanfear (ELIXIR Hub).

# Table of contents





Horizon 2020 grant n. 676559

Horizon 2020 grant n. 676559

# 1. Executive Summary

EXCELERATE WP5 targets the ELIXIR Interoperability Platform (EIP). This deliverable originally aimed to present “Interoperability Implementation Regulations” as “a set of standards, rules, controlled vocabularies, authorized unique identifiers and interoperable service APIs for the data repositories and biological knowledge bases agreed and implemented in the field with WP6 to 9 (Use Cases), and WP3 (Data)”, as well has ensuring that resources were registered in appropriate ELIXIR registries.

During the execution of EXCELERATE we have significantly refined and adapted this vision to offer a flexible and practical open framework for developing and deploying interoperability across the WPs, working with the WPs, and to more appropriately pump-prime sustainable and scalable approaches to interoperability in ELIXIR.

The main aim of this deliverable is to present our progress for the implementation of our interoperability framework for EXCELERATE Use Cases WP6-9 and the ELIXIR Data Platform (WP3). This deliverable does not seek to *dictate or prescribe* the rules, regulations or vocabularies. Rather, it addresses the necessary interoperability infrastructure to support the use cases through **standards, services and know-how**.

WP5 adopted four guiding principles: (i) **FAIR**: services and practices for Find, Access Interoperate and Reuse of Data; (ii) **Interoperability for a Purpose** driven by need rather than idealism; (iii) **Interoperable Interoperability**: adopting emerging practices and technologies aligned with global standardisation efforts and communities; and (iv) **Reuse not reinvention**: identifying pre-existing ELIXIR and external services. WP5 does not develop services: it coordinates their adoption and supports their development by their owners.
We developed and continuously refine an **EIP Framework,** using this as a roadmap to define the components and processes that comprise and define the wider ELIXIR Platform. The framework helps us organise and consolidate work for structured collaboration with other projects and initiatives, as well as providing a means to interact with Use Cases, other ELIXIR Platforms and EXCELERATE WPs and with ELIXIR Nodes. It also supports the work on service sustainability, service maturity and business cases to be reported in D5.2, and BYODs to be reported in D5.3.

Task 5.2 set about provisioning this EIP framework with services, standards and know-how to build the EIP Backbone. Task 5.1 uses these for Use Cases and the Core Data Resources and Deposition Databases of the Data WP5. Each task informs the other. As the use cases use the services, we summarise Task 5.2 before Task 5.1.

## 1.2 Task 5.2 Interoperability Implementation Services

The Framework to be provisioned covers six areas:

- **Services Framework**: Reference framework, service selection; metrics and maturity models.
- **Metadata**: Registry services; metadata services and resource mark-up (Bioschemas).



Horizon 2020 grant n. 676559

- **Standards**: Formats, minimal information models and ontologies; identifiers; API and tool description; and workflow description and portability.
- **Approaches**: Workflows and Linked Data as prime mechanisms for interoperability.
- **Capacity building**: BYODs and hackathons; workshops and tutorials; and a Knowledge Hub.
- **Global initiatives**: engagements with global activities in all of the above.

Highlights include:

- The development of a **Service Reference Framework**, a robust objective selection process for **Recommended Interoperability Resources** (RIRs), analogous to the CDRs of WP3 currently in its first call and a preliminary FAIR Capability Maturity Model (tasks 5.2.1, 5.2.2, 5.2.5).
- A working set of **Interoperability Registry Services** have been identified and promoted: Standards registry (FAIRSharing); Ontology registry (OLS) and Identifiers registry (Identifiers.org). All are actively used by the Use Cases. We have near to complete registered coverage of the WP3 Core Data Resources and Deposition Databases. All registries have undertaken work to improve their interoperation with each other and with other ELIXIR Registries, and to provide better capabilities. For example, our work on robust support for machine-resolvable, persistent compact identifiers in biomedical data citation, interoperating identifiers.org with N2T.org in the USA (tasks 5.2.1, 5.2.2, 5.2.5).
- An ongoing full audit and gap analysis of services offered by ELIXIR nodes is underway. Key **EIP metadata services** have emerged as providing important services for the Use Cases and our own Interoperability Registries: ontology mapping and development, identifier mapping, metadata annotation and validation, and metadata search and harvesting. Services for metadata validation and ontology mapping and development have been deployed for the Use Cases; others are on the priority list for gaps. The full audit and gap analysis of services will be completed for D5.2 and related to the Service Framework (task 5.2.5).
- The design and implementation of **Bioschemas** (http://bioschemas.org) - **a new data discovery mechanism and universal markup standard for dataset descriptors** to support dataset lifecycle management, and to assist integration, aggregation and search tools using ELIXIR Resources. We established a Bioschemas community of 200+ people and 18 working groups. 12 ELIXIR Nodes, all Use Cases and many of the CDRs/DDs participate (tasks 5.2.3, 5.1.2, 5.1.1).
- The **development of ontologies, reporting standards and standard APIs for all Use Cases, relevant to their communities**. **Identifier best practice guidelines** have been published and disseminated (tasks 5.2.1, 5.2.2, 5.2.3).
- The recommendation of standardised best practices for the design and documentation of APIs using OpenAPI and the EDAM Ontology, with WP1 (task 5.2.3).
- The adoption, design and implementation of the **Common Workflow Language** (http://commonwl.org), a community driven workflow standard for workflow interoperability and portable execution (additional task to the DoW).
- A **demonstration of practical Interoperability practices using Workflows and Linked Data for WP6 and WP8** respectively. We worked alongside WP6 to



Horizon 2020 grant n. 676559

standardise workflow descriptions, and with WP1 and WP4 to build fully containerised workflows (additional task to the DoW). With WP8 we work on Node capacity building and access to technical infrastructure from the ELIXIR-NL Node (the Data FAIR Point toolkit) that supports Linked Data interoperability (task 5.2.4).

- **Bring Your Own Data (BYODs), hackathons, summer schools, workshops and tutorials** commonly held with other WPs, global initiatives and other project partners. (tasks 5.3.1, 5.3.2, 5.3.3).

## 1.3 Task 5.1 FAIR Principle Interoperability Implementation Agreements

### Task 5.1.1 Use Cases

Engagements with the Marine Metagenomics (WP6), Plants and Crops (WP7) and Rare Disease (WP8) Use Cases have determined the requirements and interoperability service needs, delivery of Bring Your Own Data Workshops (BYODs).

Interoperability requirements were elicited for the different Use Cases through meetings, interviews and targeted hackathons. Each WP has used services identified in Task 5.2; developed and used standards; engaged with global interoperability initiatives; and adopted different interoperability approaches. WP5 has assisted each Use Case in developing ontologies, metadata annotation and validation services; tools and workflow descriptions, API documentation, organising BYODs and supporting specialised services for workflows and linked data. Much of this work is reported by the Use cases in their own deliverables.

Standards and best practices guidelines have been published by the Use Cases for metagenomics studies[1], plant phenotyping[2] and rare disease dataset FAIRification[3].

Highlights include:

- Universal engagement by Use Cases in Bioschemas. Resource markup of the MarRef database using Bioschemas for easy metadata publishing to BioSamples in WP6; recommended for standardising concept description by WP7; and for markup of multiple patient registries in WP8.

[1] Petra ten Hoopen, Robert D. Finn, Lars Ailo Bongo, Erwan Corre, Bruno Fosso, Folker Meyer, Alex Mitchell, Eric Pelletier, Graziano Pesole, Monica Santamaria, Nils Peder Willassen, Guy Cochrane; The metagenomic data life-cycle: standards and best practices, GigaScience, Volume 6, Issue 8, 1 August 2017, gix047, https://doi.org/10.1093/gigascience/gix047

[2] Pommier, C., Cornut, G., Letellier, T., Michotey, C., Neveu, P., Ruiz, M., Larmande, P., Kersey, P.J., Cwiek, H., Krajewski, P. and Coppens, F. (January, 2018) Data standards for plant phenotyping: MIAPPE and its implementations [W785]. Proceedings Plant and Animal Genome XXVI Conference. PAG. San Diego : PAG, Résumé,

[3] Carta, C., Roos, M., Jacobsen, A., Thompson, M., Wilkinson, M.D., Cornet, R., Waagmeester, A., Van Enckevort, D., Jansen, M., Licata, L. and Via, A. (2017) January. The FAIRification of data and the potential of FAIR resources demonstrated in practice at the Rome Bring Your Own Data workshop. In CEUR Workshop Proceedings (Vol. 2042).



Horizon 2020 grant n. 676559

- Universal use of EIP registries by all Use cases. Use of EIP metadata tools: for ontology making (Slim-O-Matic) by WP6, ontology mapping, term search, and term conflict resolution (OLS, OxO) by all WPs; metadata validation (ISATools) by WP7.
- Support in the development of MIAPPE and BrAPI reporting standard and API, and the adoption and extension of the ISA framework to support MIAPPE validation (WP7) and ontologies suitable for WP6 metagenomics pipelines (a GO-Slim for InterPro, and enseqlopedia in OLS for sequencing application information).
- Adoption of Common Workflow Language (CWL) for formal description of computational analysis processes by WP6 (Deliverable D6.3)
- A Linked Data FAIRification process developed by WP8 for data harmonisation best practice to make rare disease resources interoperable (FAIR) 'at the source', using the Data FAIRPort services. (Deliverable 8.2).

Pump-primed by EXCELERATE, 7 related ELIXIR Implementation Studies have followed-on and strengthened the work. See Appendix 8.

### Task 5.1.2 Data

**WP3 (Data)** takes responsibility for the Core Data Resources (CDR) and Deposition Databases (DD) of ELIXIR. With WP5 we support steps towards interoperability.
Highlights include:

- Ubiquitous, minimal web-based mark-up of resources with machine processable metadata using **Bioschemas** for specifications at the data resource level DataCatalog and Dataset, which apply to all WP3 CDR and DD databases. Five Core Data Resources (CATH, EGA, Human Protein Atlas, MINT, PDBe) and three Deposition Databases (BioSamples, PDBe, EGA) are marked-up so far in their Live deploys. Many more are in the pipeline.
- A selection process of **Recommended Interoperability Resources** that mirrors and adapts the WP3's CDR selection process
- Recommendations for **standardised identifier practices**[4], in cooperation with EU CORBEL project. A simplified identifier hygiene checklist will be used to check all CDRs and DDs.
- Registries for standards, ontologies and identifier resolution used by WP3's CDRs and DDs. All CDRs and DDs are registered in FAIRsharing and Identifiers.org.
- CDRs and DDs are being examined for compliance to "FAIR Metrics" in an ELIXIR Implementation Study [5]that is complementary to EXCELERATE.

### Task 5.3.3 Node Capacity Building

**WP10** aims to build capacity in the Data Nodes to disseminate interoperability practice. The Interoperability Services are provided by the Nodes.

Highlights include:

[4] https://doi.org/10.1371/journal.pbio.2001414

[5] https://www.elixir-europe.org/platforms/data/fairness-core-resources



Horizon 2020 grant n. 676559

- A **Recommended Interoperability Resources** selection process open to all Nodes and an audit of Node resources to be reported in D5.2. Currently only 3 Nodes (UK, EBI and NL) provide the Interoperability registries and metadata services used by WP5.
- Registration of Node resources in Interoperability Registries: 57/69 eligible registered in FAIRSharing; 59/70 eligible registered in Identifiers.org.
- 12 Nodes involved in the Bioschemas initiative.
- Node Staff Exchange programme awarded to support Node uptake of Bioschemas.
- In at least six of the Bring Your Own Data meetings (BYODs) focusing on data and five focusing on (repository) software, a total of more than 100 practitioners have been taught practical steps[6] to make data interoperable by FAIR data modeling using linked data and ontologies.

### Task 5.1.3 Global Engagement

To adhere to our third principle **Interoperable Interoperability** we have systematically engaged with global initiatives and sought to adopt pre-existing and emerging standards rather than "roll our own". WP5 has participated in over 61 meetings in 15 countries with a reach of over 3,600 people (see EIP dissemination log for details).
Throughout Appendix 1 and 2 in situ references are made to Global Engagements and Collaborations. Appendix 4 gives a matrix of WP5 activities to the global initiatives and standards bodies we have collaborated with.

Highlights include: Research Data Alliance, GA4GH, Pistoia Alliance, NIH Data Commons, Force 11, Common Workflow Language, FDA BioCompute Object, Research Object and the Galaxy Community. Significant EU H2020 project collaborations include: CORBEL, EOSCPilot, BioExcel, FREYA, openAIRE, CHARME Cost Action, RD-Connect. Commercial collaborators include: Google, DataCite, and Springer Nature Scientific Data.

# 2. Impact

| D5.1 related Activity | Impact Indicators |
|---|---|
| Task 5.1<br>FAIR Principle Interoperability Implementation Agreements | WP6<br>● 4 containerised pipelines in CWL (METAPipe, ITSoneDB, EMG, MetaShot) that are a gamechanger for pipeloine comprehension and portability.<br>● 350 terms added to OLS for sequencing<br>● GO-Slim for InterPro now encapsulates over 97% of all GO-terms currently identified compared with the previous version that only mapped 83% of GO-terms. |

[6] For instance, see the FAIRification workflow in Deliverable D8.2



Horizon 2020 grant n. 676559

| | |
|---|---|
| | • Published standards recommendations in leading journal (ten Hoope, et al 2017).<br><br>Outreach<br>• Poster and oral presentations at international/European rare disease and bioinformatics events reported by WP6.<br><br>Proposals<br>• Grant proposal focusing on Tools and Workflow Collaboratory using CWL, Call: H2020-INFRAEOSC-2018-2020 (Implementing the European Open Science Cloud), INFRAEOSC-04-2018, SEP-210489595: EOSC-Life (accepted).<br><br>WP7<br>• BrAPI endpoints to 20 datasets for a total of 1,000+ studies: INRA/GnpIS (ELIXIR-FR): 20 datasets with 1 000+ studies (maize, wheat, forest trees, grape, etc); tropgene (ELIXIR-FR): 5 to 10+ datasets (rice); EU-SOL BreeDB (ELIXIR-NL): (various solanaceae datasets); PIPPA (ELIXIR-BE): one maize dataset; (ELIXIR-SI): 2 solanaceae datasets and PHENO (ELIXIR-PT): woody plant and rice datasets.<br>• Published standards recommendations (Pommier et al , 2018)<br><br>Outreach<br>• Poster and oral presentations at international/European rare disease and bioinformatics events reported by WP7.<br><br>WP8: also see Deliverable D8.2<br>• 99 resources 'Rare Disease' collection in bio.tools[7]<br>• 8 WP5 BYODs supported the development of the FAIRification process for WP8. Annual BYOD as part of the Rome Summer school for rare disease registry managers (Rome, 2015-2018), and the 'Sample & Data Banking' Course, Bologna, 2017 & 2018.<br>• Preliminary version of the FAIRification workflow was the basis for the cross-project 'rare disease data linkage plan' that runs on support from RD-Connect, ELIXIR, ELIXIR-NL, ELIXIR-EXCELERATE, BBMRI-ERIC, BBMRI-ADOPT, BBMRI-NL, Dutch FAIR-development grants (ODEX4All, FAIR-dICT), and contributions from rare disease stakeholders.<br>• The rare disease data linkage plan is the basis for organising FAIRification in the European rare disease field. A Global Open FAIR implementation network is in preparation to further support FAIRification in the rare disease community in the context of |

[7] https://bio.tools/?page=1&collectionID=%27Rare%20Disease%27&sort=score

| | |
|---|---|
| | implementing the European Open Science Cloud for rare disease research.(from D8.2)<br><br>Outreach<br>● Poster and oral presentations at international/European rare disease and bioinformatics events, including the European Conference for Rare diseases and Orphan Drugs (Edinburgh, 2016 ; Vienna, 2018 ), the IRDiRC conference (Paris, 2017) , RD-ACTION Workshop Co-hosted by DG SANTE: Using standards and embedding good practices to promote interoperable data sharing in ERNs (Brussels, 2017) , RD-Connect annual meetings (2015, 2016, 2017, 2018) , E-Rare workshop (Berlin, 2017) , European Conference of Human Genetics (2016, 2017) , Semantic Web Application and Tools for Health Care and Life Science (2016, 2017) , ISMB/ECCB (Prague, 2017) , also presented as ELIXIR webinar .<br><br>Proposals<br>● Grant proposals in the rare disease domain that include FAIRification, particularly the European Joint Program-COFUND (SC1-BHC-04-2018, accepted). |
| Task 5.2 Interoperability Implementation Services | Established Bioschemas<br>● Community of 200+ people and 18 working groups.<br>● 12 ELIXIR Nodes participate. All Use Cases use markup.<br>● 15 specifications cover multiple 'types'<br>● 3 new schema.org types<br>● 30 live deployments of datasets, 6 Million + pages marked up<br>● 5 ELIXIR Core Data Resources and 3 Deposition Databases marked up<br>● EIP Registries and 2 other ELIXIR Registries (bio.tools, TeSS) use Bioschemas markup.<br>● 11 Hackathon events[8] developing markup, over 250 people<br>● 8 Dissemination events (see http://bioschemas.org/publications/)<br>● 4 tutorials (SWAT4LS 2018, NETTAB 2018, ECCB 2018, ELIXIR All Hands 2018).<br>● 5 posters (ELIXIR All Hands, SWAT4LS, ISMB, ISWC)<br><br>EIP Registry Services capture ELIXIR Node Data Resources<br>● 57/69 eligible registered with FAIRsharing (85%) (total 1067 resources registered)<br>● 59/70 eligible registered on Identifiers.org (84%) (total 655 collections, 814 resources registered) |

[8] http://bioschemas.org/meetings/



Horizon 2020 grant n. 676559

| | |
|---|---|
| | • 200+ Ontologies registered on OLS<br><br>Metadata<br>• OxO 1.8 million ontology mappings validated by Pistoia Alliance Ontologies Mapping Project<br>• Zooma 100,000 manually curated text to ontology mappings from 9 databases to predict ontology annotations<br>• ISAtools adopted by WP7, and Metabolomics User Community<br>• Identifier recommendations (McMurry et al PLoS 2016) accessed 19K times<br><br>Registry Services featured in EOSC and NIH Data Commons<br>• FAIRsharing featured in the EU FAIR Data Experts' Report on Turning FAIR Data into Reality and NIH Data Commons projects<br>• Bioschemas, FAIRsharing and Identifier.org pilots for the EOSC Datasets Minimum Metadata Guideline (EDMI) developed by the EOSCPilot.<br>• Force11 and NIH Data Common Identifiers resolution harmonisation between Identifier.org and N2T.net<br><br>Workflow standardisation, portability and discovery (FAIR Workflows)<br>• ELIXIR is a flagship Common Workflow Language adopter ad has funded standardisation efforts in the Community.<br>• CWL adopters include NIH BD2K, (including 3 Cancer Genomics Cloud Pilots), GA4GH Cloud execution, the European Open Science Cloud and the EU RI IBISBA for Industrial Biotechnology<br>• CWL proposed as the workflow description component for the Food and Drug Administration's BioCompute Objects[28], which is undergoing IEEE Standardisation (IEEE P2791 BioCompute Working Group[9]).<br><br>Proposals<br>• Grant proposal focusing on Tools and Workflow Collaboratory using CWL, Call: H2020-INFRAEOSC-2018-2020 (Implementing the European Open Science Cloud), INFRAEOSC-04-2018, SEP-210489595: EOSC-Life (accepted).<br>• Grant proposal in FAIRification: Proposal number: 802750-1, IMI2-RIA, Topic: IMI2-2017-12-02, FAIRPlus (accepted). |

[9] IEEE P2791 BioCompute Working Group (BCOWG): Standard for Bioinformatics Computations and Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication, http://sites.ieee.org/sagroups-2791/

| | |
|---|---|
| | • 7 ELIXIR Implementation Studies (see Appendix 8).<br><br>Impact on Global initiatives<br>• RDA Metadata IG, FAIRSharing WG, Pistoia Alliance, Force 11, etc see Table 4.<br><br>Outreach<br>• Poster and oral presentations at 34+ international events in USA, Europe and Asia. See Appendix 6.<br>• Organised 21 workshops and 5 forthcoming |
| Task 5.3 Bring Your Own Data" (BYOD) & Capacity Building Workshops | BYODs to be reported in Deliverable D5.3<br><br>BioHackathon 2018 organised by ELIXIR WP5 EIP (120 people expected).<br><br>Organised 26 events (see Appendix 6). |

# 3. Project objectives

| No. | Objective | Yes | No |
|---|---|---|---|
| 1 | Objective 1: Define agreements on identifiers and machine processable (meta)data descriptions with data providers (WP3, WP6 to 9), for (i) data repositories and (ii) knowledge bases, using established technical and domain standards, working with global initiatives to ensure broader interoperability | Yes | |
| 2 | Objective 2: Consolidate existing Interoperability Services to support: (i) machine processable identity, data formats, experimental reporting guidelines, knowledge representations, and (ii) resource operational practices for transparent releases, versioning, provenance, updates. | Yes | |
| 3 | Objective 3: Implement data interoperability between prioritized resources, in priority areas. | Yes | |
| 4 | Objective 4: Run a programme of Bring Your Own Data (BYOD) "bootcamps" and coordinate with WP11. Build capacity with WP10: Data Nodes Network. | | No |



Horizon 2020 grant n. 676559

# 4. Delivery and schedule

The delivery is delayed: ☑ Yes • No

# 5. Adjustments made

The deliverable reporting has been quite delayed due to staffing issues, material availability and a decision to wait for Use Case deliverables to be completed. The work described has not been delayed.

# 6. Background Information

WP as originally indicated in the Description of Action (DoA) is included here for reference.

| | | | |
|---|---|---|---|
| Work package number | 5 | Start date or starting event: | Month 1 |
| Work package title | **The ELIXIR Interoperability Backbone** | | |
| Lead | Barend Mons (up to 1/3/2017) (NL)<br>Chris Evelo (since 1/3/2017) (NL) and Carole Goble (UK) | | |
| **Participant number and person months per participant**<br>1 – EMBL 12.00; 2 – UOXF 10.40; 3 - TGAC 10.60; **4 - UNIMAN 19.00**; 6 - NBIC 23.00; LUMC 4.00; UMAAS 10.00; DLO 1.00; 8 - CRG 2.20; 10 - IRB 13.00; 12 - BSC 15.00; 13 - CSIC 2.80; 17 - INESC-ID 6.90; 24 - UiO 4.00; 25 - SIB 6.00; 26 - CNRS 10.00; 30 - CNR 4.26; **32 - UL 6.00**; 34 - UOCHB 16.19; 38 - DTU 10.00; 45 - UU 20.00 (KTH 2.00; SU 10.00); 46 - HWU 11.00 | | | |
| WP5 - The ELIXIR Interoperability Backbone [Months: 1-48]<br>UL, EMBL, UOXF, TGAC, UNIMAN, NBIC, CRG, IRB, BSC, CSIC, INESC-ID, UiO, SIB, CNRS, CNR, UOCHB, DTU, UU, HWU<br>Optimal Interoperability is attained when data access and use can be completely automated: programming and interfaces conform to standards that specify consistent syntax and formats; and data are associated with metadata and terminology identifiers and codes that support computational aggregation and comparison of information that reside in separate resources. As an exemplar implementation of this principle, The Interoperability Backbone in WP5 is data and Use Case driven, implementing FAIR principles, working in partnership with the custodians of the datasets. We will support cross-resource questions, for instance bridging genomic and phenotypic data for variant | | | |



Horizon 2020 grant n. 676559

identification, using machine processable (RDF/XML) representations of the metadata. We use this as a reference for building Node capacity in skills and knowledge for data interoperability and access to technical infrastructure that supports data interoperability. For example: Rare disease data described with the ORDO (Orphanet Rare Disease ontology) to integrate rare and common disease (WP8); Different plant phenotypic data described with a variety of specialist and overlapping terms (e.g. Plant Ontology, Crop Ontology, Plant Trait Ontology) with sample environment descriptors drawn from the Environment Ontology (EnvO) and the eXtensible Experiment Markup Language (XEML) (WP7); Marine metagenomics data described using EnvO, XEML, plus descriptions to screen environmental metagenomic sequence datasets (e.g. FOAM), and the Genomic Standards Consortium's reporting checklists and provenance aspects(such as environment from which samples originate).
For the data resource and service management of Named and Core Resources (WP3) we emphasize self-described datasets and explicitly described and published life-cycle metadata, using machine processable representations and common APIs for accessing it. The approach taken here will be an iterative one, working on tasks in parallel, to ensure that continuous but stepwise improvements for interoperability are yielded, and forged in practice against expressed need. We will not attempt comprehensive perfection: instead we aim for a principled "Just Enough and Just in Time" interoperability "on demand", while raising the bar on general data service quality along with WP1. No standard is committed to without an example implementation.

The objectives for WP5 are:

1. Define agreements on identifiers and machine processable (meta)data descriptions with data providers (WP3, WP6 to 9), for (i) data repositories and (ii) knowledge bases, using established technical and domain standards, working with global initiatives to ensure broader interoperability.
2. Consolidate existing Interoperability Services to support: (i) machine processable identity, data formats, experimental reporting guidelines, knowledge representations, and (ii) resource operational practices for transparent releases, versioning, provenance, updates.
3. Implement data interoperability between prioritized resources, in priority areas.
4. Run a programme of Bring Your Own Data (BYOD) "bootcamps" and coordinate with WP11. Build capacity with WP10: Data Nodes Network.

Work Package Leads: Barend Mons (up to 1/3/2017) (NL) Chris Evelo (since 1/3/2017) (NL) and Carole Goble (UK)

**Task 5.1: FAIR Principle Interoperability Implementation Agreements (52PM)**
Many of the necessary services already exist across ELIXIR but need alignment and support. This task is aimed at identifying critical services, defining the interfaces between them, and developing a long-term integration strategy. By doing so, this will survey national (Node) practices and needs. Work is based on existing ELIXIR infrastructure and community emergent conventions and the forging of partners in and outside Europe. The WP partners have established track records of metadata



Horizon 2020 grant n. 676559

management for biology datasets or metadata mechanisms, notably using Linked Data (e.g. Semantic Web) technologies for publishing self-described data (e.g. RDF), resolving identifiers(e.g. URIs) and defining ontologies.
For datasets we will agree:

- Practices of data management and data publishing; Managed APIs and message formats, with agreed APIs for access to dataset descriptors. We tackle data repositories and biological knowledge bases differently, reflecting their different content and their content lifecycles.
- Common exchange formats: such as RDF and XML data schemes.
- Common reporting guidelines: submission, curation and validation tools using data templates (example: ISATAB and tools), focusing on interoperability of standards via common data element mappings.
- Common terminologies: general (e.g. VoID, PROV for provenance); data type/community specific (examples: ORDO, Plant Trait Ontology); cross-community common elements (example: EnvO environment descriptors).
- Common APIs: for common data types.
- Best practices for publishing data as Linked Data, leveraging the EMBL-EBI's RDF Platform, LinkedISA, and resources of other platforms (example: IMI Open PHACTS Discovery Platform), as a semantic interoperability platform in addition to the use of APIs.

For biological knowledge bases the task will concentrate on common conventions for:

- Descriptions using common terminology, standard data formats, and mappings between common data elements and standard ontologies.
- Descriptions of the dependencies, curation and computational processes used to generate the current record for the biological entity, where appropriate.
- Good practices for publishing data as Linked Data, leveraging the EMBL-EBI's RDF Platform and resources of other platforms (example: IMI Open PHACTS Discovery Platform), as a semantic interoperability platform in addition to the use of APIs.

Partners: all partners will be involved in sprints, focussing on specific dataset combinations to achieve WP6 to 9 questions.
Managers/Core: UK, NL, EMBL-EBI, SE

**Subtask 5.1.1: Use Cases (WP6 to 9) (22PM)**

Outcome: Interoperability, common APIs and descriptors workable in the field. First on WP7 (Genomic and Phenotypic Data for Crop and Forest Plants) and WP8 (Rare Disease) followed by WP6 (Marine metagenomics). Jointly identify data-driven interoperability examples, cross-dataset questions, bottlenecks and common descriptions. Survey national (Node) practices and needs.

**Subtask 5.1.2 Core and Named Resources (WP3) (15PM)**

Outcome: Common APIs and dataset descriptors workable in the field. Focus on interoperability at the dataset level.



Horizon 2020 grant n. 676559

Workshops to co-produce common APIs and dataset descriptors leveraging proposed standards (example: W3C HCLS
Dataset descriptor). Contribute to dataset metrics/quality criteria for data service life-cycle management in WP3. Survey national (Node) practices and needs.

**Subtask 5.1.3 Global engagement: international organisations and multilateral forums (15PM)**

Outcome: Establish a leading role for ELIXIR internationally on this aspect and compatibility with other international interoperability initiatives. Engagement with European initiatives (examples: IMIs, RIs, EUDAT EUON), global initiatives (examples: NIH BD2K, GA4GH, Force11), commodity standards (examples: W3C, DataCite, ORCID, VIVO, ORE) and community standards (example: RDA). Feedback between external forces and ELIXIR resources, maintaining ELIXIR visibility in key meetings with interoperability initiatives.

**Task 5.2 FAIR Interoperability Implementation Services (55PM)**

Integrating complex datasets requires services to handle identifiers for data and biological concepts (phenotypes, diseases), and tools that allow users to map data between different sources and help users find and apply standards. Many of these tools exist. This task brings these services together.

Partners: UK, NL, EMBL-EBI, SE, FR

**Subtask 5.2.1: Identity Management, Mapping and Tracking services (14PM)**

Outcome: Make explicit the scope and limitations of identifiers, the mappings between identifiers for entities (example: Ensembl gene identifiers can be mapped to Uniprot identifiers) and provide identifier services used by data resources in the field. Work includes: identity authorities for specific data types and concept categories; identity resolution, identity mapping, and entity resolution.

**Subtask 5.2.2: Reporting Guidelines, Formats, Controlled Vocabulary Services (8.2PM)**

Outcome: The best of breed services assembled, organized into a coherent tool suite, and used in practice in WP3 and WP6 to 9. Workshops of service providers and users to “bake-off” alternatives and integrate complementary services.

**Subtask 5.2.3: Dataset publishing for API interoperability (10.8PM)**

Outcome: Common and standardised practices to dataset lifecycle management and release management, including: Distributed revision control (example: GIT), dependency management; and well described, validated and maintained APIs registered in catalogues (example: BioCatalogue, BioSharing).

**Subtask 5.2.4: Biological knowledgebase publishing for Linked Data interoperability (12PM)**

Outcome: Data published as Interoperable Linked Data for some biological knowledge bases. We will develop services for the creation and management of mappings, as first class artefacts, between data entities to describe the curation and computational processes used to generate the current record for the biological entity (leveraging work at EBI/SIB, DTL, SciLifeLabs and the UK).

**Subtask 5.2.5: Sustainability of Interoperability Implementation Services (12PM)**

Outcome: A strategy for sustaining key services.



Horizon 2020 grant n. 676559

**Task 5.3: "Bring Your Own Data" (BYOD) & Capacity Building Workshops (105PM)**
Implement data interoperability between resources for WP6 to 9 and data publishing of resources for WP3. Provide practical support and guidance through a programme of Bring Your Own Data bootcamps. BYODs are a mix of tutorials and hands on, practical "hackathons" with specific datasets: database custodians' work with experts in semantic web and linked data technologies to make their data available in a FAIR way using machine processable (meta) data and APIs. During the BYOD Task 5.1's specifications, standards and data templates are refined in practice and Task 5.2's tools and services are exercised.
Partners: NL, UK, ,FR, PT, SI, SE, CZ, IT, ES
**Subtask 5.3.1: Manage and Run BYOD Workshops (65PM)**
Outcome: Organise and manage workshops synergistically with WP3 and WP6 to 9 work plans, as well as aligned with WP11 training activities. Monitor workshops and feedback to Tasks 5.1 and 5.2. After the first 6 months of the project we anticipate a BYOD every 3 months.
**Subtask 5.3.2: Create and manage BYOD training materials (20PM)**
Outcome: Materials for BYOD, updating in step with Tasks 5.1 and 5.2. Deploy BYOD materials on the WP11 TeSS Portal. Develop BYOD materials for Data Carpentry training (WP11).
**Subtask 5.3.3: Data Node Capacity Building (20PM)**
Outcome: Build capacity in the Data Nodes (WP10, Task 10.2) using 5.3 as a reference example for interoperability practice and the BYOD methodology rolled out across the Nodes. Have capability to independently run BYOD's for new or national datasets in one third of the Nodes.


Horizon 2020 grant n. 676559

# Appendix 1: FAIR Principle Interoperability Implementation Agreements (Task 5.1)

## A1.1 Introduction

EXCELERATE aims to drive coordination efforts at both national and international levels in a harmonised approach to data management in life sciences, which is a priority for Europe.

The ability to retrieve published scientific data declines over time, highlighting the importance of a long term strategy for data stewardship. Addressing such challenges requires an approach that is consistent across institutional and geographic boundaries, maximising data sharing and use for the scientific community. Furthermore, there is a need to engage policy makers and research funding bodies, at both national and European level, to collectively respond to this challenge.

The ELIXIR Interoperability Platform (EIP), through WP5, aims to provide **Services, Standards and Expertise** in order to maximise the value and benefit of disparate resources across disciplines and borders, and align with activities in other platforms. The work of EIP enables allow partners (e.g. other ESFRI Research Infrastructures, national resources, institutional archives) to make use of each others' existing data and services, and to connect and interoperate between their own resources and allow them to leverage evolving resources as they develop.

The WP5 strategy has been refined[10] to accelerate the implementation of interoperability services necessary for the WP6-9 Use Cases and Core Data Resources prioritized in WP3.

- Provisioning a coordinated "Backbone" of services and resources by identifying and consolidating **existing interoperability services** to support machine processable identity data formats/ reporting guidelines/ knowledge representations, and resource operational practices
- Consolidating and sharing sustainable standards and practices by defining agreements on **identifiers** and **machine processable (meta)data** descriptions with data providers (WP3, WP6 to WP9)
- Driving "standards as the default" across the Use Cases and Data WPs, working with other WPs to **co-identify, co-provide and co-disseminate:** appropriate services, standards, knowledge and best practices (rather than prescribe "a set of regulations[11]").
- Engaging with international standards efforts and ensuring interoperability resources are **registered** in appropriate ELIXIR registries and portals for findability and accessibility.

[10] https://docs.google.com/document/d/1pm9PZIivXjiF33ofbOT77R7YgB2YslLOoGmWE-x6UUY/edit

[11] as specified in the Proposal deliverable



Horizon 2020 grant n. 676559

This deliverable originally aimed to present “Interoperability Implementation Regulations” as “a set of standards, rules, controlled vocabularies, authorized unique identifiers and interoperable service APIs for the data repositories and biological knowledge bases agreed and implemented in the field with WP6 to 9 (Use Cases), and WP3 (Data)”, as well has ensuring that resources were registered in appropriate ELIXIR registries.

During the execution of EXCELERATE we have significantly refined and adapted this vision to offer a flexible and practical open framework for developing and deploying interoperability across the WPs, working with the WPs, and to more appropriately pump-prime sustainable and scalable approaches to interoperability in ELIXIR.

## A1.2 EIP Principles and Roadmap (Task 5.2)

The WP has adopted four guiding principles:

- **FAIR Services**: services and practices for Find, Access Interoperate and Reuse of Data. A key objective over the course of EXCELERATE, and in perpetuity, is to make data ‘FAIR’ (Findable, Accessible, Interoperable, Re-usable). While each principle may itself have different contextual meaning within specific domains, the philosophical intent is to enhance and maximise research and data outputs, many of which have been generated through public funding.
- **Interoperability for a Purpose**: interoperability services and resources on an “as necessary” basis, driven by need rather than theory and idealism, for: Use Cases (WP6-9), Data Platform (WP3) and Nodes (WP10).
- **Interoperable Interoperability**: adopting emerging communities, practices and technologies, avoiding ad-hoc and proprietary implementations and engaging with, contributing to, and using key international standardisation efforts. We aimed to ensure and demonstrate the interoperability of EIP services and the harmonised use of interoperability resources in practice, for example: we have fully engaged with and adopted the Common Workflow Language; Bioschemas builds on the Schema.org web markup standard the use of; Bioschema dataset descriptions are combined with ontologies and identifiers.
- **Reuse not reinvention**: identifying pre-existing ELIXIR and external services, resources, ontologies, and solutions. WP5 does not develop services; it coordinates services developed by the ELIXIR Nodes. For example, our registries are all pre-existing services offered by the Nodes.

To frame, organise and prioritise the work needed, WP5 has developed an ELIXIR Interoperability Platform (EIP) Framework and uses it as a roadmap. Reported in the mid-term review as 7 sub-projects, this has since been further organised into 4 major sectors: Metadata, Standards, Approaches, Capacity (Figure 1)[12]. Each sector has services, which contribute to and are supported by a Services Framework, and each are subject to FAIR Metrics and Maturity Models. All components are aligned with Global Initiatives, in many of which WP5 members are active participants

[12] See Appendix 3



Horizon 2020 grant n. 676559

**The EIP Framework's** parallel yet integrated sectors have enabled the WP to improve the granularity of tasks and better define specific interim milestones towards the deliverables. Dependencies, drivers and interactions with other work packages, are also defined. This approach ensures maximal connectivity between different elements which are identified in advance. This mapping also allows what were previously independent tasks to be incorporated with better precision into their larger contextual projects. The global engagement activities and BYODS have been incorporated into specific projects, where engagement can be appropriately exploited & managed and BYODs better tailored (to be reported in D5.3). Throughout, the promotion and implementation of good practice in interoperability services is encouraged through the formalization of metrics and quality indicators and their adherence to the FAIR principles in collaboration with the ELIXIR Data Platform.

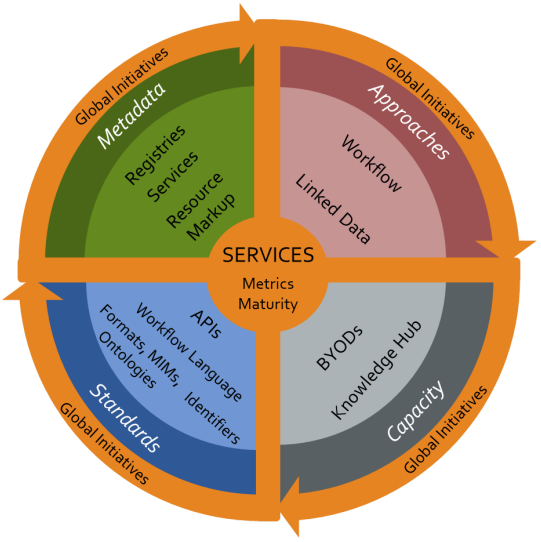


**Figure 1: ELIXIR Interoperability Platform (EIP) WP5 Framework.**

A brief summary of the components are below and are further expanded in Appendix 2.

**Services, Metrics and Maturity Models**

**FAIR Services Framework:** sets out the interoperability services needed to operate FAIR data and assist the implementation of the FAIR-data needs of the Use Cases, and includes: landscape and gap analysis, audit of Node and external resources, supporting processes for identifying Recommended Interoperability Resources, Interoperability Resource Maturity Model, interoperability needs between the interoperability services themselves and interactions between the services. Services are not developed by WP5.

**Metrics and Maturity Models for Data and Processes:** works towards a design framework and exemplar metrics to measure and compare FAIR resources and a preliminary FAIR Capability Maturity Model Integration to help quantify the investment needed to reach different levels of FAIR and assess the gaps to reach different FAIR maturity levels.

**Metadata**



Horizon 2020 grant n. 676559

**Interoperability Registry Services:** identifies, promotes and integrates registries provided by ELIXIR Nodes. Interoperability Registry Services have been identified for standards (FAIRsharing), ontologies (Ontology Lookup service) and identifier resolution (Identifiers.org). Work also integrates these registries with external registries and other ELIXIR registries and we participate in international efforts for catalogue metadata and exchange, notably the European Open Science Cloud.

**Metadata Services:** sources and coordinates the range of services needed for ontology and identifier management and mapping, resource annotation, annotation validation and semantic search. Work has chiefly focused on ontology, annotation/validation and linked data services driven by the needs of the Use Cases.

**Resource markup – Bioschemas:** specifies and promoted the adoption of dataset descriptors with common data elements, to support dataset lifecycle management and release management, and to assist integration tools using ELIXIR Resources. Bioschemas, an extension Schema.org specification has been developed to improve Findability and Accessibility of ELIXIR resources and serve ELIXIR registries, indexes, integration and aggregation tools and search engines.

**Standards**

**Formats, Minimal Information Models and Ontologies:** provides advice and assistance and supplies support services to the Use Cases who specify and adopt standards for their specific data types:

**Identifiers:** identifies resource requirements and aims to harmonise existing practice in identifier assignment and resolution, to support resources in the implementation of community standards and in the adoption of best practice and identifier services.

**APIs and Tools Description Standards:** promotes the use of standardised best practices for the design and documentation of APIs and helping communities in their definition and adoption of standardised APIs for data types.

**Workflow Standards:** participates as a driving member of the Common Workflow Language community which promotes a new standard to describe workflow recipes and analysis tools in a platform-neutral way, making them portable and scalable across multiple computing environments (with the Tools platform and increasingly the Compute Platform).

**Approaches to Interoperability**

**Workflows: w**ork alongside WP6 to standardise workflow descriptions, and WP1 and WP4 to build fully containerised workflows, coordinate support services and develop and promote best practice and knowledge transfer across WP.

**Linked Data:** work alongside WP8 (Rare Disease) on building Node capacity in skills and knowledge for data interoperability (through FAIRification process and BYODs), and for access to technical infrastructure that supports Linked Data interoperability using the resources of RD-Connect and the Data FAIRPort toolkit provided by ELIXIR NL and GO-FAIR.

**Capacity Building**

**BYODs and Hackathons**: Bring Your Own Data (BYODs), hackathons, summer schools, conventional workshops and tutorials are the mainstay of our face to face training. BYODs will be reported in D5.3.



Horizon 2020 grant n. 676559

**Knowledge Hub**: a "one stop shop" website for guidelines, pointers to guidelines, templates, best practice papers etc for EIP's interoperability work, complementing training materials and events listed in the TeSS Portal.

## A1.3 Relationship to EXCELERATE Use Cases (WP6-9) and Data (WP3) (Task 5.1)

To adhere to our second principle (**Interoperability for a Purpose**) tasks cut across several Use Cases but are strongly coordinated with one EXCELERATE Use Case for a deep focus, linking use cases to specific interoperability services, providing concrete outcomes for each.

### A1.3.1 Use Cases (Task 5.1.1)

Engagements with the Marine Metagenomics (WP6), Plants and Crops (WP7) and Rare Disease (WP8) Use Cases have determined the requirements and interoperability service needs, delivery of Bring Your Own Data Workshops (BYODs), the design and implementation of Bioschemas (http://bioschemas.org) - a new data discovery mechanism and universal markup standard and the design and implementation of the Common Workflow Language (http://commonwl.org) - a community driven workflow standard for workflow interoperability and portable execution.

Interoperability requirements were elicited for the different Use Cases through meetings, interviews and targeted hackathons. A Use Case EIP Framework Engagement Map was developed, presented in the mid-term review and subsequently revised (Figure 2).

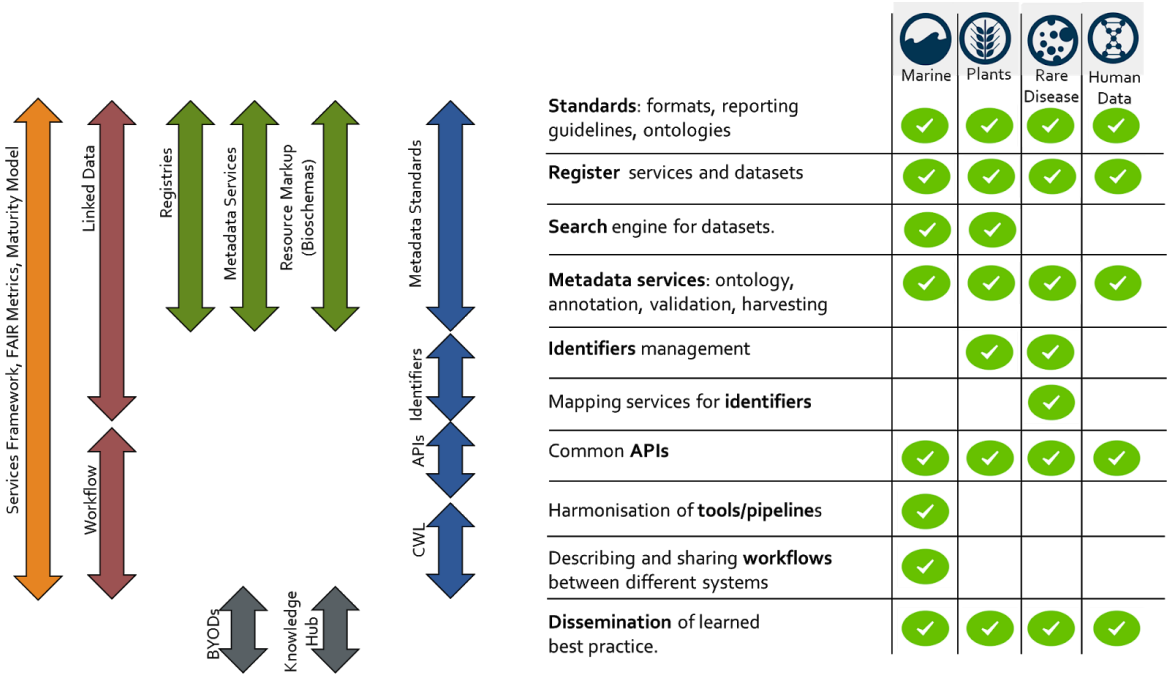


**Figure 2: Use Case and Interoperability Services mapped against the EIP Framework**

This deliverable addresses the necessary interoperability infrastructure to support the Use Cases as indicated by the mapping, and to help identify, consolidate and disseminate *common* activities and services. Each Use Case requires scientific and technical experts,



Horizon 2020 grant n. 676559

and collaborations to define the ways in which this can be accomplished and the means to deliver such 'FAIR' data and services. Thus we do not seek to *dictate or prescribe* the rules, regulations or vocabularies that should be used in the context of individual Use Cases.

We summarise the key interoperability activities undertaken with Use Cases. The details of the EIP services are given in Appendix 2. A summary publication was presented at ICBO 2018[13]

*WP6: Marine Metagenomics*
The goal of the Use Case is to develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain. It does this in partnership with other European infrastructure projects and international partners including MS-RAST in the USA.

The Use Case is developing standards and best practices for the marine domain driving the use of the Common Workflow Language for the description of several metagenomics analysis pipelines including automated pipelines for analysing and archiving metagenomics: the EBI metagenomics pipeline and the META-Pipe pipeline. It has developed contextual and sequence reference databases (MarRef, MarDB, MarCat, ITSoneDB, Eukaryotic gene catalogue). MarRef, MarDB, MarCat are available from the MGnify Portal.

The Interoperability approach has been to use Workflows to increase transparency, portability and reproducibility.
Highlights:

- Adoption of Common Workflow Language (CWL) for formal description of computational analysis processes.
- Standards and best practices guidelines for metagenomics studies. Checklists currently operate at ~60% adoption in BioSamples for sample sequencing records, ~ 50% across all soil samples in ENA (superset), ~25% across all ENA. Specific ontologies have greater adoption (e.g. ENVO at 83%).
- Developed ontologies suitable for the metagenomics pipelines including: a GO-Slim for InterPro and recommendation of inclusion of enseqlopedia into the Ontology Lookup Service (OLS) to provide sequencing application information. The archiving of results for identification data analysis of biological entities entails annotation of sequence data with metadata (eg. GO terms, gene identifiers/symbols, etc.). ENA has been recently extended to accommodate such information. Referenced-based taxonomy classification criteria have been identified as a challenging, multi-dimensional system.

[13] Sarntivijai, S, Juty, N, Goble, C, Parkinson, H, Evelo, C, Lanfear, J, Blomberg, N, Corvallis, , Interoperability: Standardisation of identifiers, schemas, and ontologies for scientific communities in Proc International Conference on Biological Ontology 2018, Oregon, USA, August 7-10 2018



Horizon 2020 grant n. 676559

- Resource markup of the MarRef database using Bioschemas for easy metadata publishing to BioSamples.

WP5 assists WP6 in developing ontologies, service interoperability, tools and workflow descriptions, API documentation and the development and the development and use of the Common Workflow Language (Table 1).

| WP5 Activities | WP6 Marine Metagenomics |
|---|---|
| Registries | |
| - FAIRSharing | Pipeline legacy databases (e.g. InterPro) mainly registered.<br>MarRef, MarDB, MarCat, ITSoneDB, Eukaryotic gene catalogue not.<br>MIxS, MIMS, MIMARKS and MIGS standards registered<br>M2B3, MISAG and MIMAG not |
| - Ontology Lookup Service | Pipeline ontologies registered<br>(ENVO, GO etc) |
| - Identifiers.org | MarRef, MarDB, MarCat registered<br>ITSoneDB, Eukaryotic gene catalogue not. |
| Metadata Services | |
| - Ontology | Slim-O-Matic used to produce new GO-Slim for EMG pipelines |
| - Annotation / Validation | -- |
| - Search | MGnify Portal[14] facetted term and sequence similarity search |
| - Identifier | -- |
| - Other | CWL Viewer[15] for visualising the EMG pipeline in GitHub |
| Resource Markup: Bioschemas | Design of protein, sample and data resource specs.<br>Mar databases expose additional metadata using Bioschemas markup; ELIXIR BioSamples Deposition Database harvests this. |
| Standards | |
| - Formats, MIMs, Ontologies | Standards specified[16]<br>~350 terms added[17] to OLS from for sequencing application information |
| - Identifiers | Hygiene check to be completed. |

[14] https://www.ebi.ac.uk/metagenomics/

[15] https://view.commonwl.org/workflows/github.com/mr-c/ebi-metagenomics-cwl/blob/viz/workflows/emg-pipeline-v3.cwl

[16] Petra ten Hoopen, Robert D. Finn, Lars Ailo Bongo et al; The metagenomic data life-cycle: standards and best practices, GigaScience, Volume 6, Issue 8, 1 August 2017, gix047, https://doi.org/10.1093/gigascience/gix047

[17] From http://enseqlopedia.com/enseqlopedia/



Horizon 2020 grant n. 676559

| | |
|---|---|
| - API Standards and Standard APIs | MGnify API in OpenAPI format |
| - Workflow standards | Participated in CWL standardisation<br>EMG, ITSone and MetaShot pipelines specified in CWL<br>Annotations using EDAM<br><br>META-Pipe and furthers workflows being developed in an ELIXIR Implementation Study |
| Approaches | |
| - Linked Data | -- |
| - Workflows | Galaxy, AWE and Toil workflows described using CWL, containerised using Biocontainers, annotations using EDAM |
| Capacity | |
| - BYODS | Training workshops organised by WP6<br>CWL workshops, tutorials and hackathons organised by WP5+WP6 |
| - Knowledge Hub | TBD |
| Global Engagements | GSC, Ocean Sampling Day, MS-RAST |

**Table 1: Summary of WP5 for WP6 Interoperability Activities**

*WP7: Plant, Crop and Forest Phenotyping*

The goal of the Use Case is to establish a technical infrastructure and associated social practices to allow plant genotype-phenotype analysis for crop and tree species based on the widest available public datasets. By making data interoperable, in accordance with the FAIR principles, plant genotypic and phenotypic data will be easier to find, integrate and analyse. It does this in partnership with other European infrastructure projects, including EMPHASIS.

The Use Case is developing a Data Lookup service dedicated to genetic and phenomic data, using these reporting standards and APIs, and a database annotation and validation submission process.

The Interoperability approach has been to use common APIs, schemas, and workable formalised descriptors for Genomic and Phenotypic Data.

The ELIXIR Plant Informatics community examined different plant phenotyping data and identified challenges, requirements, and tools needed to build the research metadata infrastructure for Plant Informatics[18]. The results indicated that there was a non-harmonised use of nomenclatures in the different sectors of the community where different plant phenotypic data were described with a variety of specialist and overlapping terms from multiple ontologies: Plant Ontology, Crop Ontology, and Plant Trait Ontology.

[18] Pommier, C., Cornut, G., Letellier, T., Michotey, C., Neveu, P., Ruiz,M., Larmande, P., Kersey, P.J., Cwiek, H., Krajewski, P. and Coppens, F. (January, 2018) Data standards for plant phenotyping: MIAPPE and its implementations [W785]. Proceedings Plant and Animal Genome XXVI Conference. PAG. San Diego : PAG, Résumé,1 p.



Horizon 2020 grant n. 676559

Additional descriptors such as plant sample descriptors were drawn from Environment Ontology and the eXtensible Experiment Markup Language.

Highlights:

- Recommendation for the adoption of Bioschemas to standardise concept description
- Ontology term search, ontology term conflict resolution, through OLS to standardise identifiers
- Development and adoption of the MIAPPE (Minimum Information About a Plant Phenotyping Experiment) reporting standard and the adoption and extension of the ISA framework to support MIAPPE validation, including a BRAPI2ISA conversion toolkit.
- Implementation of MIAPPE in a web service using the Breeding API (BrAPI) to connect participating repositories. Each partner has built an access controlled BrAPI endpoint to enable seamless access to datasets and 'trials', including 20 datasets for a total of 1,000+ studies: INRA/GnpIS (ELIXIR-FR): 20 datasets with 1 000+ studies (maize, wheat, forest trees, grape, etc); tropgene (ELIXIR-FR): 5 to 10+ datasets (rice); EU-SOL BreeDB (ELIXIR-NL): (various solanaceae datasets); PIPPA (ELIXIR-BE): one maize dataset; (ELIXIR-SI): 2 solanaceae datasets and PHENO (ELIXIR-PT): woody plant and rice datasets.
- The BrAPI network will be merged with WheatIS, a SolR based data discovery system extending the work done in transPLANT.

Interoperability work spans all WP5's resources: identifier normalisation, ontology term mapping, descriptor formalisation for tools and concepts, schema standardisation, and outreach for a wider adoption of defined best practice. WP5 assists through the development of ontologies and their support, MIAPPE-based annotation and validation using the ISAFormat, tools and framework, and services supporting the BrAPI (Breeding API) and its documentation (Table 2).

| WP5 Activities | WP7 Plants and Crops |
|---|---|
| Registries | |
| - FAIRSharing | MIAPPE registered<br>BioSamples, transPLANT, Brassica Information Portal registered |
| - Ontology Lookup Service | Implemented a data import pipeline for the Crop Ontologies<br>Ontology term search, ontology term conflict resolution, through OLS to standardise identifiers |
| - Identifiers.org | Annotates and submits datasets to BioSamples, registered |
| Metadata Services | |
| - Ontology | -- |
| - Annotation / Validation | ISA framework adopted for annotation and validation. |
| - Search | Data Lookup service [19] using Elastic Search and BrAPI, MIAPPE and ISA format (See Deliverables D 7.2, D7.3) |

[19] https://github.com/elixir-europe/plant-brapi-etl-data-lookup-gnpis



Horizon 2020 grant n. 676559

| | |
|---|---|
| - Identifier | -- |
| - Other | -- |
| Resource Markup: Bioschemas | Design of protein, sample and data resource specs.<br>Recommendation for the adoption of Bioschemas to standardise concept descriptions |
| Standards | |
| - Formats, MIMs, Ontologies | MIAPPE v1.0 specification<br>MIAPPE implemented in ISA format<br>Crop Ontology<br>New ontologies: Woody Plant Ontology, Plant Phenotype Experiment Ontology. |
| - Identifiers | Hygiene check to be completed. |
| - API Standards and Standard APIs | BrAPI for plant phenotype/genotype databases to serve their data to crop breeding applications.<br><br>BrAPI specification documented on SwaggerHub and Apiary.<br><br>A BRAPI2ISA conversion is being developed in an ELIXIR Data Validation Implementation study. |
| - Workflow standards | -- |
| Approaches | |
| - Linked Data | MIAPPE 1.1 formalised in OWL and RDF |
| - Workflows | -- |
| Capacity | |
| - BYODS | BrAPI BYOD Hackathon organised by WP7 |
| - Knowledge Hub | TBD |
| Global Engagements | CGIAR, EMPHASIS |

**Table 2: Summary of WP5 for WP7 Interoperability Activities**

*WP8: Rare Disease*

The goal of the Use Case is to create a federated infrastructure that will enable researchers to discover, access and analyse different rare disease repositories across Europe and hence supports the development of new therapies for rare disease. Population-based studies have generated a large volume of data that immediately set the framework of requirements for interoperability services. It does this in partnership with other European infrastructure projects, including RD-CONNECT, BBMRI-ERIC, FAIR-dICT, ODEX4ALL and GO-FAIR.



Horizon 2020 grant n. 676559

The Use Case has developed registries of data resources and tools and a rare disease data linkage plan that allows it to test that tools, models, and protocols to standardise data services in the rare disease domain conform to FAIR data principles. The work has focused on building Node capacity in skills and knowledge for data interoperability, and for access to technical infrastructure that supports Linked Data interoperability working with resources of RD-Connect and the Data FAIRPort toolkit provided by ELIXIR NL and GO-FAIR.

The Interoperability approach has been to use Linked Data Services to answer cross-resource questions such as bridging genomic and phenotypic data for variant identification using machine processable (RDF/XML) representation of the metadata to, for example, integrate rare disease data described with the Orphanet Rare Disease Ontology with common disease and create variant-gene mappings to allow functional evaluation of (rare disease) variants.

Highlights:

- Utilisation of WP5 resources (Identifiers.org, OLS, and Bioschemas) to accommodate concept mapping, and concept linking.
- BYOD (Bring-Your-Own-Data) workshop series to serve the Rare Disease community as a data harmonisation service
- Development of the data linkage plan to testing datasets and tools. The defining of a FAIRification process[20] to extend and generalise the bespoke BYOD service into a best practice guideline for adoption of a BYOD process by a wider community. A FAIRification strategy for datasets and tools has been developed and is in execution phase, to be piloted with local and 'roving' data stewards. The aim is to make rare disease resources interoperable (FAIR) 'at the source' and to speed up the FAIRification process with better tooling and procedures.
- Adoption and co-development of the Data FAIRPort services.

WP5 assists by supporting the data linkage plan, organising BYODs, supporting data stewardship and by supporting the Linked Data Services.

| WP5 Activities | WP8 Rare Disease |
|---|---|
| Registries | Developing a data resources and tools 'Rare Disease' collection in bio.tools[21] of 99 resources<br>Findability solutions are currently Orphanet and the RD-Connect Biobank and registry finder[22]. |
| - FAIRSharing | Used for sharing of general purpose semantic application models, such as HOOM, and semantic application models created in BYODs and in |

[20] Carta, C., Roos, M., Jacobsen, A., Thompson, M., Wilkinson, M.D., Cornet, R., Waagmeester, A., Van Enckevort, D., Jansen, M., Licata, L. and Via, A. (2017) January. The FAIRification of data and the potential of FAIR resources demonstrated in practice at the Rome Bring Your Own Data workshop. In CEUR Workshop Proceedings (Vol. 2042).

[21] https://bio.tools/?page=1&collectionID=%27Rare%20Disease%27&sort=score

[22] https://rd-connect.eu/phenotypic-data/rb-finder-for-registries/



Horizon 2020 grant n. 676559

| | |
|---|---|
| | FAIRification projects as part of the default FAIRification workflow (Deliverable 8.2).<br>Many of the datasets in the bio.tools collection are registered. A full audit and synchronisation is to be completed. |
| - Ontology Lookup Service | Ontologies used are registered.<br>Used by data stewards during FAIRification, and as part of training data managers in BYODs and other training events. |
| - Identifiers.org | Many of the datasets in the bio.tools collection are registered.<br>Identifiers.org's mapping between different URI schemes has proved essential.<br>A full audit and synchronisation is to be completed.<br>Used by data stewards during FAIRification, and as part of training data managers in BYODs and other training events. |
| Metadata Services | |
| - Ontology | See OLS, OxO. |
| - Annotation / Validation | -- |
| - Search | -- |
| - Identifier | A gap has been indicated, for an identifiers mapping service. |
| - Other | Linked Data Services<br>WP8, Task 8.2/3 make extensive use of the Data FAIRport service suite for BYODs and FAIRification projects in the context of the rare disease data linkage plan.<br><br>The FAIRifier is used for linking a semantic application model to existing data and subsequent conversion to a Linked Data representation. Metadata models are applied via the metadata editor that instantiates a FAIR data point that serves the Linked Data. A basic FAIR search tool is available to register FAIR data points. The index is used for a FAIR demonstrator via which predefined cross-resource queries are executed (see Deliverable 8.2)<br><br>EMBL-EBI RDF Platform. |
| Resource Markup: Bioschemas | Design of protein, sample and data resource specs.<br><br>Orphanet adopting Bioschemas (at Biohackathon 2018).<br>Rare disease sample catalogue markup [23]<br>Mark up of multiple patient registries: DEB-Central, CHD7 Database, Microvillus Inclusion Disease Patient Registry, AIP Mutation Database[24]. |
| Standards | |
| - Formats, MIMs, Ontologies | Orphanet Rare Disease Ontology (ORDO), Human Phenotype Ontology (HPO) are recommended for all rare disease resources. Other ontologies |

[23] https://rd-connect.eu/biosamples-data/sample-catalogue/

[24] https://f1000research.com/posters/7-1228



Horizon 2020 grant n. 676559

| | |
|---|---|
| | are used in semantic application models as appropriate for the data set at hand.<br>DCAT and Re3Data are metadata models used for the FAIR Data Point |
| - Identifiers | Hygiene check to be completed. Ideally, identifiers come from an authoritative source that also provides a URL linked to an ontology. Identifiers are to complies with the '10 simple rules ' guidelines in (PLoS 2016) . |
| - API Standards and Standard APIs | FAIR Data Point API. |
| - Workflow standards | -- |
| Approaches | |
| - Linked Data | WP8 Task 8.2/3 Linked Data is used for the description of the metadata of FAIR Data Points, and it is used to make record-level data elements linkable 'at source'. The Data FAIRPort services[25] have been used and co-developed with the Use Case. |
| - Workflows | -- |
| Capacity | |
| - BYODS | Annual BYOD as part of the Rome Summer school for rare disease registry managers. The BYOD has been and will continue to be co-funded by partners in the rare disease community. Incidental BYODs include a BYOD with registry software providers that started an activity to develop architecture to make existing software FAIR data producing tools.<br><br>8 WP5 BYODs supported the development of the FAIRification process for WP8. |
| - Knowledge Hub | TBD |
| Global Engagements | RD-Connect, BBMRI-ERIC, GO FAIR, NIH, FAIR-dICT, ODEX4ALL |

**Table 3: Summary of WP5 for WP8 Interoperability Activities**

*WP8: Human Disease*

WP9 has not been the primary focus of WP5. However, the Bioschemas beacons specification is proposed to self-describe a beacon's genetic variant cardinality service for better integration with other beacons within the beacon-network. It builds upon the Beacon service API and uses existing schema.org entities and properties. The European Genome-phenome Archive (EGA) live deploy is marked up with Bioschemas.

***Issues Arising***

Issues have emerged from Interoperability activities in the Use Cases that will be addressed in future work (discussed in detail in Appendix 2). These include:

[25] https://www.dtls.nl/fair-data/find-fair-data-tools/



Horizon 2020 grant n. 676559

- **Identifier associations**: the need for identifier associations (connecting relevant concepts), and for identifier mappings (coupling of similar/identical concepts) to enable the framework for Linked Open Data for WP8's Rare Disease community's research.
- **Workflow discovery and quality**: the discovery, quality control and stewardship of workflows is raised as an issue by WP6 Marine Metagenomics, including assessment of quality for CWL tool description and the tool parameters.
- **Non-standardised API development methods**: Although we recommend API standardisation practices in reality only a subset of ELIXIR's resources adhere and there is poor understanding of good API practice. Most WPs raise this point.
- **Referenced-based taxonomy classification**: the systematic handling of taxa across samples and datasets is identified as a significant interoperability challenge by WP6.
- **Data curation burden**: the burden for data curation, missing metadata annotation in data records etc is recognised by all WPs. WP8, in addition, identifies the labour and expertise challenges to FAIRify datasets at source as part of their FAIRification process reported in Deliverable D8.2.
- **Best practice**: finding best practice advice, templates and examples.

### A1.3.2. Data (Task 5.1.2)

**WP3 (Data)** takes responsibility for the Core Data Resources (CDR) and Deposition Databases (DD) of ELIXIR. With WP5 we support steps towards interoperability. CDR, DD and Node Data Resources (Databases and Knowledge bases) have major roles in the Use Cases and are thus subject to Interoperability tasks and best practice.

WP5's work has focused on:

- Ubiquitous, minimal web-based mark-up of resources with machine processable metadata for automated harvesting, indexing, citation and processing by apps, aggregators and registries. **Bioschemas** covers the mark-up of provenance, versioning/release and licensing information as well as data type specifics using conventions of the web-standard schema.org.
    - The data resource level DataCatalog and Dataset, which cover universal properties of data resources and should apply to all WP3 CDR and DD databases.
    - Five Core Data Resources (CATH, EGA, Human Protein Atlas, MINT, PDBe) and three Deposition Databases (BioSamples, PDBe, EGA) are marked-up so far in their Live deploys. Many more are in the pipeline. InterPro and UniProt have been very active in the specifications with markup in their pre-production pipelines but not yet their live deploys.
- Selection process of **Recommended Interoperability Resources** that mirrors and adapts the WP3's CDR selection process
- Recommendations have been made for **standardised identifier practices**[26], in cooperation with EU CORBEL project, and our simplified identifier hygiene checklist will be used to check all CDRs and DDs.

[26] https://doi.org/10.1371/journal.pbio.2001414



Horizon 2020 grant n. 676559

- Registries for standards, ontologies and identifier resolution used by WP3's CDRs and DDs. All CDRs and DDs are registered in FAIRsharing and Identifiers.org.
- Work is ongoing to audit CDR and DDs for API standardised documentation and best practices.
- CDRs and DDs are being examined for compliance to "FAIR Metrics" in an ELIXIR Implementation Study [27] that is complementary to EXCELERATE.

### A1.3.3. Node Capacity Building (Task 5.3.3)

**WP10 (Node Capacity Building)** aims to build capacity in the Data Nodes to disseminate interoperability practice. The Interoperability Services are provided by the Nodes.
Highlights:

- A Recommended Interoperability Resource (RIR) selection process open to all Nodes and audit of Node resources to be reported in D5.2. Currently only 3 Nodes (UK, EBI and NL) provide the Interoperability registries and metadata services used by WP5.
- Registration of Node resources in Interoperability Registries: 57/69 eligible registered in FAIRSharing; 59/70 eligible registered in Identifiers.org.
- 12 Nodes involved in the Bioschemas initiative.
- Node Staff Exchange programme awarded to support Node uptake of Bioschemas.
- 26 Capacity Building workshops
- 8 BYODs for WP8, 1 BYOD WP6, 1 BYOD WP7

## A1.4 Global Engagement (Task 5.1.3)

To adhere to our third principle **Interoperable Interoperability** we have systematically engaged with global initiatives and sought to adopt pre-existing and emerging standards rather than "roll our own". WP5 has participated in over 60 meetings in the USA, Europe and Asia with a reach of 1500+ people.

Throughout Appendix 1 and 2 in situ references will be made to Global Engagements and Collaborations. Appendix 4 gives a matrix of WP5 activities to the global initiatives and standards bodies we have collaborated with.

A summary of the chief stakeholders and players is given below.

| **Global engagement** | **WP5 activity** |
|---|---|
| Research Data Alliance<br>rd-alliance.org | Metadata Interest Group on identifier properties<br>FAIRsharing Working Group |
| GA4GH<br>ga4gh.org | Common workflow Language, workflow portability, API standardisation and tools markup |
| Pistoia Alliance<br>pistoiaalliance.org | Ontology mapping |
| NIH Data Commons<br>commonfund.nih.gov/commons | FAIRSharing, FAIR Metrics, Identifier resolution, research object portability |

[27] https://www.elixir-europe.org/platforms/data/fairness-core-resources



Horizon 2020 grant n. 676559

| | |
|---|---|
| Force 11<br>Force11.org | FAIR Metrics, Identifier resolution and standards |
| Common Workflow Language<br>Commonwl.org | Common workflow Language, workflow interoperability and portability, API standardisation and tools markup |
| Galaxy Community<br>galaxyproject.org | Common workflow Language, workflow portability, API standardisation and tools markup |
| FDA BioCompute Object<br>osf.io/h59uh | Common workflow Language, workflow interoperability, API standardisation and tools markup |
| FAIRmetrics.org / GO-FAIR | FAIR metrics |
| Research Object<br>Researchobject.org | Common workflow Language, workflow interoperability and portability, API standardisation and tools markup, metadata standards |
| EU H2020 project collaborations | |
| CORBEL<br>Corbel-project.eu | Identifier resolution and standards |
| EOSCPilot<br>Eoscpilot.eu | Identifier resolution and standards, Data Catalogue metadata exchange (EDMI), bioschemas |
| BioExcel<br>Bioexcel.eu | Common workflow Language, workflow interoperability, API standardisation and tools markup |
| FREYA<br>project-freya.eu | Identifier resolution, standards and citation |
| openAIRE<br>openaire.eu | Identifier resolution, standards and citation, Data Catalogue metadata exchange (EDMI), bioschemas |
| CHARME Cost Action<br>cost-charme.eu | FAIR services |
| RD-Connect<br>rd-connect.eu | Registries, FAIRification of datasets |
| Commercial collaborators | |
| Google / Schema.org | Bioschemas |
| DataCite<br>Datacite.org | Identifier resolution, standards and citation, Bioschemas |
| Springer Nature Scientific Data<br>nature.com/sdata | Identifier resolution, standards and citation |
| Centres | |
| California Digital Library<br>www.cdlib.org | Identifier standards and resolution, Bioschemas |
| NIH BD2K CEDAR Center<br>metadatacenter.org | FAIRSharing, ontology, validation and annotation services |
| NIH BD2K BioCADDIE Center<br>biocaddie.org | FAIRSharing, ontology, validation, annotation and indexing services, Bioschemas |

**Table 4: Global Engagement**

Three H2020 projects warrant special mention:



Horizon 2020 grant n. 676559

- **CORBEL** (http://www.corbel-project.eu). This project, which brings together 14 BMS Research Infrastructures, has been a close collaborator with EXCELERATE, particularly with regard to work on identifier standards[28]. CORBEL extends EXCELERATE's work to BioBanking, imaging and other domains to examine service and identifier needs and is informed by service provision and gap analysis performed in collaboration with EXCELERATE. The two projects address cross domain identifier use cases such as sample provision from BioBanking (BBMRI, CORBEL) into Rare Disease projects (WP8, EXCELERATE).
- **The European Open Science Cloud Pilot** (https://eoscpilot.eu). The WP members have been active in the EOSCPilot, notably the Data Interoperability WP and the specification and piloting of the EOSC Datasets Minimum Metadata Guideline (EDMI)[29]. The WP has also actively contributed to the consultations of the HLEG Rules of Participation and the HLEG FAIR Data Action Plan.
- **BioExcel Centre of Excellence** (https://bioexcel.eu). The WP work closely with BioExcel on workflow interoperability and portability, and standardisation and tools that support the Common Workflow Language, notably to support WP6.

## A1.5 Relationship to other WPs

- WP1 and WP2 (Tools): Cooperations have focused the use of Bioschema by the bio.tools registry and OpenEbench benchmarking initiative; integrations between the Interoperability registries and bio.tools; and markup of tools with the EDAM ontology for workflow discovery, annotation and validation.
- WP11 (Training): BYODs are reported in D5.3. Other co-operations have focused on the use of Bioschema by the TeSS Portal; integrations with the Interoperability registries (notably FAIRSharing) and the registration of WP5 events in the TeSS.

## A1.6 Relationship to other ELIXIR User Communities

During EXCELERATE the number of ELIXIR Communities has widened. Work undertaken in EXCELERATE has been transferred to these communities, notably:

- **Galaxy:** the community associated with the most popular workflow analysis platform. Wp5 relevant work includes: workflow standardisation, tool description standardisation, EDAM extensions (undertaken by WP1) and native ISA tools support for Galaxy.
- **Metabolomics:** ISAtools power the submission tool of the ELIXIR DD MetaboLights.
- **Structural Bioinformatics, Microbial Biotechnology and Metabolomics:** are active users of workflows and will benefit from our work with WP6 in workflow development and standardisation.

[28] Goble et al CORBEL Review of identifier schemes, standards and interoperability maps and proposed harmonization strategy https://zenodo.org/record/376236#.W4QpuLhG318

[29] https://www.eoscpilot.eu/sites/default/files/eoscpilot-d6.3.pdf

# Appendix 2: Interoperability Implementation Services (Task 5.2)

## A2.1. Services, Metrics and Maturity Models

The efficient function of the EIP requires a core set of services that deliver and support interoperability components. These existing services must be consolidated to support: machine processable identity, data formats, experimental reporting guidelines, knowledge representations, and resource operational practices for transparent releases, versioning, provenance, updates. Prioritized services were first reported in MS26 (Roadmap) and these are now being extended collaboratively.

### A2.1.1 FAIR Services Framework

The service framework sets out the interoperability services needed to operate FAIR data and assist the implementation of the FAIR-data needs of the Use Cases.
Progress has been made towards the following primary objectives:

- identification of services integral to EIP (core infrastructure requirements)
- establishing a service selection process for EIP (assess current and candidate)
- identifying gaps in services for the current model (gap filling as required)

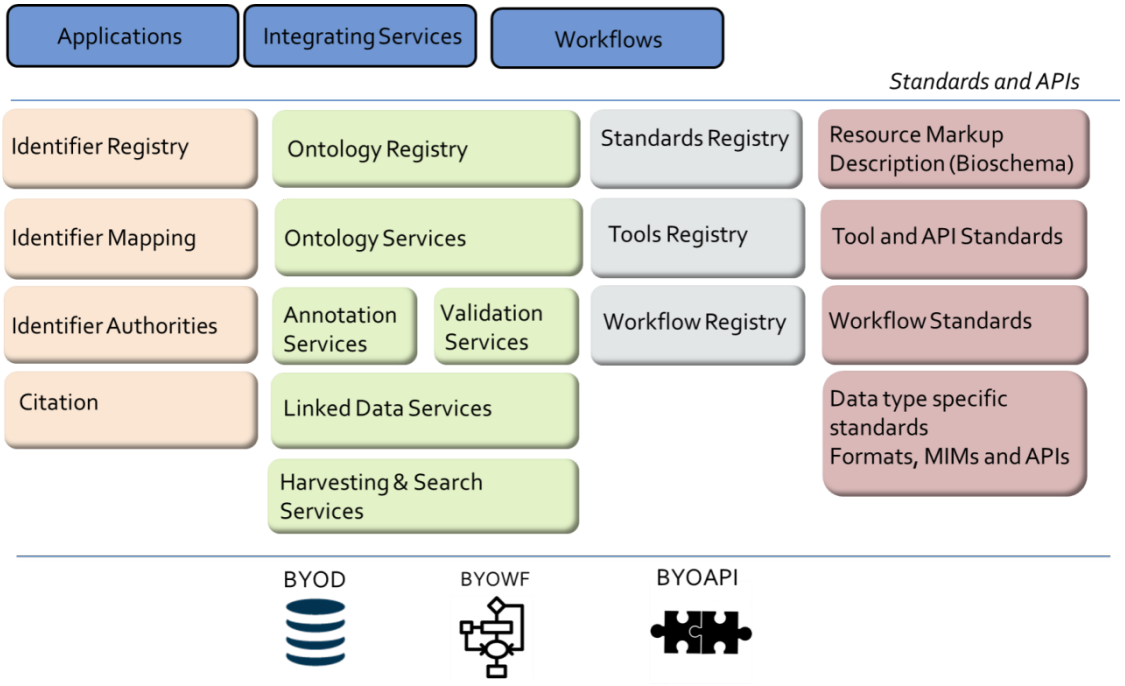


**Figure 3: Services Reference Framework**

***Results***

A Service Reference Framework (Figure 3) has been developed and refined to organise our services. Services chiefly comprise those 'owned' by ELIXIR (being in a Node's Service Delivery Plan (SDP), or are undergoing preparation to become part of SDP), ELIXIR Commissioned Services. Reports on external 3rd party services (independent of



Horizon 2020 grant n. 676559

ELIXIR, but deemed critical for the success of ELIXIR) are being compiled and will be delivered under a separate deliverables (D5.2).
Ongoing work has highlighted the need not only to identify services that are essential to the operation of the EIP, but also a way to recognise the importance and reliance upon those services, and determine which need to be sustained beyond the limits of ongoing projects. In part, this requires an objective (FAIR and measurable) set of criteria by which existing, current, and future candidate services may be evaluated, particularly with respect to their suitability for incorporation into the interoperability platform).
Meanwhile, an interim working set of mission critical Interoperability Registry Services have been identified (FAIRSharing, Ontology Lookup Service, Identifiers.org) and a small set of Metadata Services identified as valuable to the Use Cases and for future ELXIIR activities. These are further described below. A currently limited view of these Interoperability services (ELIXIR website).

Services outside the scope of ELIXIR framework are also critical to FAIR activities such as secondary services that underlie the operation of the RIR services. An ongoing audit and gap analysis of services (i) offered by ELIXIR nodes (visible through the ELIXIR website), (ii) available externally, and (iii) mirror services fulfilling a similar task, to achieve cross stack interoperability and resilience (e.g. BD2K/Force11 and ELIXIR identifier resolution services) will be reported in D5.2 (service landscape examination and business cases).

Collectively, these services (RIR recommended, secondary) will build the interconnected FAIR service infrastructure capability.

Dissemination of ELIXIR interoperability services through has been extensive, through presentations at meetings, posters and through webinars (See Appendix 6 and 7).

Recommended Interoperability Resource (RIR) process
A robust objective service selection process is demands services demonstrate their capability in facilitating studies that integrate, reuse, and (re)analyse dynamic data in a meaningful reproducible way. Suitable criteria and a robust objective selection process for Recommended Interoperability Resources, analogous to the CDRs of WP3 have been defined. The first open call closed July 2018 and the first list will be available end 2018.
The Recommended Interoperability Resource (RIR) selection process and accompanying documents (criteria and case) for publication, subsequent to which there has been the first open call for candidate submissions for 2018 in June. The call will be open until mid-July 2018 before undergoing the review process aiming for the Heads of Nodes' approval in December 2018. The RIR selection process is built around key properties/criteria; the service must be fully sustainable by being part of an ELIXIR Node Service Delivery Plan (SDP) or ELIXIR-Commissioned Work, and must serve an interoperability requirement that either exists or is forecast to exist in the ELIXIR Interoperability Roadmap. The selection process itself, based on these documented criteria, follows standard procedures, as practiced for scientific appraisals, incorporating independent and external 3rd party evaluators. The call for submission of the candidate services to be considered for this


Horizon 2020 grant n. 676559

selection was announced through the usual ELIXIR channels (EIP website, Key contributor announcements, and community dissemination such as through mailing lists). The open call process will live in a dynamic space, where periodic calls will be made for new service submissions. Appraisals will also be performed for previously selected services, to ensure they continue to meet the required standards.

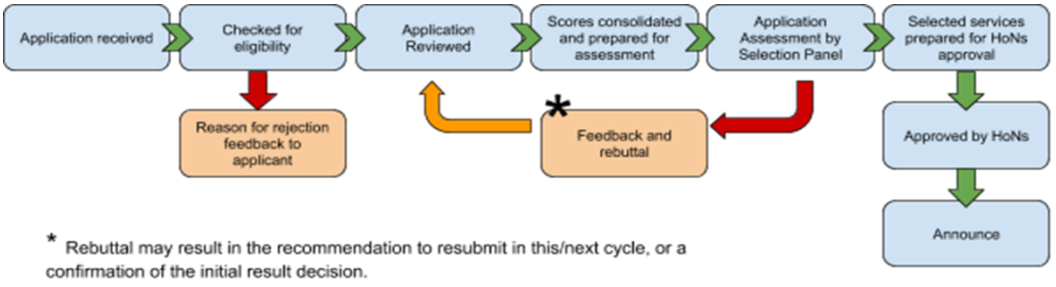


**Figure 4: RIR Selection Process Summary**

For the purposes of the selection process, an ELIXIR Interoperability Resource is defined as a technical or administrative service provided through an ELIXIR Node or ELIXIR Commissioned Service agreement, for use by the Life Sciences community. This set of services, together, will form the backbone of the ELIXIR EIP infrastructure. EIP services could include, for example, those that map between reference standards, generate or process metadata, or container-ise appropriate software. Interoperability services may, besides function, be categorised by role; they may act as a 'glue' service, to provide a stopgap solution prior to the implementation of a sustainable solution, or may provide a meta-service required in infrastructure scale up. Other interoperability services may be more closely tied to existing or developing interoperability components, such as registries, metadata (description though Bioschemas), or with frameworks (e.g. Intermine, MOLGENIS).

### A2.1.2 Gap analysis

A defined interoperability service landscape enables the identification on gaps in the EIP, for instance enabling necessary services to be placed on our roadmap for future solicitation or development, either though ELIXIR, or in collaboration with external partners. Some highlighted gaps (services, resources and procedures/processes) have been identified:

- *Identifier mapping*: No services at the production and maintenance level that fulfils the need to map identifiers between datasets, i.e. an identifier in one dataset and its relationship to a different identifier in a different dataset.
- *Process for maturation of EIP components*: There is currently no development plan to coherently address the differing maturity levels of interoperability components, and drive their performance to production quality.
- *Standards mapping:* Need to support and sustain the mapping of standards (models, formats, reporting guidelines, identifier schema and ontologies) across the ELIXIR ecosystem to ensure greater understanding of standards provision and relationship between standards and the knowledge-bases and repositories that implement them.



Horizon 2020 grant n. 676559

- *Data type specific services:* for example: synonym searching, chemical substructure matching, identifier resolution for chemical entities and image annotation with links to ontology lookup.

***Future Work***

The first round of Recommended Interoperability Resources completed and process review and the audit and gap analysis of services (D5.2) needs to be completed. We need to improve the listing of interoperability services in the ELIXIR web site and create comprehensive entries on the EIP Knowledge Hub. A plan to address Gaps is to b developed.

An Interoperable Interoperability Services Interactions Profile needs to be completed reflecting various interoperability scenarios that have emerged through the EXCELERATE work. For example, compact identifiers in publications (Citation) resolve to resources with metadata marked up with Bioschemas (Dataset Description), and which can be harvested and viewed through services from Identifiers.org (Identifier Registry and Resolution).

An Interoperability Resource Maturity Model is needed to determine the 'state' of individual services, with a view to progressing them in their lifecycle (maturation, productionisation, sunsetting, etc. Figure 5 is our starting point.

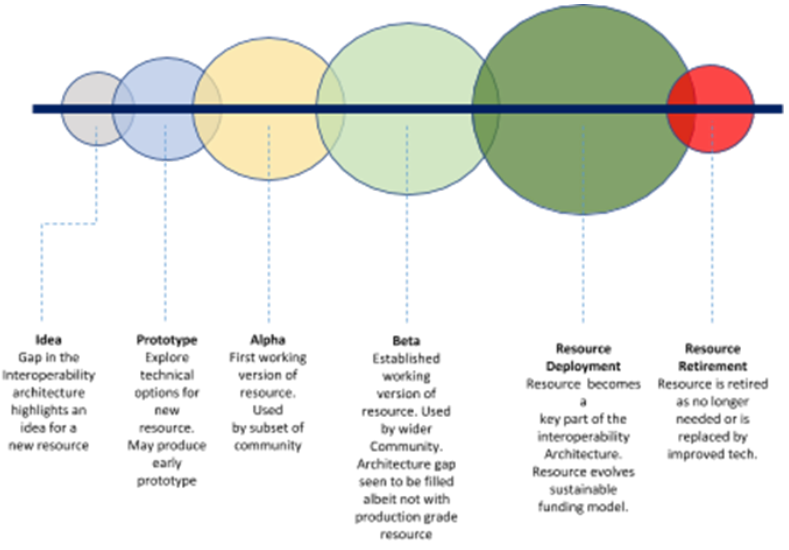


**Figure 5: Recommended Interoperability Resource Maturity Model**

### A2.1.3 Metrics and Maturity Models for Data and Processes

FAIRification of data resources (the adoption of FAIR principles by those resources) necessitates a framework for evaluating "FAIRness". While each of the FAIR principles may differ in their interpretation in different contexts, each of the principles themselves should be measurable in some way within a specific domain, for a specific resource. Work has been done to develop a mechanism which can be used to suggest which criteria could be used to generate metrics to evaluate the FAIRness of a resource[30]. For

[30] https://identifiers.org/doi/10.1101/225490



Horizon 2020 grant n. 676559

instance, FAIRsharing is working with a number of groups (including the NIH Data Commons KC1) to develop FAIR metrics and a FAIR rubric, drawing on resource metadata provided by FAIRsharing.

The GO-FAIR FAIRmetrics Group[31], whose members are also members of WP5, has developed the first steps towards a design framework and exemplar metrics[32]. A preliminary FAIR Capability Maturity Model Integration (FAIR-CMMI) framework has been developed (figure 6) to help quantify the investment needed to reach different levels of FAIR and assess the gaps to reach different FAIR maturity levels.

| Level | Process | Datasets and Linksets |
|---|---|---|
| 1 Initial | Processes are disorganized.<br>Success is dependent on specialised, heroic and one-off efforts, considered unrepeatable, because processes are not sufficiently defined and documented to be replicated. | Datasets are disorganised and may well be unstructured.<br>No Linksets (i.e. explicitly published mappings between datasets). |
| 2 Repeatable | Basic processes are established, created and maintained.<br>Successes could be repeated, because the processes are defined, and documented. | Basic levels of FAIR are implemented by the dataset.<br>Linksets are implied as the links are intermingled with the data. Descriptions of the links are not available. |
| 3 Defined | An organization has its own process through greater attention to documentation, standardization, and integration. | Datasets have limited metadata and access capabilities.<br>Linksets are identified, i.e. there are descriptions of the datasets that are linked to. |
| 4 Managed | Organization monitors and controls its own processes through data collection and analysis | Datasets have further metadata and access capabilities.<br>Linksets are explicitly managed (but can be idiosyncratic). |
| 5 Optimizing | Processes are constantly being improved through monitoring feedback from current processes and introducing innovative processes to better serve the organization's particular needs | Datasets are fully annotated with metadata and access capabilities, i.e. they fully satisfy all the FAIR principles.<br>Linksets are managed as first class objects, i.e. regarded as datasets in their own right, and accessible to mapping services |

**Figure 6: FAIR-CMMI Levels** - moving from present to future data integration

***Results and Future Work***

BYODs were originally embedded within WP8 Rare Disease use case and have evolved from “one-off” heroic efforts into a generalisable methodology for processing FAIR data, becoming an interoperability foundation. This methodology could be employed to develop new software, and guide best practice adoption, as well as address use cases beyond the original remit of Rare Disease.

[31] http://fairmetrics.org

[32] https://www.biorxiv.org/content/early/2017/12/01/225490



Horizon 2020 grant n. 676559

Leveraging upon the activities of our CORBEL partners, we will follow guidelines proposed through the FAIR capability maturity model integration (FAIR-CMMI). This will include metrics that (a) quantify the investment needed to reach different levels of data discovery, accessibility, interoperability and reusability and (b) evaluate how FAIR data are and will subsequently become (success metrics). There will be a need to establish a formal collaboration for this to happen. A FAIR-CDR ELIXIR Implementation Study has been funded to put proposed FAIR Metrics into practice for the ELIXIR community by starting to "FAIRify" ELIXIR Core Data Resources, and an IMI2-2017-12-02 FAIRPlus IMI proposal (802750-1) has been accepted to further develop FAIR Metrics and FAIR Capability Maturity Model Integration (FAIR-CMMI) framework building on these foundations.

The FAIR-CDR ELIXIR Implementation Study proposes to put FAIR Metrics into practice by starting to FAIRify ELIXIR Core Data Resources ArrayExpress, ENA, PDBe, PRIDE, CatH, CHEMBL, ChEBI, UNIPROT, HPA, INTERPRO, MINT, and STRING-db. The study will first establish effective guidelines for implementation and then involve hands-on FAIRification workshops, in which FAIRness will be assessed before and after the work done.

## A2.2 Metadata

Registries are data resources collecting and integrating metadata from several resources to facilitate the discovery of third party resources. The ELIXIR Hub has created the ELIXIR Registries Working Group to implement a common registry strategy that interlinks registries to drive findability and contextualisation, and exploit registries and their applications. In addition they have undertaken a preliminary analysis of the registries landscape.

Registries play a key role within ELIXIR, the EIP, and potentially within EOSC. Of the 20 registries supported by ELIXIR nodes, 10 are already supported by Service Delivery Plan. Recent analysis revealed coverage of the major content types (databases, datasets, cloud resources, containers, tools, workflows and user credentials) and identified potential dependencies on other external registries describing resources like 'Virtual Machines' and 'cloud resources', which are not currently addressed. .

One of the objectives in the EIP is to better integrate information from existing registries, allowing the sharing of equivalent content (from the same source, and exchanged across registries), use of similar standards (common vocabularies and guidelines), and better inter-linking of information across registries.

An interim working set of Interoperability Registry Services have been identified, provided by an ELIXIR Node and subject to a Service Delivery Plan (See table 5):

- **Standards registry: FAIRSharing** (http://fairsharingorg)(ELIXIR-UK). A registry that provides an interlinked summary of available reporting guidelines, models/schemas and terminologies used within individual repositories, as well as funding sources. FAIRsharing is featured in the EU FAIR Data Experts' Report on Turning FAIR Data into Reality.

- **Ontology registry: Ontology Lookup Service** (https://www.ebi.ac.uk/ols) (ELIXIR-EBI). A portal for biomedical ontologies that provides a single point of access to the latest ontology terms and versions, as well as providing ontology based services such as mapping of terms between ontologies, and providing tools to request new and visualise existing terms.
- **Identifier registry: Identifiers.org** (http://identifiers.org) (ELIXIR-EBI). A registry that provides unique, stable, resolvable and location independent compact URIs to identify and locate scientific data, as well as services to request alternative locations for equivalent data records, and to map between alternative identifier schemes (using the same entity identifier).

In addition three other cross-WP registries have been identified as important to WP5

- **Training registry** of materials and BYODs: The TeSS Portal (WP11): see deliverable D11.2.
- **Tools registry**: bio.tools with EDAM markup (WP1): see deliverables D1.3, D1.7 and D1.8.
- **Workflow registry/repository for CWL-compliant workflows** (with WP1). The workflow registry has become increasingly important as workflow-based interoperability has gained prominence in ELIXIR Use Cases (pioneered by WP6). Work is underway with WP1 (Tools Platform) to plan and develop a CWL workflow registry based on prior developments, notably Biocontainers developed as part of WP1, CWL Viewer developed by the EU BioExcel project, and myExperiment which is supported by members of ELIXIR-UK. Future development will be principally undertaken as part of H2020 EOSCLife project. Some work has already been undertaken to link CWL Viewer with bio.tools using EDAM ontology markup.

***Results***

The Interoperability Registry Services (FAIRSharing, Ontology Lookup Service and Identifiers.org) are currently undergoing formal evaluation as RIRs using the processes outlined above. All are actively used by all the Use Cases, have near to complete coverage of the WP3 Core Data Resources and Deposition Databases and widespread use by the Node Resources (Table 5). All registries are active in our Bioschemas resource markup activity.

All registries have undertaken work to interoperate their services to offer an integrated metadata registry integrate with external registries (Table 6) and made software updates and feature improvements to support interoperation and the Use Cases.

FAIRSharing and Identifiers.org are pilots for the EOSC Datasets Minimum Metadata Guideline (EDMI)[33] developed by EOSCPilot. WP5 has been highly active in this work, to develop simple metadata exchange mechanisms and elements between catalogues.

| | |
|---|---|
| FAIRsharing | http://fairsharing.org Provided by ELIXIR-UK Node |

[33] https://eoscpilot.eu/sites/default/files/eoscpilot-d6.3.pdf



Horizon 2020 grant n. 676559

| | |
|---|---|
| Description | Provides an interlinked summary of available reporting guidelines, models/schemas and terminologies used within individual repositories, as well as funding sources.<br><br>Endorsed by a community of over 68 organizations, including publishers (embedded in the data policies of 600 Springer Nature's journals, PLOS, EMBO press, BMJ, F1000Research, BioMedCentral, Oxford University Press, Wellcome Trust Open Research and others), standardization groups, and research data management support initiatives and libraries (such as those at JISC, Stanford, Cambridge and the Oxford Universities). |
| Global engagement | Operates as a Working Group in the Research Data Alliance and Force11 and has recently released their recommendations, signed by a number of adopters.<br>NIH Data Commons[34]<br>Featured in the EU FAIR Data Experts' Report on Turning FAIR Data into Reality<br>Pilot for the EOSC Datasets Minimum Metadata Guideline (EDMI)[XX] developed by EOSCPilot. |
| ELIXIR profile | ELIXIR resources are organised into a named and searchable FAIRsharing collection.<br>Of 1067 resources registered, 57 are ELIXIR node contributed resources.<br>All ELIXIR CDRs and DDs are registered (WP3).<br>57 /69 eligible ELIXIR Node Data Resources with SDPs are registered |
| WP5 developments | An interactive network graph of standards, databases and policies in collections and publisher recommendations, an embeddable widget to displays any collection or recommendation on a 3rd party website; DOIs minting for each FAIRsharing record allow formal citations; and an application ontology has been developed to classify and search records by disciplines and topics. |
| API | REST Read API<br>SWAGGER documentation |
| Bioschemas | DataCatalog and DataSet Bioschemas markup in FAIRSharing web pages<br>FAIRSharing harvest dataset and ontology markup. |
| **Ontology Lookup Service** | **https://www.ebi.ac.uk/ols Provided by EMBL-EBI Node** |
| Description | A portal for biomedical ontologies that provides a single point of access to the latest ontology terms and versions, as well as providing ontology based services such as mapping of terms between ontologies, and providing tools to request new and visualise existing terms. |
| Global engagement | The Pistoia Alliance Ontologies Mapping project |
| ELIXIR profile | 200+ Ontologies registered |

[34] https://commonfund.nih.gov/commons/awardees



Horizon 2020 grant n. 676559

| | |
|---|---|
| | Prime registry for the ontologies used by WP3 and WP6-9 Use Cases, including those of the Crop Ontology project to support the annotation of agricultural resources.<br>Enseqlopedia incorporated into OLS to marine sequencing application information (WP6) |
| WP5 developments | A data import pipeline for the Crop Ontologies to support the Plant Use Case (WP7).<br>Ontology browsing implementation extended to cover non-OBO compliant ontologies into a harmonised view with OBO ontologies. |
| API | REST Read API<br>Home-grown documentation |
| Bioschemas | Bioschemas markup on the Ontology Lookup Service pages connects to Bioschemas ontology markup in the BioSamples web pages.<br>Harvesting bioschema markup of ontology usage on the Web feeds into the Zooma knowledgebase to improve OLS search and ranking and promote the widely used ontologies |
| **Identifiers.org** | **http://identifiers.org Provided by the EMBL-EBI Node** |
| Description | A registry that provides unique, stable, resolvable and location independent compact URIs to identify and locate scientific data, as well as services to request alternative locations for equivalent data records, and to map between alternative identifier schemes (using the same entity identifier).<br><br>Identifiers.org records all known resource specific access URLs for a given scientific record, as well as document alternative URI schemes through which datasets can be accessed. This information is crucial for integration activities, as it allows numerous, but equivalent, URIs and URLs to be considered as a single (Identifiers.org as the canonical URI) lexical string.<br><br>Registry covers 655 collections (814 resources) and is embedded in key data indexing services such as omicsDI.org. Offers significant help to publishers and others implementing persistent, machine-resolvable citation of research data in compliance with emerging science policy body recommendations and funder requirements. |
| Global engagement | RDA Metadata Interest Group on property definitions for license and identifier<br>Pilot for the EOSC Datasets Minimum Metadata Guideline (EDMI) developed by EU H2020 EOSCPilot.<br>Co-operation with N2T.net for Force11 and NIH Data Commons.<br>Co-operations with EU H2020 FREYA for PID infrastructure. |
| ELIXIR profile | All ELIXIR CDRs and DDs are registered.<br>59/70 of eligible ELIXIR Node Resources are registered on identifiers.org |
| WP5 developments | Harmonisation with its sister service in the USA n2t.net (supplied by California Digital Library for NIH BD2K and Force11)[35] .<br>A Nature Scientific Data editorial announced the adoption of harmonised resolution services. This work is jointly undertaken with the EU CORBEL project.<br><br>Identifiers.org has recently been extended to support giving credit to individual data providers/databases through the compact identifier system. This |

[35] https://www.nature.com/articles/sdata201829

| | |
|---|---|
| | supports the Force11 JDDCP (Joint Declaration of Data Citation Principles), released in 2014, for the publishing of research objects, targeted particularly to scholarly literature. |
| API | Cloud API Library, REST Read API<br>Home-grown documentation |
| Bioschemas | DataCatalog and DataSet Bioschemas markup in Identifiers.org web pages<br>Alpha service for harvesting Bioschemas markup into identifiers.org. |

**Table 5: ELIXIR Resource coverage by Interoperability Registry Services**

| | **FAIRSharing** | **Ontology Lookup Service** | **Identifiers.org** |
|---|---|---|---|
| FAIRSharing | - | n/a | Done |
| OLS | On-going | - | Done[36] |
| Identifiers.org | On-going | On-going[37] | - |
| TeSS | Done | | Under discussion |
| Bio.tools | On-going | EDAM mappings[38] | EDAM mappings |
| External registries | BioPortal,<br>OBO Foundry<br>NBDC Integbio catalogue<br>AgriPortal (ELIXIR-FR, WP6), | n/a | N2T.net[39], BioPortal |

**Table 6: Registry Interoperability Matrix**. A workflow registry is to be adopted and further developed in EOSC-Life and part of the 2019-2023 ELIXIR Workplan.

***Future work***

All registries are to work to improve the interoperation of their services to offer an integrated metadata registry infrastructure, and to complete the implementation of the Registry Interoperability Matrix. Following on from the EXCELERATE WP5, work to integrate FAIRSharing with bio.tools and the TeSS is supported by an ELIXIR Implementation Study (see Table 2). Work is needed to fully align the registries metadata, for example, mapping terms from the FAIRSharing application ontology to the bio.tools EDAM ontology and identifiers.org. Work is also underway to put in place joint standardised procedure/practices between Identifier.org and the USA's N2T.net service.

We aim for universal registration of all ELIXIR resources in the Interoperability Registries and universal adoption and active use of Bioschemas by all Interoperability Registries.

Each registry has future work plans arising from WP5 activities:

- Identifiers.org's harmonisation process with n2t.net is being refined to put in place standardised procedure/practices to define how the request process is managed by collaborating parties, and to define its precise scope (for instance, a strategy on

[36] OLS is listed as a resolving location for all ontologies listed in identifiers.org
[37] Identifiers.org and OLS are in the process of synchronising their recommended canonical URIs on a per ontology basis
[38] Mapping the bio.tool EDAM Ontology to OLS and the Identifiers.org data collections.
[39] Identifiers.org and N2T.net act as resolvers for an agreed set of prefixes serving life sciences



Horizon 2020 grant n. 676559

identifier assignment for transitory datasets and their currently used temporary identifiers, and deciding appropriateness of assigning prefix schemes to transitory projects, specifically how this should be verified or assessed). The newly implemented compact identifier system needs to undergo maturation including decentralisation / 'cloudification' of data and associated services and adoption to targeted communities promoted. Finally, improved metadata including harvest/display (bio)schema.org for findability of resources; improved rdf-typing of identifiers.org to facilitate Linked Data use cases and applications, and incorporation of EDAM (or other relevant) terms to better standardise 'types' in identifiers.org (replace or supplement current 'tag' mechanism), through API tools.

- OLS plans to provide access to ontology terms using alternative prefixes and provide a canonical URI for ontology terms (in collaboration with identifiers.org), provide access to historic versions of ICD, which use different identifiers in subsequent versions and include topic and species metadata on ontologies to support filtering and faceting of search within OLS for specific communities.
- FAIRsharing.org has potential for supporting the modelling step of the FAIRification process pioneered by WP8. Data stewards first search and reuse application models (application ontologies) that were used in previous FAIRifications or are commonly used in FAIR data repositories like the ELIXIR recommended repositories. FAIRsharing.org offers features to store and search such application models. FAIRsharing.org can support this step more specifically.

WP6 reports one of its outstanding challenges to be workflow description discovery and the assessment of quality for CWL tool description and the tool parameters. The development and stewardship of a workflow registry is to be undertaken in the forthcoming H2020 EOSC Life project, commencing January 2019.

## A2.3 Metadata Services

Metadata is a key strategy leveraged by the EIP to enable and improve FAIRification of data and services. Metadata services cover the range of services needed for ontology and identifier management and mapping, resource annotation, annotation validation and semantic search (see Service Framework). Some services have been already been identified and used in WP6-9. Work has chiefly focused on ontology, annotation/validation and linked data services driven by the needs of the Use Cases.

***Results***

An ongoing full audit and gap analysis of services offered by ELIXIR nodes is underway and results will be reported in D5.2. Meanwhile several EIP metadata services have emerged as providing important services for the Use Cases and our own Interoperability Registries (Table 7).

| | |
|---|---|
| Ontology services | In addition to OLS, the following services are offered via the ELIXIR-EMBL-EBI Node:<br>- OxO, views relations between ontology terms in different vocabularies. The latest release includes 1.8 million ontology |



Horizon 2020 grant n. 676559

| | |
|---|---|
| | mappings. A release of a Python library for predicting new ontology mappings uses OLS and OxO and includes a number of newly curated mappings between disease ontologies relevant for WP8 and WP9. This work was developed as part of the Pistoia Alliance Ontologies Mapping project.<br>- Zooma & Webulous aid in annotation procedures. Zooma maps data to ontologies, whereas Webulous supports extensions to ontologies via online spreadsheets. Zooma contains over 100,000 manually curated text to ontology mappings from 9 databases that can be used to predict ontology annotations. The Pistoia Alliance Ontologies Mapping[40] project produced a systematic evaluation of existing tools and algorithms.<br>- Slim-O-Matic, is a semi-automated tool for GO slimming used by WP6. GO slims are cut-down versions of the GO that contain a subset of terms. Slims are useful for summarising the processes or functions mediated by groups of proteins |
| Annotation and validation services | Services needed to annotate and validate marked up resources against expected minimum information models (aka reporting guidelines)<br>The ISA framework provided by the ELIXIR-UK Node). ISA is built around the '**I**nvestigation' (the project context), '**S**tudy' (a unit of research) and '**A**ssay' (analytical measurement) data model and serializations (tabular, JSON and RDF). The ISA format and framework have been adopted for metadata capturing by the WP7 Plant Use Case, and two new ELIXIR Communities outside EXCELERATE: Metabolomics and Galaxy (ISA is a native Galaxy data type). ISAtools power the submission tool of the ELIXIR DD MetaboLights.<br>ISA's adoption in ELIXIR Use Cases has been aided by a release of several JSON-LD context files in addition to the ISA-TAB format, and by the ISA-API that accelerates reporting, validation and conversions to other metadata formats.<br>ELIXIR Data Validation Implementation Study is undertaking further investigation and prototyping open validation services for archetype archival databases and knowledge bases for the Plant Community (WP7), whereby content validation is according to minimum information checklists, GA4GH file formats, and phenotyping data. |
| Identifier services | The management and serving of mapping between identifiers has been recognised as a gap. The ELIXIR-NL Node proposes BridgDB as a framework for finding and mapping equivalent database |

[40] http://www.pistoiaalliance.org/projects/ontologies-mapping/



Horizon 2020 grant n. 676559

| | |
|---|---|
| | identifiers, chiefly for genes, proteins, and metabolites. Identifier mapping.<br><br>Further work on identifiers mapping will be undertaken in EU IMI FAIRPlus follow-on project, which focuses on the FAIRification of drug discovery datasets. |
| Linked Data Services | The ELIXIR-NL Node has pioneered a Linked Data approaches to Interoperability, focused on supporting WP8 (Rare Disease). These "Data FAIRport" services will be reported more fully below, and include: FAIRifier and Metadata Editor (to create); FAIR Data Point (to publish); FAIR Search Engine (to find); and ORKA (to annotate). |
| Search and Harvesting services | Harvesting services harvest metadata from resources to build indexes an support search. Semantic search services incorporate semantic terms in search.<br>BioSolr facilitates indexing of ontologies, ontology driven faceting, and ontology enriched Solr search. BioSolr is provided by the ELIXIR-EMBL-EBI Node and potentially available to Wp7.<br>Harvesting, indexing and search enrichment services are being primarily developed as part of the Bioschemas activity. |

**Table 7: Metadata services emerging from the Use Cases.**

***Future Work***

The full audit and gap analysis of services offered by ELIXIR nodes is to be completed and related to the Service Framework presented earlier. As highlighted earlier, an Interoperable Interoperability Services Interactions Profile needs to be completed reflecting various interoperability scenarios that have emerged through the EXCELERATE work.

## A2.4 Resource markup – Bioschemas

One of the prime tasks of WP5 is the specification and adoption of dataset descriptors with common data elements, to support dataset lifecycle management and release management, and to assist integration, aggregation and search tools using ELIXIR Resources. For example, provision of Registry, Dataset and sample level metadata are crucial for interoperability, and facilitate the realisation of machine-processable metadata to mark-up resources to support the FAIR Principles (see appendix 5).

These efforts are focused primarily through Bioschemas, an extension of the well-established Schema.org specification employed in commonly used search engines and indexing, as the mechanism we have developed to achieve universal markup of resources is with machine processable metadata. Working with the Bioschemas group, we have initially targeted metadata to improve Findability and Accessibility, since Interoperability and Reusability require richer metadata.

Bioschemas (bioschemas.org) is a simple and powerful way to support resource discovery (Find), resource citation and indexing and inter-resource metadata exchange (Reuse), strengthening collaborations and integration among ELIXIR Data Resources. It

leverages the widely used web-based, commercially supported schema.org, providing a lightweight, long-term sustainable approach to metadata mark-up of the data eco-system that can be both retrospectively and prospectively applied. Bioschemas describes 'types' of information (profiles), each of which are composed of a number of 'properties'. For example, 'Dataset' includes properties such as 'identifier', 'measurementTechnique', and 'variableMeasured'. The main output from Bioschemas is a collection of specifications that provide guidelines to reduce the complexity of including schema.org markup in life sciences resources and to facilitate a more consistent adoption of schema.org markup within the life sciences.

Bioschemas embeds lightweight machine readable metadata in the human facing web pages to enable indexing by web search engines. Resources use the profiles to publish the mark-up in their web-page to then be harvested and indexed by registries, aggregators and search engines (figure 5). Thus web-based metadata can published and acquired without the implementation of costly APIs, retrospectively deployed in legacy data publishing workflows and uses a well-established web standard used by standard search engines and more than 40% of the world's web pages.

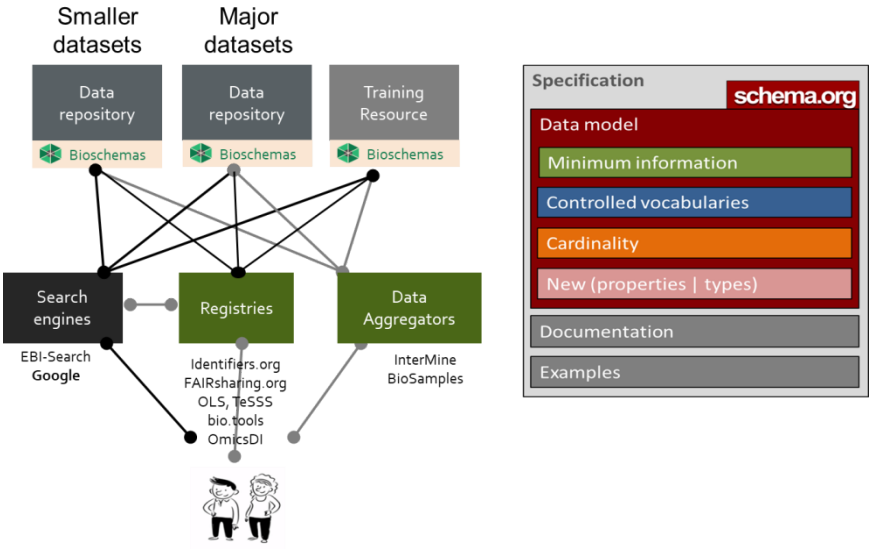


**Figure 6: Bioschemas**: Resources publish the mark-up in their web-page to be harvested and indexed by registries, aggregators and search engines; Bioschemas specifications defines profiles of schema properties and the conventions for using them.

Bioschemas aligns with and promotes ELIXIR recommendations on identifiers, metrics, citation, release, versioning, programmatic interfaces, etc. facilitating the (re)use of data resources. It is a WP5 flagship activity and has been set up as a "semi-independent" community project and has proved an excellent platform for collaboration, bridging with ELIXIR's Data Resources, use cases, platforms and partners internal and external.

Pump-primed by EXCELERATE, Bioschemas has been further supported by an ELIXIR Node Staff Exchange award in 2018, and two ELIXIR Implementation Studies: the first in 2017 to establish draft specifications and pilots and the second in 2018 to build adoption across ELIXIR Core and Node Data Resources and to establish the community.

Bioschemas has captured the imagination of the community beyond ELIXIR and Life Sciences: already a biodiversity specification is being developed and the notion of a "ResearchSchema" has been proposed for the European Open Science Cloud (EOSC).



Horizon 2020 grant n. 676559

Bioschemas is being driven by ELIXIR as an open community project involving and engaging external stakeholders from the outset. This work has had contributions from Pistoia Alliance, GOBLET, TeSS, FAIRSharing, BBMRI, Google and the BD2K BioCADDIE project as well as schema.org.

***Results***

We have established a Bioschemas community of 200+ people and 18 working groups, a steering committee and an independent website (bioschemas.org). 12 ELIXIR Nodes are involved (which is relevant to WP10).

Fifteen Bioschema specifications cover multiple 'types', whilst others in various draft stages, and all presented on a comprehensive website. Specifications fall broadly into three categories:

- The data resource level DataCatalog and Dataset, which cover universal properties of data resources and should apply to all WP3 CDR and DDs and all WP6-9 databases.
- The data type specific: protein: Protein, ProteinAnnotation, ProteinStructure; samples: Sample; human data: Beacon; and (in draft) chemistry: Chemical, Chemical Structure
- Resources other than data resources: Event, Lab Protocol, Training Material, Tool

Three new schema.org types have been found necessary to be defined: BioChemEntity, DataRecord and LabProtocol. These are currently in preparation for formal Schema.org submission.

Thirty live deployments of datasets have been made and 6 Million + pages have been marked up: a live deploys list uses the Google structured data testing tool. Further tools for harvesting and search are in development (http://bioschemas.org/software/).

All the Interoperability Registries have adopted Bioschemas resource markup, by publishing Bioschema markup themselves and by harvesting markup as described in Table 5. The benefits of consistent markup in life sciences resources as provided by the Bioschemas specifications is beginning to be demonstrated.

Additional ELIXIR Interoperability services that engaged with Bioschemas to benefit from markup include: InterMine, MOLGENIS and OmicsDI.

Bioschemas is a cross-WP activity:

- WP3 (Data) Five Core Data Resources and three Deposition Databases are marked-up so far and many more in the pipeline.
- WP1 (Tools) biotoolsSchema (the bio.tools data model) is compatible with the Bioschema which paves the way for mark-up of web pages, improving tool findability. Benchmarking activities through OpenEbench (WP2) are also now gathering software-related data from, for example, Biocontainers/Bioconda and Galaxy initiatives. This information is being collated and exposed through a Bioschemas mechanism.
- WP11 (Training) The TeSS portal pioneered Bioschemas markup for training materials to enable the harvesting of training materials from ELIXIR Nodes without the need for APIs or fragile feeds. TeSS aggregates Event and TrainingMaterial markup from a variety of sources, i.e. the websites of the training providers who are advertising their

events, enabling sharing of data between EeLP (eLearning Platform) and TeSS (in a collaboration between UL (ELIXIR-SI) and the TeSS group, as part of EXCELERATE). The Bioschemas markup is then used to provide a specialised search application, e.g. you can search TeSS for training material about Phylogenetics. Six of the 39 content providers for TeSS are Bioschemas based (8 are HTML scrapers, 7 are JSON feeds, 8 are other). TeSS has registered 497 materials marked up in Bioschemas (nearly half), and 1310 events marked up in Bioschemas (1/6th). Bioschemas is planned to be prime mechanism for at-source metadata markup. WP11 has also developed and utilized Bioschemas for marking up announcements of workshops, conferences, and other academic events for the TeSS portal.

From the Use Cases perspective, all the WPs have participated:

- WP6 (Marine Use Case): Three new and manually curated reference databases (MarRef, MarDB and MarCat) expose additional metadata using Bioschemas markup. We have demonstrated the ability to harvest MarRef data into the ELIXIR BioSamples Deposition Database without the requirement for an API[41]. This has demonstrated an increase in data interoperability in a sustainable way.
- WP7 (Plants Use Case): Recommendation for the adoption of Bioschemas to standardise concept descriptions of datasets.
- WP8 (Rare Disease Use Case): Bioschemas can be used alongside other mechanisms recommended for making specifically rare disease resources findable, such as registration at OrphaNet, registration at the RD-Connect Biobank and Registry finder, and genomics data submission through the RD-Connect platform. These mechanisms, as well as the FAIR data point metadata editor, support the generation of rich metadata to improve Findability and Reuse, and improve data Interoperability through the use of machine readable data. The MOLGENIS scientific data integration system is widely used by the community and has been active in the Bioschemas community, marking up multiple patient registries (DEB-Central, CHD7 Database, Microvillus Inclusion Disease Patient Registry, AIP Mutation Database) with Bioschemas[42].
- WP9 (Human Data Use Case): The Bioschemas beacons specification is proposed to self-describe a beacon's genetic variant cardinality service for better integration with other beacons within the beacon-network. It builds upon the Beacon service API and uses existing schema.org entities and properties. The European Genome-phenome Archive (EGA) live deploy is marked up with Bioschemas.

Bioschemas has had major presence in the ELIXIR and general Life Science community, including:

- 11 Hackathon events[43] developing markup and involving over 250 people
- 8 Dissemination events promoting Bioschemas: (see http://bioschemas.org/publications/)

[41] https://github.com/EBIBioSamples/bioschemas_marref_demo/blob/master/Summary.md

[42] https://f1000research.com/posters/7-1228

[43] http://bioschemas.org/meetings/

- 4 tutorials (SWAT4LS 2018, NETTAB 2018, ECCB 2018, ELIXIR All Hands 2018).
- 5 posters (ELIXIR All Hands, SWAT4LS, ISMB, ISWC)
- Topic focus of the ELIXIR sponsored BioHackathon 2018, supported by an ELIXIR Staff Node Staff Exchange award.

Beyond EXCELERATE, we seek to impact the European Open Science Cloud, notably work on data catalogue metadata exchange. A close collaboration with EOSCpilot has influenced the development of the EOSC Datasets Minimum Metadata Guideline (EDMI)[44] , for describing common and minimum metadata for finding and accessing data. There are a number of groups and international efforts committed to the use of metadata, and in specifying minimal standards for the description of data. In addition, there are a number of metadata properties that are common to different metadata models (for example, DataCite); these properties while conceptually identical seem often to be defined incongruously, leading to potential ambiguity in their assignment. While these properties are earmarked for use across efforts internationally, the precise details around what each property entails (with respect to attributes for the property, and what constitutes valid values for these attributes) is vague, at best.

WP5 participated in a breakout sessions at EOSCpilot All Hands (Pisa, Q1 2018) to identify (Bioschemas) properties that were a priority for Interoperability purposes (facilitating FAIR), and begin to draft properties and attributes, and what would be valid/expected values for those attributes. These properties are being further developed, and will be used in collaborative work with, for example, EOSCpilot and RDA Groups: license and identifier

***Future work***

An ELIXIR Implementation Study is ongoing to promote adoption of Bioschemas and best practices among ELIXIR data resources and Use Cases; develop tools to ease mark-up adoption; facilitate indexing of data resources via ELIXIR registries, aggregators and search engines like Google; and make ELIXIR CDRs compliant with EOSC guidelines to increase visibility in H2020. The key effort is to demonstrate of benefits for the resources and the use of Bioschemas markup by applications, search engines and registries. We are working with the Tools Platform (WP1) to action Bioschema markup of tools in bio.tools, and we aim to have comprehensive uptake of Bioschemas by the EIP services and registries: improved metadata including harvest/display (bio)schema.org for findability of resources for identifiers.org.

Events are being run to encourage further adoption, e.g. tutorials at ECCB 2018, NETTAB 2018, and SWAT4HCLS 2018, as well as a hackathon topic at the ELIXIR-sponsored Biohackathon 2018 to further enhance the benefits of including Bioschemas markup. An ELIXIR Node Staff Exchange award is supporting the adoption of Bioschemas by Node resources, supports attendance at these events.

A high impact publication is expected in the next year, before the end of EXCELERATE. We have held back until the developments described above to be able to really

[44] https://eosc-edmi.github.io/properties



Horizon 2020 grant n. 676559

demonstrate the impact of Bioschemas in practice. This means we need the specifications, tooling and extensive resource mark-up to all be in place.
Work continues on establishing a self-managing Bioschemas community. Over the next year, the community will be formalising its governance structure with the goal of making the community self-sustaining. Additional specifications will be developed to enable markup of additional resource types, e.g. markup to represent chemicals is at an early drafting stage but is needed to markup core data resources such as ChEBI and ChEMBL. Proposals for extensions to schema.org will be submitted.
Finally we will continue our work with RDA and EOSC on metadata properties to: deliver property definitions on license and identifier; propose properties to RDA working group; develop tools for adding / consuming Bioschemas and develop a validator for markup.

## A2.5 Standards

### A2.5.1 Formats, Minimal Information Models and Ontologies

Standards for richly describing the interoperable metadata of datasets are closely tied to the data types and tasks they relate to, and must be undertaken by the communities that the datasets serve. Thus, WP5 does not define the standards for specific data types: these are undertaken by the Use Cases in WP6-9. However, WP5 provides advice and assistance and supplies support services.

***Results***

All standards used by WP3 CDRs and DDs and most standards (but not all) used by the Use Cases are registered in FAIRSharing. All the ontologies used are registered with OLS.
For WP6 (Marine Use Case), a GO Slim (a condensed view of GO terms to give higher-level, broader overview of the ontology) is used for high-level visualisation of the functional profile of metagenomic samples, which lets users run a quick comparison of samples. Using the Slim-O-Matic system, WP5 has worked with WP6 to generate a new, more comprehensive GO-slim for EMG pipelines, which now encapsulates over 97% of all GO-terms currently identified compared with the previous version that only mapped 83% of GO-terms[45]. ~350 terms were added to OLS from http://enseqlopedia.com/enseqlopedia/ for sequencing application information. Referenced-based taxonomy classification criteria have been identified as a challenging, multi-dimensional system.
For WP7 (Plants Use Case), the Minimum Information about Plant Phenotyping Experiment (MIAPPE) standard is a list of attributes for plant phenotyping experiments. The MIAPPE specification has been implemented in the Breeding API (BrAPI) and is currently being extended. With WP5, MIAPPE has been implemented in ISA. WP7 has also developed new ontologies to address gaps in existing vocabularies e.g. Woody Plant Ontology, Plant Phenotype Experiment Ontology (PPEO) which are registered in the OLS and supported by a data import pipeline especially developed for the Use Case.

[45] https://www.ebi.ac.uk/about/news/service-news/metagenomics-go-slim-2016



Horizon 2020 grant n. 676559

For WP8 (Rare Disease Use Case), WP5 has been supporting the development of a data linkage plan to test tools, data models, and protocols to standardize rare disease data services to conform to FAIR data principles 'at source' (with standards, guidelines, software, data stewardship services). Progress is being made with close collaboration with stakeholder partners (RDConnect, BBMRI(-NL), FAIR-dICT, ODEX4ALL, GO-FAIR).

For WP9 (Human Data Use Case), the Data Use Ontology (DUO) describes the uses to which health datasets may be used, and allows search and query based on the use-case for the data. This work is in partnership with the GA4GH Initiative, as is the development of standards for clinical & phenotype data capture: to enable computable and interoperable phenotypes to produce a set of recommended phenotype & disease ontologies, standardized mappings between ontologies, and best practices for their use in genomic medicine, and information models for exchanging clinical data and genomic samples. This includes the development of Phenopackets- a computable phenotype bundle for use inside and outside of clinical settings - and interoperability with the FHIR (Fast Healthcare Interoperability Resources) standard.

***Future Work***

We will continue to support the activities of WP3 and WP6-9. We will also prioritise the development of ontology guidelines and know-how for our planned Knowledge Hub. We will also complete a comprehensive sweep through all the Use Cases to pick up unregistered standards.

### A2.5.2 Identifiers standards and best practices

The FAIR Data principles place significant emphasis on identifiers. The Interoperability platform's Identifiers Working Group seeks to identify resource requirements and to harmonise existing practice in identifier assignment and resolution, to support resources in the implementation of community standards and in the adoption of best practice and identifier services. This work is targeted through use cases and communities, and actively engages international partners. Agreements on identifiers are being formalised in collaboration with the CORBEL project.

***Results***

An identifier recommendations paper[46] published in PLoS (June 2017) provides guidelines on '10 simple rules ' for identifiers best practices and has been viewed over 19K times. It was developed in partnership with a wide international community. Our work with N2T.net on robust support for machine-resolvable, persistent compact identifiers in biomedical data citation was published (May 2018, Nature Scientific Data[47]), and recommended by publisher for Nature Scientific Data (editorial[48]). Based on these papers an identifier harmonisation strategy has been developed with CORBEL, including an "identifiers

[46] https://doi.org/10.1371/journal.pbio.2001414

[47] https://www.nature.com/articles/sdata201829

[48] https://www.nature.com/articles/sdata201895



Horizon 2020 grant n. 676559

hygiene" checklist for datasets. This allows data providers to ascertain whether their resources comply with identifier best practice.

Work with publishers aims to define and develop core persistent identifiers and associated services (with the H2020 project FREYA)

With global initiatives and projects Research Data Alliance , DataCite and EOSCpilot we are establishing identifier metadata property definitions for use in specifications (presented at EOSCpilot All Hands, Pisa 2018) and defining metadata specifications to facilitate findability (*FAIR*) of resources.

***Future Work***

Work ongoing with Use Cases to address identifier requirements and the identifier harmonisation strategy to be tested (this has taken longer than expected to get underway due to staff availability). In collaboration with NIH Data Commons and CORBEL, the Identifier 'hygiene' checklist for repositories has been simplified and will shortly be made available through Knowledge Hub. We aim to target specific communities (e.g. FREYA) to adopt compact identifier formulation particularly targeting publishers (follow on activity from adoption by Nature Scientific Data).

### A2.5.3 APIs and Tools Description Standards

Standards for APIs form a core part of interoperability services. Our API standardisation follows two tracks:

- The use of standardised best practices for the design and documentation of APIs. This is with close collaboration with WP1/WP2 and the Tools Platform, and relates to practices of describing tools and datasets for access from applications and workflows, and for registration in the bio.tools registry. The bulk of this responsibility for this track falls to the ELIXIR Tools Platform (WP1/WP2) and we refer you to their deliverables D1.3, D1.7 and D1.8 for further details.
- The definition and adoption of standardised APIs for data types. Like the standards described above, this work is undertaken by the communities that the datasets serve with support from WP5.

***Results***

With WP1 we have highlighted the use of standardised best practices for the design and documentation of APIs. OpenAPI (https://www.openapis.org/) is recommended for the documentation of APIs. The OpenAPI specifies machine-readable interface files for describing, producing, consuming, and visualizing RESTful web services. It is overseen by the Open API Initiative, an open source collaborative project of the Linux Foundation. Swagger (https://swagger.io/) is recommended as a useful framework for API development which includes open source and professional tooling. Swagger and some other tools can generate code, documentation and test cases given an OpenAPI interface file. OpenAPI was originally part of the Swagger framework donated to the Linux Foundation 2016.

The EDAM ontology provides a controlled vocabulary for the systematic annotation of bioinformatics resources with machine-readable statements, in terms for common topics, specific operations, types of data, data identifiers and data formats. It is used to describe, register and find tools in bio.tools and other resource collections, and a means of describing what software does and what formats it consumes and generates. This supports the FAIR principles and enables better record keeping of the provenance of processed results. In EXCELERATE, EDAM's coverage has been extended to encompass all formats known to the Galaxy platform (EDAM releases 1.18-1.21) and mapped to many external repositories. Galaxy is an important, new User Community for ELIXIR.

EDAM mark-up of tools enables CWL standardised workflow description and validation (see below). See ELIXIR EXCELERATE deliverable reports D1.3, D1.7 and D1.8 for further details.

The definition and adoption of standardised APIs for data types has primarily focused on three Use Cases. For WP7 (Plants Use Case), BrAPI specifies a standard interface for plant phenotype/genotype databases to serve their data to crop breeding applications. A BRAPI2ISA conversion is being developed in an ELIXIR Data Validation Implementation study. For WP8 (Rare Disease Use Case), the REST-based FAIR data point API specifies a metadata standard based on DCAT and Re3Data. It is used to control access and provide basic metadata about a data resource in BYODs and FAIRification projects in the context of the rare disease data linkage plan (see Deliverable 8.2). It can also accommodate Bioschema tags as part of its metadata.
For WP9 the focus has been on GA4GH and Beacon API standards for accessing secure health data.

***Future Work***

WP6 (Marine) reports that non-standardised API development methods are a significant impediment to the development of their EMG pipelines. Moreover, although we recommend these API standardisation practices in reality only a subset of ELIXIR's resources adhere and there is poor understanding of good API practice. An audit of tools and dissemination campaign of best practices for API design, development and documentation through the Knowledge Hub and BYO-API workshops, with WP1 is needed and highlighted in the ELIXIR Tools Strategic Plan for 2019-2023.

### A2.5.4. Workflow Standards

Computational Workflows are a major mechanism of automating data processing and analysis, orchestrating simulations and compute access, and operating data collection and stewardship pipelines. The goal is to create specifications that enable data scientists to describe the sequential analysis that they perform, using tools etc., which constitute a workflow. While the workflow within a use case may be very specific, the formalism of how these processes are described (metadata) and registered (catalogue/registry) will facilitate identification and definition of components, and hence interoperability across use cases.
Workflows are increasingly used to describe and link reusable tools, capture reproducible multi-step processes and represent the step-by-step know-how and provenance of



Horizon 2020 grant n. 676559

processes. Oddly, workflows were not represented at all in the ELIXIR EXCELERATE proposal, despite their long-term and widespread use by the members of the project. Nevertheless, WP5 has undertaken significant work, alongside WP1 (Tools) and WP6 (Marine Use Case) on workflow standards for interoperability, execution and portability.

Communicating and sharing data analyses is a critical problem. A recent Nature article[49] notes that more than 70% of researchers cannot reproduce another scientist's experiments, and startlingly more than 50% cannot repeat their own experiments. A major bottleneck has been the lack of standards mechanisms for communicating, interoperating and reusing analysis methods[50]. An analysis of the metagenomics domain in WP6 also identified major obstacles to interoperability, in large part due to a lack of a methodology to accurately describe computational analysis pipelines (workflows) in use.

A standard for describing analysis methods would provide the foundation for sharing, processing, finding and porting analyses: i.e. FAIR workflows. Such a system would describe analytic processes encapsulates tools (including versions and parameters), reference databases, and computational resource requirements, as well as being flexible enough to evolve with emerging technologies, and possess scalability to handle growth of input data.

***Results***

Common Workflow Language (CWL) (https://commonwl.org) is a relatively new specification for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments. CWL is a community-led[51] standard to describe workflow recipes and analysis tools in a platform-neutral way, making them portable and scalable across multiple computing environments. It covers inputs, outputs, analysis steps, software requirements, resource needs and tool descriptions. EDAM identifiers are used to make these descriptions and link to bio.tool entries (work with WP1)[52]. This enables users to not only find, but also use and connect tools deployed in workflow environments.

A CWL command line is described with parameters and wired together to form a workflow template, which can be executed repeatedly on implementing workflow platforms such as Toil, Arvados, Galaxy, the Seven Bridges Platform or Broad Institute's Cromwell by specifying input files and workflow parameters (as included in a BCO). The command-line tools link to containers like Docker, packaging systems like BioConda and container/tool registries like GA4GH's Dockstore[53] and Biocontainers[54] and identified using PIDs such as RRID. Containers[55] allow lightweight[56], easy and interoperable sharing through capturing and distributing the execution environment of tools and services, with particular

[49] https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

[50] Alterovitz G., Dean D, Goble C., Crusoe M., Soiland-Reyes S., et al . bioXriv. doi: https://doi.org/10.1101/191783

[51] CWL founders and supporters include Seven Bridges Genomics, Broad Institute, Curoverse, Galaxy, Wellcome Trust Sanger Institute, Institut Pasteur, EMBL-EBI, University of Melbourne Center for Cancer Research, Harvard T.H. Chan School of Public Health, IBM Spectrum Computing and Apache Taverna.

[52] https://www.commonwl.org/user_guide/16-file-formats/

[53] PMID:28344774

[54] https://doi.org/10.1093/bioinformatics/btx192

[55] https://doi.org/10.1007/978-3-319-38791-8_58

[56] https://doi.org/10.7717/peerj.1273

focus on cloud deployment (Docker, Singularity, Kubernetes). They are key to portability and reproducibility.

- ELIXIR has joined forces with the EU BioExcel Centre of Excellence to actively sponsor the development and take-up of CWL. The CWL Viewer, developed by BioExcel, is used to visualise and browse workflows described in CWL. By adopting Common Workflow Language ELIXIR has tapped into a growing open and global community of bioinformatics workflow developers and will enable portable pipelines to be created, executed and reused by a multitude of workflow engines and platforms.
- CWL has been pioneered in EXCELERATE with the WP6 Marine Metagenomic Use Case to support containerised pipelines for portability and reuse, further developed with the support of an ELIXIR Implementation Study. Feedback suggests that CWL is relatively quick and easy to adopt if the pipeline is well described and documented. CWL has significantly matured through its association with EXCELERATE. There are an increasing number of workflow execution frameworks and tools for producing CWL tool descriptions.
- Other CWL adopters include NIH BD2K, (including by the three Cancer Genomics Cloud Pilots), Genomics and Health (GA4GH) Cloud execution, the European Open Science Cloud (EOSC) and the EU RI IBISBA for Industrial Biotechnology. CWL is central to planned efforts to develop a Tool and Workflow Collaboratory in the recently accepted EOSC-Life project, coordinated by ELIXIR.
- CWL is proposed as the workflow description component for the Food and Drug Administration's BioCompute Objects[28], which is undergoing IEEE Standardisation (IEEE P2791 BioCompute Working Group[57]). This is a significant development and paves the way for adoption in regulated environments such as commercial vendors and pharmaceutical companies.
- Workflows are a central method for three new ELIXIR Communities: Structural Bioinformatics, Microbial Biotechnology and Galaxy.
- CWL tutorials have been given at the ELIXIR All Hands in 2017 and 2018, and CWL has been central to hackathons and workflows lead WP1. CWL is highlighted in the forthcoming BioHackathon that is being organised by EIP. EXCELERATE's work with CWL is a flagship project of the CWL community and has been featured in many presentations and outreach venues including BOSC 2016, BOSC2017, BOSC2018, Galaxy Community Conference 2018, BioCompute Objects PoC Workshop 2018, HTS COMPUTATIONAL STANDARDS FOR REGULATORY SCIENCES WORKSHOP 2017, National Cancer Institute (NCI) Center for Biomedical Informatics and Information Technology (CBIIT) Speaker Series 2017, presentations to Elsevier and Seven Bridges, to at the Melbourne Bioinformatics (May, 2017), and NITRD

[57] IEEE P2791 BioCompute Working Group (BCOWG): Standard for Bioinformatics Computations and Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication, http://sites.ieee.org/sagroups-2791/

MAGIC team meeting[58] (2018) and webinars for ELIXIR, openPHACTS and Bioexcel[59] .

***Future Work***

Work on WP6 has highlighted extensions necessary to CWL notably specifications to enable workflow branching to handle different input types, for example. CWL 1.1 is in draft[60] .

CWL describes prospective workflows to be executed in multiple ways, and so does not capture other critical aspects of an already executed analysis, such as the actual input and output data, or necessary metadata such as the domains in which the workflow is valid. Sharing workflow specifications alone is not sufficient to guarantee successful reuse. Insufficient documentation, missing example data, unreliable third-party resources and execution environment setups are cited as obstacles by WP6 and others. With BioExcel and the BioCompute Object / IEEE P2791 community, we are combining CWL for workflow specification and the Research Object RO (http://www.researchobject.org) specification for rich metadata with modern container environments for capturing and distributing the execution environment of tools and services[61]. The RO approach originated in the desire for workflow reuse, reproducibility and preservation by applying standards-based formats for packaging and describing the complete digital contents of an experiment. The CWL workflow (the interoperable method) and RO metadata (the reproducible container) is a machine-actionable details view of BCOs, enabling them, to be portable and repurposable across systems. CWLProv standardises the collection of runtime provenance. All this work is highly relevant to ELIXIR's workflow efforts.

CWL describes workflow recipes and analysis tools in a platform-neutral way that must them be executed by CWL-compliant workflow systems. Work with Galaxy continues, as Galaxy is a widely used analysis platform and an ELIXIR Community. Other workflow systems well-used in the Life Science community such as Knime for biomolecular simulation are developing CWL support in partnership with the CWL community including ELIXIR.

Improved metadata to support workflow description includes the incorporation of EDAM (or other relevant) terms to better standardise 'types' in identifiers.org (replace or supplement current 'tag' mechanism).

Finally, work with WP4 and the Compute Platform will address and align with GA4GH standards for Workflow Execution (WES) and Tool Execution (TES).

## A2.6 Approaches

### A2.6.1 Workflow

[58] https://www.nitrd.gov/nitrdgroups/images/a/ao/Common_Workflow_Language.pdf

[59] https://www.slideshare.net/BioExcel/bioexcel-webinar-series-introduction-to-the-common-workflow-language-cwl-project

[60] http://www.commonwl.org/v1.1.0-dev1/

[61] https://doi.org/10.1016/j.future.2011.08.004



Horizon 2020 grant n. 676559

As highlighted above, computational workflows are a major mechanism of automating data processing and analysis, orchestrating simulations and compute access, and operating data collection and stewardship pipelines. Galaxy, arguably the most popular workflow analysis platform in Bioinformatics, has been formally recognised as an ELIXIR Community and is extensively used across the ELIXIR Node members and by other, new, ELIXIR Communities such as Metabolomics and Proteomics. Several other workflow systems are also used by the ELIXIR community: for example, Nextflow, AWE and Toil by Marine Metagenomics and Knime by Structural Bioinformatics and Microbial Biotechnology.

These systems need support of WP5 for: standardisation (through the Common Workflow Language), portability through containerisation for execution over cloud and clusters, and accessibility through shared metadata for search and registration in public catalogues. Finally, workflow design, deployment, optimisation, maintenance and preservation is non-trivial and requires technical support and best practice training and guidelines.

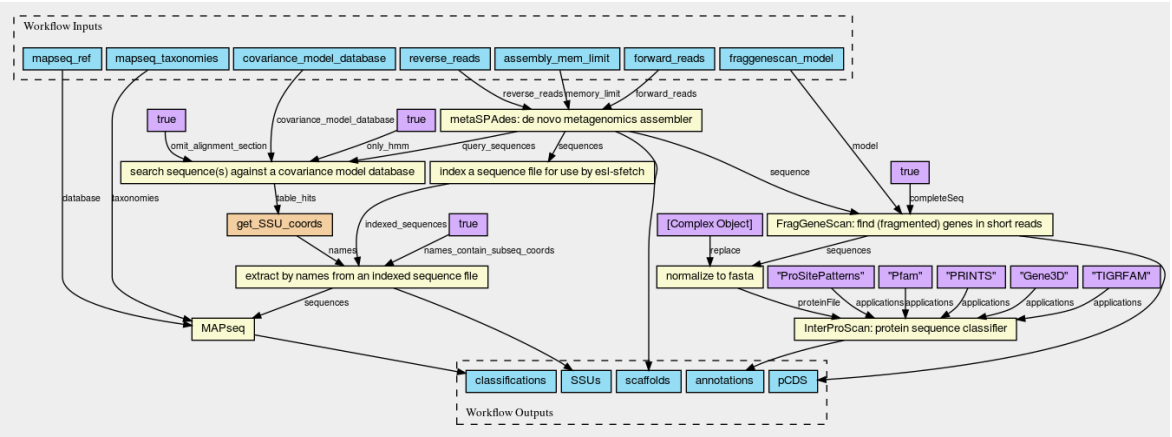


**Figure 7: CWL workflow describing EMG assembly for paired end Illumina rendered using the CWL Viewer[62].**

***Results***

WP5 has worked alongside WP6 to standardise workflow descriptions, and WP1 and WP4 to build fully containerised workflows. An ELIXIR Implementation study has further developed this work[63]. Workflows have been developed and are publicly available for the EBI metagenomics pipeline, ITSone and MetaShot pipelines and shortly for the META-pipe pipeline. These pipelines are reported in Deliverable D6.3. CWL has significantly matured as a direct result of this work.

The adoption of workflows has in turn impacted data archives and their annotation. The archiving of results for identification data analysis of biological entities entails annotation of sequence data with metadata (e.g. GO terms, gene identifiers/symbols, etc.). ENA has been recently extended to accommodate such information

[62] https://w3id.org/cwl/view/git/ca6ca613fod3728d9589a6ca6293e66dfde87bfb/workflows/emg-assembly.cwl

[63] https://view.commonwl.org/workflows/github.com/mscheremetjew/workflow-is-cwl/blob/master/workflows/TranscriptsAnnotation-wf.cwl gives an example of a workflow



Horizon 2020 grant n. 676559

In WP6 a number of workflow execution environments are used cross the partners. The increasing number of workflow execution frameworks and tools for producing CWL tool descriptions allows the capture of sample metadata, provenance information, and analysis workflows as a Research Object with a single identifier and significantly decreases the code base/maintenance, and is starting to be adopted externally to the project; for example MG-RAST have transitioned from their proprietary workflow language to CWL. Currently, different pipelines are at different stages with respect to CWL adoption.

***Future Work***

The discovery, quality control and stewardship of workflows has been raised as an issue by WP6 and others. Work is underway with WP1 (Tools Platform) to plan and develop a CWL workflow registry based on prior developments, notably Biocontainers developed as part of WP1, CWL Viewer developed by the EU BioExcel project, and myExperiment which is hosted and developed by members of ELIXIR-UK and used by EU RI IBISBA1.0. This work will be co-developed with the ELIXIR Galaxy User Community. Future development will be principally undertaken as part of H2020 EOSCLife project. Some work has already been undertaken to link CWL Viewer with bio.tools using EDAM ontology markup.

Having demonstrated the value of workflows and their standardisation, we aim to develop greater adoption in other Use Cases and to have workflow practices and support embedded in ELIXIR across all work packages. For example, in the RD-Connect bioinformatics pipeline[64] of WP8, data from sequencing experiments is submitted by participating research projects, processed with a standard pipeline and made available for online analysis through a user-friendly interface to authorised users. Some of this pipeline may be appropriate for CWL encoding.
The reporting of best practice with examples through the EIP Knowledge Hub is a priority.

### A2.6.2. Linked Data

The FAIR principles do not prescribe any specific implementation or standard, which occasionally has led to disagreement about what it means to be FAIR and what data resources are justified to carry the label "FAIR". However, any have interpreted the publishing and consuming of machine-processable metadata to be synonymous with Linked Data – that is publishing structured data so that it can be interlinked and become more useful through semantic queries that builds upon standard Web technologies such as HTTP, RDF, SPARQL and URIs.

WP8 Rare Diseases has pioneered Linked Data Services to answer cross-resource questions such as bridging genomic and phenotypic data for variant identification using machine processable (RDF/XML) representation of the metadata. RD-Connect Linked Data and Ontology Task Force (LDOTF) advocates the use of Linked Data and Ontologies for managing and integrating rare disease information and make biomedical data easier to use for computational research for the benefit of patients.

[64] https://platform.rd-connect.eu, based on https://www.ncbi.nlm.nih.gov/pubmed/27604516

Linked Data approaches requires: rich, common ontology-based metadata and persistent identifiers to describe the Data and the relationships (Links) between them; tools to undertake the annotating, publishing and processing; and considerable data stewardship know-how. The process has been dubbed “FAIRification”.

***Results***

Work has focused on building Node capacity in skills and knowledge for data interoperability, and for access to technical infrastructure that supports Linked Data interoperability working, using the resources of RD-Connect and the Data FAIRPort toolkit provided by ELIXIR NL and GO-FAIR. The work is summarised in figure 8.

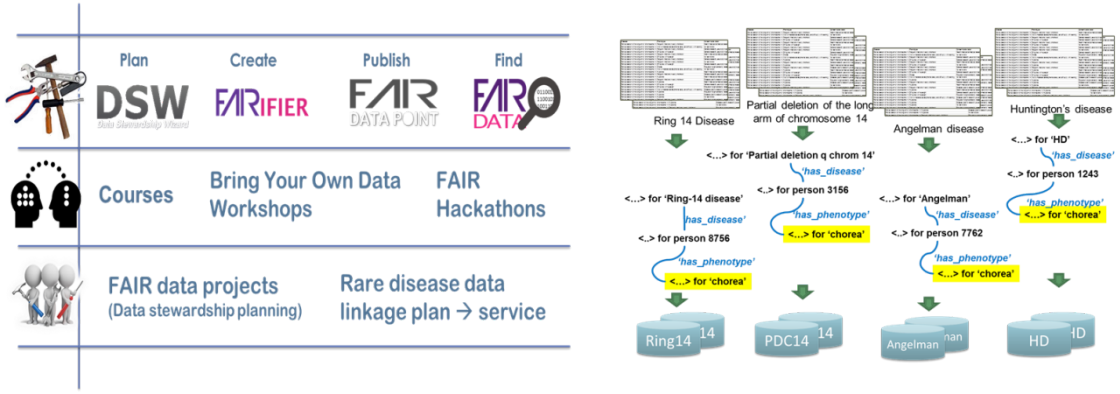


**Figure 8: The tools, capacity building and stewardship activities (left) to build Linked Data SPARQL endpoints for Rare Disease data sets (right).**

The Dutch Techcentre for Life Sciences (part of ELIXIR-NL) has been working towards a suite of tools and standards - Data FAIRPort[65] - that serves the dual purpose of being a reference platform (demonstrating what a FAIR data infrastructure could look like) and a set of reusable components which may be adopted (whole or in part) by any infrastructure that wants to adopt FAIR (or certain aspects of FAIR). These tools have been extensively used by WP8:

- FAIRifier and Metadata Editor (to create). The FAIRifier is an extension of the Open Refine software that supports making data more FAIR using Linked Data.
- FAIR Data Point (to publish). These give a standard metadata interface for a data repository using existing metadata standards such as DCAT and HCLS, and REST APIs. An implementation of the FAIR Data Point was written as a standalone example application. In FAIR hackathons other repository software such as MOLGENIS, tranSMART and Mendeley has been extended with the FDP interface.
- FAIR Search Engine (to find). The FAIR Data Search Engine harvests the metadata available on FAIR Data Points or compatible data repositories, indexes them, and provides a search interface.
- ORKA (to annotate). The Open, Reusable Knowledge graph Annotator (ORKA) supports human curation of knowledge graphs by offering graph annotation as a

[65] https://www.dtls.nl/fair-data/find-fair-data-tools/



Horizon 2020 grant n. 676559

service and capturing the provenance of the annotator and the original statement. The ORKA prototype has been developed in the context of the ODEX4all project.

- For the Rare Disease community, a user application was developed that presents a friendlier interface for non-technical end-users and brings together several of the FDPs created for FAIRified Rare Disease data and can support complex federated queries over the distributed data.

FAIR procedures and expertise are based on experiences from FAIRifying various rare disease resources, such as registries for Duchenne muscular dystroph, vascular anomalies, and osteogenesis imperfecta for FAIRification projects under the 'rare disease data linkage plan' (all interdisciplinary collaborations between FAIR experts and managers of a rare disease data source). The FAIRification process[66] has been defined to extend and generalise the bespoke BYOD service into a best practice guideline for adoption of a BYOD process by a wider community. It is currently piloted with local and 'roving' data stewards, to make rare disease resources interoperable (FAIR) 'at the source' and to speed up the FAIRification process with better tooling and procedures. This is reported in Deliverable 8.2.

A 'data steward' is a person who contributes to increasing the long-term reusability of data, here by making data comply with the FAIR guiding principles. A prerequisite for knowledge transfer is that a data steward working on the side of the domain expert is committed to each case. Eight Bring Your Own Data (BYOD) meetings have been held that have brought linked data experts together with data owners and subject matter experts. BYODs are reported in Deliverable 5.3.

WP5 resources (Identifiers.org, OLS, and Bioschemas) have been used to accommodate identifier resolution, concept mapping, and concept linking. Identifiers.org records all known resource specific access URLs for a given scientific record, as well as document alternative URI schemes through which datasets can be accessed. This information is crucial for integration activities, as it allows numerous, but equivalent, URIs and URLs to be considered as a single (Identifiers.org as the canonical URI) lexical string. This 'mapping' feature is integral to the EMBL-EBI RDF platform, where Identifiers.org URIs act as a semantic glue in distributed queries. For new data elements, managers of a data source may consider generating their own identifiers. Generating new identifiers is supported by the FAIRifier tool. These are formed as URIs, but may or may not be proper globally unique and resolvable identifiers that can be published on the web. This could be a temporary step, before mapping these provisional (or internal) identifiers to authoritative identifiers, or a manager decides to become the authoritative source (e.g. OrphaCodes from OrphaNet).

***Future Work***

[66] Carta, C., Roos, M., Jacobsen, A., Thompson, M., Wilkinson, M.D., Cornet, R., Waagmeester, A., Van Enckevort, D., Jansen, M., Licata, L. and Via, A. (2017) January. The FAIRification of data and the potential of FAIR resources demonstrated in practice at the Rome Bring Your Own Data workshop. In CEUR Workshop Proceedings (Vol. 2042).

The Data FAIRPort tools are still in Alpha and are hard to use outside their originating group. Work is needed to bring them to TRL6 and beyond.

A review of identifier use cases has provided mapping use cases (the need for mapping services) and identifier generation use cases around ontologies. Improved rdf-typing of identifiers.org is needed to facilitate Linked Data use cases and applications, and the incorporation of EDAM (or other relevant) terms to better standardise 'types' in identifiers.org (replace or supplement current 'tag' mechanism).

FAIRsharing.org has potential for supporting the modelling step of the FAIRification process pioneered by WP8. Data stewards first search and reuse application models (application ontologies) that were used in previous FAIRifications or are commonly used in FAIR data repositories like the ELIXIR recommended repositories. FAIRsharing.org offers features to store and search such application models but can support this step more specifically. The FAIRifier currently has rudimentary support for sharing its transcribed data models in FAIRsharing.org. As the data model does not contain the data itself, it may in most cases be shared freely, thus supporting model reuse, which in turn reduces modeling effort and increases interoperability for other datasets of a similar type. It is future work to extend application model sharing facilities for FAIR data stewards.

A FAIRification BYOD process needs to be generalised and un-embedded from WP8 to become transferrable and robust. The recently accepted EU IMI FAIRPlus project, led by ELIXIR, aims to build on the work of EXCELERATE to further build and deploy a transferrable FAIRification methodology.

## A2.7. Capacity Building (Task 5.3)

### A2.7.1 BYODs and Hackathons

Bring Your Own Data (BYODs), hackathons, summer schools, conventional workshops and tutorials are the mainstay of our face to face training. BYODs will be reported in D5.3.

***Results***

| | |
|---|---|
| BYODs | BYODs in WP5 have focused on the WP8 Rare Disease Use Case and the FAIRification of resources using the ELIXIR-NL Data FAIRPoint services and Linked Data. The process is described further in Deliverable D8.2.<br>● 8 Bring Your Own Data (BYOD) meetings have been held for EXCELERATE that have brought linked data experts together with data owners and subject matter experts. These will be reported on in another deliverable. In total, 19 BYOD meetings have been held.<br>● 50+ different people that have visited a BYOD meeting and thereby applied different aspects of Linked FAIR Data for the first time is.<br>● 37 people functioned as experts in BYOD meetings.<br>● 5 industry partners of ELIXIR-NL (3 small and 2 large companies) have been involved in BYOD events, as experts, as well as a tool provider. Industry partners include: DSM and Elsevier |



Horizon 2020 grant n. 676559

| | |
|---|---|
| | • 4 implementations of the Fair Data Point protocol in different repositories (commercial as well as open source) have been realized.<br><br>BYODs for other Use Cases<br>• BrAPI Gent 30 May- 1 June 2017 (WP7)[67]<br>• BYODs at Bio-IT World 2017 and 2018. |
| Tutorials | CWL<br>• Galaxy Community Conference[68]<br>• ELIXIR All Hands 2017 and 2018<br><br>Bioschemas<br>• ELIXIR All Hands 2017 and 2018<br>• SWAT4LS 2018 (future)<br>• ECCB 2018 (future)<br>• NETTAB 2018 (future) |
| Summer Schools | In the WP8 rare disease community, a BYOD organised as part of the popular Rome Summer school for rare disease registry managers that is organised by the Istituto Superiore di Sanita (ISS). ISS and DTL/ELIXIR-NL co-organise the BYOD. The summer school+BYOD has become a household name and will continue to be organised annually for at least the next 5 years. |
| Hackathons and Workshops | • Bioschemas has held 11 meetings since March 2017<br>• BioHackathon 2018 (www.bh2018paris.info) coordinated by the ELIXIR Interoperability Platform. Over 140 people are expected, and 30 places are dedicated to Bioschemas.<br>• March 2018, Pisa, Italy (EOSCpilot All Hands); approximately 30 attendees across 4 breakout sessions on metadata<br>• In FAIR hackathons repository software such as MOLGENIS, transMART and Mendeley has been extended with the FAIR DataPoint interface |
| Staff exchange | An ELIXIR Staff Node Exchange award in 2018 was granted to support Bioschemas adoption, to support attendance at tutorials and attendance at hackathons, notably the ELIXIR sponsored BioHackathon in Paris, November 12th - 16th 2018 (above). |

**Table 8: Capacity Building activities. Further reporting in D5.3.**

***Future Work***

BYODs were originally embedded within WP8 Rare Disease use case evolved from "one-off" heroic efforts into a generalisable methodology for processing FAIR data. A transferable permanent methodology evolution is needed, as opposed to "wizarding" a resource as a one off solution.This methodology could be employed to develop new

[67] https://drive.google.com/file/d/0B5leXRSs64hUck9QY3NQWF91bnM/view

[68] https://gccbosc2018.sched.com/event/Dn9R/introduction-to-common-workflow-language



Horizon 2020 grant n. 676559

software, guide best practice adoption, as well as address use cases beyond the original remit of Rare Disease.

A FAIR-CDR ELIXIR Implementation Study has been funded to put proposed FAIR Metrics into practice for the ELIXIR community by starting to "FAIRify" ELIXIR Core Data Resources, and the recently accepted IMI FAIRPlus, led by ELIXIR, aims to develop FAIR Metrics for datasets of value to the pharmaceutical commercial sector and to develop a transferrable "FAIRification" methodology. The bulk of the work will be through BYODs. Lessons learned in EXCELERATE and the experiences of the process developed in W8 and reported in D8.2 will be the foundation.
The emerging importance of API and Workflow knowledge has created a need for BYO-API and BYO-Workflow workshops. These and the extended BYODs need a closer relationship with the WP11 Training metrics and practices. Workflow design, deployment, optimisation, maintenance and preservation is non-trivial, needing best practice training and guidelines.

### A2.7.2 Knowledge Hub

The Knowledge Hub is intended a "one stop shop" website for guidelines, pointers to guidelines, templates, best practice papers etc for EIP's interoperability work, complementing training materials and events listed in the TeSS Portal.

In February 2018, the European Research Council (ERC) published the Open Research Data and Data Management Plans recommending all ERC grantees to consult with the ELIXIR Interoperability Platform (EIP) Knowledge Hub and website for the references on metadata resources. With all incoming ERC awardees referencing EIP metadata resources, we will impact life sciences research in major proposal calls such as H2020. Within the scope of H2020 alone, ERC is allocated with the budget of € 13.1 billion for 2014-2020 period. This is projected for 7,000 grantee funding 42,000 H2020 team members[69].

Moreover, all the Use Case WPs report on the need for best practices, guidelines, templates, and pointers to materials, services and communities and resources where further knowledge can be found. A Knowledge Hub is the place for our identifier hygiene checklists and guides to developing standards.

***Results***

Thus far little work has been undertaken on the Knowledge Hub, largely down to prioritising work on other activities and changes in coordination support at the Hub, and some false starts. A minimal web presence has been assembled thus far.

***Future Work***

An online ELIXIR Knowledge Hub is planned to be the official ELIXIR Interoperability reference of the EIP Framework, and the centralised resource for the reference of the

[69] https://erc.europa.eu/projects-figures/facts-and-figures

ELIXIR mission-critical services identified by the EIP service selection process. It will act as a place where infrastructure and tool developers can be informed about current ELIXIR Interoperability Platform activities, assess current best practice and determine which interoperability services and standards to implement under specific circumstances. This is to be developed in collaboration with the ELIXIR Hub as a priority.

Materials with which to populate the KH are currently being collated and developed. We anticipate best practice and checklist documents (identifiers checklist being developed in collaboration with NIH Data Commons KC2) to begin appearing in Q4 2018. The extensible solutions driven by use cases will be disseminated on the EIP knowledge hub and EIP will continue to publish the evolving services and methodologies in collaboration with the other ELIXIR technical platforms and our strategic partners such as GA4GH, Human Cell Atlas, and RDA.

The TeSS Training Portal is the flagship training material and events portal for ELIXIR.WP5 BYODs, Hackathons and so on are promoted through the TeSS. WP5 will work with WP11 to produce training materials and complement the Knowledge Hub. Work has already begun to produce Bioschemas training materials.

# Appendix 3: EIP Framework Mapped to WP Tasks

| WP 5 Task | WP5 Task Name | Sectors | Sub-sector |
|---|---|---|---|
| 5.1.1 | Use Cases. Outcome: Interoperability, common APIs and descriptors workable in the field. | Services<br>Metadata<br>Standards | Service Framework<br>Registries, Metadata Services<br>Resource Markup<br>Standards |
| 5.1.2 | Core and Named Resources. Outcome: Common APIs and dataset descriptors workable in the field. | Services<br>Metadata<br>Standards | Service Framework<br>Registries, Metadata Services<br>Resource Markup<br>Standards |
| 5.1.3 | Global engagement | Global Initiatives | Every Sector has global engagement.<br>See Appendix 4 |
| 5.2.1 | Identity Management, Mapping and Tracking services | Metadata<br>Standards | Registries, Metadata Services<br>Identifier standards |
| 5.2.2 | Reporting Guidelines, Formats, Controlled Vocabulary Services | Metadata<br>Standards | Resource markup<br>Standards |
| 5.2.3 | Dataset publishing for API interoperability | Metadata<br>Standards | Resource Markup (Bioschemas) |
| 5.2.4 | Biological knowledge base publishing for Linked Data interoperability | Approaches | Linked Data |
| 5.2.5 | Sustainability of Interoperability Implementation Services. | Services | Service Framework |
| 5.3.1 | Manage and Run BYOD Workshops | Capacity | Knowledge Hub, BYODs |



Horizon 2020 grant n. 676559

| | | | |
|---|---|---|---|
| 5.3.2 | Create and manage BYOD training materials | Capacity | Knowledge Hub, BYODs |
| 5.3.3 | Data Node Capacity Building | Capacity | Knowledge Hub, BYODs |

**Table 9: Mapping of the original tasks (WP task number) to the Framework.**



Horizon 2020 grant n. 676559

# Appendix 4: Global Engagement Matrix

| | | FAIR Services & Metrics | | | Metadata | | | Standards | | | | Approaches | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Services | Metrics | MaturityModels | Registries | Services | ResourceMarkup | Formats, MLM, Ontologies | Identifiers | API | Workflow | Workflows | LinkedData |
| **GlobalStandards&Commun** | GO-FAIR / FAIRMetrics.org | | X | | | | | | | | | | |
| | Research Data Alliance | X | X | | X | | | | X | | | | |
| | Schema.org | | | | | X | X | | | | | | |
| | Force11 | | | | | | | | X | | | | |
| | Common Workflow Language | | | | | | | | | X | X | X | |
| | OpenAPI / SWAGGER | | | | | | | | | X | | | |
| | GA4GH | | | | | X | | | | | X | X | |
| | NIH Data Commons | | | | | X | | | | | | | |



Horizon 2020 grant n. 676559

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **initiatives** | | | | | | | | | | | | | |
| **EU projects, Infrastructures** | EOSCPilot | | | | X | | X | | | | | | |
| | BioExcel CoE | | | | | | | | | | X | X | |
| | RD-Connect | | | | | | | X | | | | | X |
| | CORBEL | | | | | | | | X | | | | |
| | FREYA | | | | | | | | X | | | | |
| | CHARME<br>COST Action | | | | | | | X | | | | | |
| | BBMRI-ERIC | X | | | | | | | | | | | |
| | OpenAIRE | | | | | | X | X | | | | | |
| **Inter** | NIH Data Commons<br>https://commonfund.nih.gov/commons/awardees | | | | X | | | | | | | | |
| | Human Cell Atlas | | | | | | | X | | | | | |



Horizon 2020 grant n. 676559

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **National Projects & Infrastructures** | Monarch | | | | | | | | X | | | | X |
| | Galaxy | | | | | | | | | | X | X | |
| | FDA BioCompute Object | | | | | | | | | | X | X | |
| **National Centres &** | DANS (NL) | | X | | | | | | | | | | |
| | SSI (UK) | | X | | | | | | | | | | |
| | CDL (USA) | X | | | | | | | X | | | | |
| | NIH BD2K CEDAR Centre (USA) | X | | | | | | | | | | | |
| | NIH BD2K BioCADDIE Centre (USA) | | | | X | | X | X | | | | | |



Horizon 2020 grant n. 676559

| Initiatives | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Commercial | Google | | | | X | | X | | | | | | |
| | DataCite | | | | | | | | X | | | | |
| | Springer | | | | | | | | X | | | | |
| | Pistoia Alliance | X | | | | | X | X | | | | | |

**Table 10: Framework mapped to global initiatives and standards**



Horizon 2020 grant n. 676559

# Appendix 5: The FAIR Data Principles

**To be Findable:**
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

**To be Accessible:**
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

**To be Interoperable:**
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

**To be Reusable:**
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

**Figure 2: The FAIR guiding principles**, reproduced from Wilkinson *et al.*, box 2[1]

**Figure 9: FAIR :Principles**



Horizon 2020 grant n. 676559

# Appendix 6: WP5 Dissemination Events

**Oral presentation**s (partial - for full details, please see this link)
WP6-WP9 will also report their presentations in their deliverables

| | | |
|---|---|---|
| BD2K All Hands 2015 Nov 15-13 2015<br>NIH Main Campus, Bethesda, MD, USA | General | 200 |
| Science Europe: Career Pathways in Multidisciplinary Research: How to Assess the Contributions of Single Authors in Large Teams 1 Dec 2015 Brussels | General | 35 |
| PHOBIOS2 Workshop 2016, 17 -19 Feb 16, Arizona, USA<br>https://github.com/identifier-services/phoibos2/wiki<br>https://github.com/identifier-services/phoibos2/wiki | Identifiers | 50 |
| UK Ontologies network (UKON) 14 Apr 16, Newcastle, UK | Metadata, Standards | 50 |
| DCIP Identifiers Workshop 2 Jun 16, Harvard University, Cambridge MA, USA | Identifiers | 9 |
| NSF Workshop on Data and Software Citation 2016<br>6 -7 Jun 16 Boston, MA | Bioschemas, Identifiers | 20 |
| Agrohackathon , 29 Jun 16 Montpellier, France<br>https://www.meetup.com/AgroHackathon/ | Ontologies, Metadata, Standards | 50 |
| The BioSharing Registry: mapping the landscape of standards and database resources in the life sciences, ECCB presentation (ELIXIR Track) 6 Sep 16, Hague, Netherlands, http://www.eccb2016.org, | Registries | 100-200 |
| Bioschemas presentation , ECCB presentation (ELIXIR Track) 6 Sep 16, Hague, Netherlands, http://www.eccb2016.org | Bioschemas | 100 |
| Bioschemas lightning presentation, 15th Intl Semantic Web Conference, Kobe, Japan, 2016, 17 -21 Oct 16 Kobe, Japan | Bioschemas | 50 |
| Combined CHARME – EMBnet and NETTAB 25-28 Oct 16, Rome, Italy | Bioschemas | 100 |
| Leiden van Leeuwenhoek Lecture on BioScience, 24 Nov 16, Leiden, NL | Bioschemas | 45 |
| BD2K, 29 Nov -1 Dec 16 Washington, USA<br>https://datascience.nih.gov/bd2k | General | 100-200 |
| Registries of domain-relevant semantic reference models help bootstrap interoperability in domains with fragmented data resources<br>6 -7 Dec 16 Amsterdam<br>http://www.swat4ls.org/wp-content/uploads/2016/12/paper-16-3.pdf | WP5, WP8 | 15 |



Horizon 2020 grant n. 676559

| | | |
|---|---|---|
| 2017 HTS COMPUTATIONAL STANDARDS FOR REGULATORY SCIENCES WORKSHOP, FDA BioCompute Objects, 16-17 Mar 17, NIH Main Campus, Bethesda, MD | Workflows, Standards | 100 |
| ELIXIR all-hands 2017, 21-22m March 2017 | Talks on : Identifers, Bioschemas, Biosharing | 70 |
| ELIXIR Interoperability Platform: Integration of data and services Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017, ELIXIR Interoperability Platform: Integration of data and services, 21-25 July, 2017, Prague, CZ | Many talks: General, Biosharing, Bioschemas, Rare Disease FAIRification , Workflows | 100 |
| Workshop at Graphs in Life Sciences conference, sponsored by Neo4j, Berlin., June 21, 2017 | Ontologies, Tools | 23 |
| RDA 10th Plenary Montreal, 4 sessions, | FAIRsharing registry | 100 |
| CODATA Data Standards Workshop, 1 Nov 2017, London , UK | FAIRsharing registry | 40 |
| SWAT4HCLS 2017, 5-8 Dec 2017<br>Keynote (Goble), posters | Bioschemas | 50 |
| Workshop on Managing Digital Research Objects in an Expanding Science Ecosystem, 15 Nov 2017, Bethesda, USA.<br>Invited talk (Goble) | Bioschemas. General, CWL | 120 |
| UK Biobanking Showcase 2017, London, 18 October 2017 (invited talk, Goble). | Bioschemas, general | 100 |
| Open Science Fair 2017, Athens Workshop on Data Interoperability, 6-8 Sept 2017, Athens | Bioschemas, FAIRsharing | 45 |
| 16th Intl Conf Semantic Web 2017, poster presentation, 21-25 October 2017, Vienna, Austria | Bioschemas | 200 |
| ELIXIR SME Meeting, Cambridge, Jan 2018 | FAIRSharing , Bioschemas, Intermine | 45 |
| EOSCPilot Data interoperability demonstrators Workshop, 8-9 March, 2018. Pisa, Italy.<br>https://tinyurl.com/EOSCdataIntDemosPisa | Bioschemas, FAIRSharing , Identifiers.org | 35 |
| The 18th Annual Bioinformatics Open Source Conference (BOSC 2017) was held July 22-23, 2017 in Prague as part of the ISMB/ECCB meeting. | Bioschemas, CWL | 200 |



Horizon 2020 grant n. 676559

| | | |
|---|---|---|
| 14th Intl Symposium on Integrative Bioinformatics, IB2018 Rothamsted Research, Harpenden, UK, 13-15 June 2018 (featured in keynote ( Goble); posters | Bioschemas | 50 |
| The 2018 Galaxy Community Conference (GCC2018) and Bioinformatics Open Source Conference 2018 (BOSC2018), Portland, Oregon, United States, June 25-30, 2018. | CWL | 200 |
| 6th UK Ontology Network (UKON) workshop, Keele, UK, April 2018 | Bioschemas | 50 |
| BioCompute Objects PoC Workshop 2018, March 23, 2018, Washington DC | Workflows, CWL | 50 |
| International Conference on Biological Ontology 2018 August 7-10, 2018, Corvallis, Oregon, USA, 8 August 2018 | EIP services for scientific use cases | 150 |

**Organisation of workshops**

| | | |
|---|---|---|
| Workshop at RDA plenary meeting<br>RDA plenary meeting, 1PW -3 Mar 2016, Tokyo, Japan<br>https://rd-alliance.org/plenaries/rda-seventh-plenary-meeting-tokyo-japan | General | 50 |
| Metadata Searching Satellite Meeting – ELIXIR All Hands Meeting, 9-10 Mar 2016,<br>CCIB - Centre Convencions Internacional de Barcelona, Barcelona, Spain | Metadata, Bioschemas | 50 |
| Samples Club BYOD (Bring Your Own Data) workshop, 21 -23 Mar 2016, Manchester, UK<br>https://docs.google.com/document/d/1QfPjtbnbFzP8wed3f1KmadPce0h4E45c3im-cfF9lE8/edit#heading=h.n7aj89bm55x4<br>(with BBMRI-ERIC, FAIRDOM, RD-CONNECT, SynBioChem Centres and reps from NL,SE,FR,UK,EMBL Nodes ) | Metadata, Bioschemas | 15 |
| ELIXIR Curation & Usability Hackathon: Registration of Tools & Data Services, 24- 25 Mar 2016, Gif-sur-Yvette, France<br>http://tinyurl.com/registryhackathon6 | With WP1 | 20-50 |
| Data Nodes network brainstorming workshop (10.2) 13-14 Apr 2016,<br>https://drive.google.com/open?id=0B6g9Zb2twAT_Z3NETjRZVGtEek0 | With WP10, WP4, WP9 | 20 |
| FORCE2016 Searchathon, 17 Apr 2016,<br>OHSU Collaborative Life Sciences Building, Portland, Oregon, USA | Bioschemas | 30 |
| Consensus workshop PhenoHarmoIS 2016, 9-13 May 16, Montpellier, France,, http://tnyurl.com/hzsho6v | WP7 and links with WP5 | 50-100 |



Horizon 2020 grant n. 676559

| | | |
|---|---|---|
| Bioinformatics Technical Hackathon: Tools, Workflows and Workbenches, 18-20 May 2016, Paris, France, tinyurl.com/registryhackathon8 | WP1 & WP5 | 15 |
| Workshop at RDA plenary meeting, 15-17 Sep 16, Denver, USA<br><br>https://www.rd-alliance.org/rda-8th-plenary-joint-meeting-ig-elixir-bridging-force-wg%C2%A0biosharing-registrywg-data-type | General | 20 |
| Bring Your Own Data (BYOD) workshop: WikiPathways, nanopubs, and the Rett Syndrome 1-3 Nov 2016<br>various<br>http://www.dtls.nl/fair-data/byod/bring-data-workshop-wikipathways-nanopubs-rett-syndrome/ | With WP8 | 20-50 |
| BioSchemas workshop 8 -9 Nov 2016, Rothamsted, UK | Bioschemas | 20-50 |
| ELIXIR all-hands 2017: CWL workshop: Hands on with the Common Workflow Language standards:describing command line tools and workflows in a portable, interoperable, and executable manner, 22 March 2017 | CWL workflows | 50 |
| ELIXIR all-hands 2017, ELIXIR-CHARME Workshop: FAIR data and data stewardship in ELIXIR: How to write your own FAIRy tale | Bioschemas | 30 |
| ELIXIR all-hands 2017: WP5 EXCELERATE and CHARME joint meeting, 20 March 2017 | General | 25 |
| EMBL-EBI industry programme meetings and SME forum 2017, 2018 | Ontologies | 15 |
| International Conference on Biomedical Ontology (ICBO), Newcastle, UK, 13 Sept 2017 | Ontologies, by EBI | |
| ELIXIR all-hands 2018: Access to ontologies workshop, ELIXIR All-hands meeting, Berlin, Germany, 6 June 2018 | Ontologies | 20 |
| ELIXIR all-hands 2018: EIP Session, 4 June 2018 | General | 40 |
| ELIXIR all-hands 2018: Bioschemas workshop, 6 June 2018 | Bioschemas | 30 |
| ELIXIR all-hands 2018: CWL workshop<br>CWL + Marine metagenomics use case, 6 June 2018 | CWL workflows | 20 |
| Forthcoming | | |
| ECCB: Bioschemas tutorials, 10 Sept 2018 | Bioschemas | 35 (expected) |
| NETTAB: Bioschemas workshop, 22/10/2018 | Bioschemas | |
| Biohackathon EIP-focused topics, expecting min. 20 working on Bioschemas hacking, 12-16 Nov 2018 | Bioschemas, CWL Workflows | 120 expected |
| ELIXIR SME event - Interoperability services for industry presentation Frankfurt , 16 Oct 2018 | General | |
| Bioschemas tutorial, SWAT4LS , 3 Dec 2018 | Bioschemas | |



Horizon 2020 grant n. 676559

# Appendix 7: WP5 Publications

## A7.1. Publications

Sarntivijai, S, Juty, N, Goble, C, Parkinson, H, Evelo, C, Lanfear, J, Blomberg, N, Corvallis, , Interoperability: Standardisation of identifiers, schemas, and ontologies for scientific communities in Proc International Conference on Biological Ontology 2018, Oregon, USA, August 7-10 2018

Sansone, S.A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. and Thurston, M. (2018) FAIRsharing: working with and for the community to describe and link data standards, repositories and policies. bioRxiv, p.245183.

Franck Michel & The Bioschemas Community. Bioschemas & Schema.org: a Lightweight Semantic Layer for Life Sciences Websites. In proceedings of the Biodiversity Information Standards (TDWG) 2018 Annual Conference, Dunedn, New Zealand, August 2018. https://doi.org/10.3897/biss.2.25836

Alasdair J. G. Gray & The Bioschemas Community. Bioschemas community: Developing profiles over schema.org to make life sciences resources more findable. In proceedings of the 6th UK Ontology Network (UKON) workshop, Keele, UK, April 2018.

Jansen, M., Carta, C., Roos, M., & da Silva Santos, L. O. B. (2016) The Organisation of Bring Your Own Data (BYOD) Workshops to Make Life Science Data Linkable at the Source. In SWAT4LS.

Sarala M. Wimalaratne, Nick Juty, John Kunze, Greg Janée, Julie A. McMurry, Niall Beard, Rafael Jimenez, Jeffrey S. Grethe, Henning Hermjakob, Maryann E. Martone & Tim Clark, Uniform resolution of compact identifiers for biomedical data, Scientific Data volume 5, Article number: 180029 (2018)

McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. (2017) Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol 15(6): e2001414. https://doi.org/10.1371/journal.pbio.2001414

## A.7.2. Publications by Use Cases

Pommier, C., Cornut, G., Letellier, T., Michotey, C., Neveu, P., Ruiz, M., Larmande, P., Kersey, P.J., Cwiek, H., Krajewski, P. and Coppens, F. (January, 2018) Data standards for plant phenotyping: MIAPPE and its implementations [W785]. Proceedings Plant and Animal Genome XXVI Conference. PAG. San Diego : PAG, Résumé,

Sarntivijai, S., Vasant, D., Jupp, S., Saunders, G., Bento, A.P., Gonzalez, D., Betts, J., Hasan, S., Koscielny, G., Dunham, I. and Parkinson, H. (2016) Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. Journal of Biomedical Semantics, 7(1), p.8.

Carta, C., Roos, M., Jacobsen, A., Thompson, M., Wilkinson, M.D., Cornet, R., Waagmeester, A., Van Enckevort, D., Jansen, M., Licata, L. and Via, A. (2017) January. The FAIRification of data and the potential of FAIR resources demonstrated in practice at the Rome Bring Your Own Data workshop. In CEUR Workshop Proceedings (Vol. 2042).

Robertsen, E. M., Denise, H., Mitchell, A., Finn, R. D., Bongo, L. A., & Willassen, N. P. (2017) ELIXIR pilot action: Marine metagenomics–towards a domain specific set of sustainable services. F1000Research, 6.

Petra ten Hoopen, Robert D. Finn, Lars Ailo Bongo, Erwan Corre, Bruno Fosso, Folker Meyer, Alex Mitchell, Eric Pelletier, Graziano Pesole, Monica Santamaria, Nils Peder Willassen, Guy Cochrane; The metagenomic data life-cycle: standards and best practices, GigaScience, Volume 6, Issue 8, 1 August 2017, gix047, https://doi.org/10.1093/gigascience/gix047

### A.7.3. Posters

K. Joeri van der Velde, Bart Charbon, Mark de Haan, Gert-Jan van de Geijn, Mateusz Kuzak & Morris A. Swertz. Implementation of Bioschemas for multiple patient registries. F1000Research 2018, 7(ELIXIR):1228 DOI: 10.7490/f1000research.1115914.1

Robinson M, Soiland-Reyes S, Crusoe MR and Goble C. (2017) CWL Viewer: the common workflow language viewer [version 1; not peer reviewed]. F1000Research, 6(ISCB Comm J):1075 (doi: 10.7490/f1000research.1114375.1)

Leyla J. Castro, Olga X. Giraldo, Alexander Garcia, Michel Dumontier & The Bioschemas Community. Bioschemas: schema.org for the life sciences. In SWAT4HCLS Poster Proceedings, Rome, Italy, Dec 2017.

Alasdair J. G. Gray, Carole Goble, Rafael C. Jimenez & The Bioschemas Community. Bioschemas: From potato salad to protein annotation. In ISWC 2017 Poster Proceedings, Vienna, Austria, October 2017.

Carole Goble, Alasdair J. G. Gray, Rafael Jimenez, Niall Beard, Giuseppe Profiti, Norman Morrison, & The Bioschemas Community. Bioschemas.org. F1000Research, 6, July 2017. (Poster at ISMB2017). https://f1000research.com/posters/6-1226

Niall Beard, Martin Cook, eta l, Bioschemas: structured data for life science using schema.org https://f1000research.com/posters/5-2296

Egon Willighagen1, Jonathan Mélius, Brenninkmeijer Christian, Stian Soiland-Reyes, Anwesha Bohler, Martina Kutmon, Andra Waagmeester, Alexander R Pico, Anders Riutta, Alasdair Gray, Christ Leemans, Colin Batchelor, Ola Spjuth, Nuno Nunes, Chris Evelo, Carole Goble https://f1000research.com/posters/5-2274, The BridgeDb framework, 15th European Conference on Computational Biology (ECCB) 2016,

**ELIXIR All Hands 2018**

Sarntivijai, S, Juty, N, Goble, C, Parkinson, H, Evelo, C, ELIXIR: Interoperability with a FAIR Purpose, https://f1000research.com/posters/7-744 **BEST POSTER**

Pau Andrio, Adam Hospital, Josep Lluís Gelpí, BioExcel & ELIXIR: towards interoperable and reproducible biomolecular research workflowshttps://f1000research.com/posters/7-994

Michel Dumontier, Rob Hooft Assessment of the FAIRness of the Core Data Resources: implementation study applying the new FAIR metrics in identifying steps to increase, https://f1000research.com/posters/7-799

Pier Luigi Martelli1, Giuseppe Profiti, Rita Casadio, Castrense Savojardo, Integrating ELIXIR Italy with ELIXIR Interoperability platform activities, https://f1000research.com/posters/7-763

### A.7.4. Specifications

Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M. and Scales, M. (2016) Common Workflow Language, v1. 0.

The Bioschemas Community (2018) bioschema.org

### A.7.5. Publications about the Services (separate to EXCELERATE)

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O. and Neumann, S. (2010) ISA software suite: supporting standards compliant experimental annotation and enabling curation at the community level. Bioinformatics, 26(18), pp.2354-2356.

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S. and Rice, P. (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics, 29(10), pp.1325-1332.

Jupp, S., Burdett, T., Leroy, C., & Parkinson, H. E. (2015, December) A new Ontology Lookup Service at EMBL-EBI. In SWAT4LS (pp.118-119). (poster)
Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D. and Booth, T. (2015) Tools and data services registry: a community effort to document bioinformatics resources. Nucleic acids research, 44(D1), pp.D38- D47.
Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, Michel Dumontier, A design framework and exemplar metrics for FAIRness, Scientific Data doi: 10.1038/sdata.2018.118
Olga Vrousgou, Simon Jupp, Thomas Liener, Tony Burdett, Sirarat Sarntivijai, Helen Parkinson The SPOT ontology toolkit: Semantics as a service**,** 18th Annual Bioinformatics Open Source Conference (BOSC 2017) Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017 (poster)
Simon Jupp, Tony Burdett, Danielle Welter, Sirarat Sarntivijai, Helen Parkinson and James Malone, Webulous and the Webulous Google Add-On - a web service and application for ontology building from templates Journal of Biomedical Semantics2016**7**:17 https://doi.org/10.1186/s13326-016-0055-3
Sylvie Maiella, Annie Olry, Marc Hanauer, Valérie Lanneau, Halima Lourghi, Bruno Donadille, Charlotte Rodwell, Sebastian Köhler, Dominik Seelow, Simon Jupp, Helen Parkinson, Tudor Groza, Michael Brudno, Peter N. Robinson, Ana Rath, Harmonising phenomics information for a better interoperability in the rare disease field, https://doi.org/10.1016/j.ejmg.2018.01.013
Simon Jupp, Thomas Liener, Sirarat Sarntivijai, Olga Vrousgou, Tony Burdett and Helen Parkinson OxO – a gravy of ontology mapping extracts, ICBO 2017 (Poster)

# Appendix 8: ELIXIR Implementation Studies

Pump-primed by EXCELERATE, a number ELIXIR Implementation Studies have followed-on and strengthened the work:

1. Enabling the reuse, extension, scaling, and reproducibility of scientific workflows Interoperability (with ELIXIR France, ELIXIR UK, EMBL-EBI, ELIXIR Finland) for WP6
2. Data Validation Interoperability (with ELIXIR Belgium , ELIXIR France, EMBL-EBI, ELIXIR UK) for WP7
3. Bioschemas (with ELIXIR Netherlands, ELIXIR UK, EMBL-EBI) for all WPs
4. Bioschemas: Community Adoption and Training (with ELIXIR UK, EMBL-EBI, ELIXIR France, ELIXIR Belgium, ELIXIR Netherlands, ELIXIR Italy) for all WPs.
5. Implementation Study on Data Identification and Interoperability (EMBL-EBI) for all WPs.
6. FAIRness of the current ELIXIR Core resources: Application (and test) of newly available FAIR metrics, and identification of steps to increase interoperability (with ELIXIR Netherlands, ELIXIR UK, EMBL-EBI, ELIXIR Italy, ELIXIR Sweden) operated by the Data Platform
7. ELIXIR integration from a user perspective (ELIXIR UK, ELIXIR Estonia, ELIXIR Belgium , ELIXIR Denmark, ELIXIR Switzerland, EMBL-EBI, ELIXIR Norway, ELIXIR France) operated by the Tools and Training Platform, to better interoperate the ELIXIR registries.

See https://www.elixir-europe.org/about-us/implementation-studies for further details.



Horizon 2020 grant n. 676559