ISBE Infrastructure for Systems Biology Europe

WP2

Data and Model management

ISBE WP2 report

Deliverable No: D2.3

Final Recommendations for a Data and Model Management Framework

Data and Model Stewardship for the ISBE Infrastructure.

June 2015

Carole Goble, UNIMAN
Katherine Wolstencroft, UNIMAN/UL
Natalie J Stanford, UNIMAN
Jacky Snoep, UNIMAN/Stellenbosch
Stuart Owen, UNIMAN
Renate Kania, HITS
Martin Golebiewski, HITS
Wolfgang Mueller, HITS
Sarah Butcher, IC
Nick Juty, EBI
Henning Hermjakob, EBI
Nicolas Le Novere, Babraham/EBI

Project funded by the European Commission under the Seventh Framework program for Research and Technological Development

SEVENTH FRAMEWORK PROGRAMME

Project ref. no.	INFRA-2012-2.2.4: 312455
Project title	ISBE – Infrastructure for Systems Biology Europe
Nature of Deliverable	R= Report
Contractual date of delivery	Month 33
Actual date of delivery	Month 35
Deliverable number	D2.3
Deliverable title	Final Recommendations for a Data and Model Management Framework in ISBE
Dissemination Level	PU
WP relevant to deliverable	WP2
Lead Participant	UNIMAN
Author(s)	Carole Goble (UNIMAN) Katherine Wolstencroft, (UNIMAN/UL) Natalie J Stanford (UNIMAN) Jacky Snoep, (UNIMAN/Stellenbosch), Stuart Owen (UNIMAN), Renate Kania (HITS) Martin Golebiewski (HITS) Wolfgang Mueller (HITS), Sarah Butcher (IC), Nick Juty (EBI), Henning Hermjakob (EBI), Nicolas Le Novere (Babraham/EBI).
Project coordinator	Richard Kitney
EC Project Officer	Keji-Alex Adunmo

Dissemination level: PU = Public, RE = Restricted to a group specified by the Consortium (including Commission services), PP = Restricted to other programme participants (including Commission Services), CO= Confidential, only for members of the Consortium (including the Commission Services)

Nature of Deliverable: P= Prototype, R= Report, D=Demonstrator, O = Other.

Executive Summary

European life science research is undergoing major changes in research practice, with actions to maximise the benefit of research output for all members of the life science community. The mission of ISBE is complimentary to this and aims to give life scientists in Europe easy access to an infrastructure that supports Systems Biology approaches in research. Systems Biology enables researchers to comprehensively understand, predict, and affect dynamic behaviour of biological systems, from cells through to organisms and even ecosystems, the skills required are often difficult to maintain in a single group. ISBE will provide a clear path of access to vital tools that enable all European life science researchers, irrespective of their knowledge and skill background, to study biological systems through the inter and intra-disciplinary means that make Systems Biology successful.

Key areas of support within ISBE will comprise broadly of high-end expertise in modelling and data generation technologies, and the storage, access, and integration of data and models produced from systems approaches.

Stewardship of data, models, and processes produced within ISBE will be a vital crosscutting component of operations, and will ensure the availability, usability, longevity, and provenance of data and models. To do this ISBE must establish standardisation, curation and cataloguing tools and practices, to ensure that ISBE contributors and users can produce, retain, maintain and exchange data that is (re-)usable for ISBE modelling and interoperable with other Research Infrastructures.

The value of stewardship is universally recognised but often more in principle than action: some £3 billion of public money is invested annually in research in the UK alone, yet the research data resulting from this considerable investment are seldom as visible as they might be. The German Research Foundation (DFG) estimates that 80-90 % of all research data is never shared with other researchers. These results are never published in a scientific journal and often hidden in a drawer in the laboratories. Thus, a majority of research data is lost because of un-sustained storage and lack of sharing of these data. The preservation and sharing of digital materials so others can effectively reuse them maximises the impact of research inspires confidence among the research councils and funding bodies that invest in the work.

To systematically examine ISBE's capability to support FAIR (Findable, Accessible, Interoperable, Reusable) data, model and SOP management we devised: (a) a User and Sector Stakeholder Analysis; and (b) an Asset Management Capability Framework.

ISBE Research Infrastructure will be made up of distributed resources and services at two levels:

1. **Specialist public archives** managed for the international community by national or pan-national providers that are: (a) asset-specific datasets; (b) public tools; (c) catalogues of datasets and tools. Providers may be aligned with nSBCs to contribute those resources/services to the ISBE Infrastructure or they may be part of another RI (e.g. ELIXIR) and their provision to the ISBE infrastructure contributed through MoUs and Service Level Agreements.

2. **Project outcomes** with locally deployable platforms and centralised resources to: support inherently integrated, cross-asset, cross-archive Systems Biology investigations; provide a **unified Sys Bio Commons** to the outcomes of European projects; and to in-the-field supported asset management in research projects, with publishing workflows into public archives and publisher repositories.

Stakeholders

Our stakeholder analysis is organised into six user categories and nine sector categories operating across three levels: institutional, national and international.

Users are identified as: researchers that are systems biology specialists or general bioscientists; application users from clinical/healthcare and/or commercial; and end user policy makers and citizens.

Sector stakeholders are identified as: funding agencies; vendors/commercial interests; employer/host institutions; scientific societies/community groups/networks; standards bodies/groups; research infrastructures; training initiatives; resource/service providers; and public and commercial scholarly communication bodies (notably publishers and libraries).

Asset Management Capability Framework

The Asset Management Capability Framework is a tool to: profile the current readiness / capability of ISBE; highlight priority areas for change and investment; and develop roadmaps. This Framework will serve as a systematic device for planning the Interim Phase of ISBE.

We extended an established framework, including the incorporation of the influence of users/sector stakeholders and their case studies and recognition of the Systems Biology method and the related stewardship lifecycle of Systems biology assets. For stewardship to be effective we identified technical, social, cultural and environment aspects of its implementation must be well managed.

Technical aspects include: how data, models and SOPs should be managed and exchanged within ISBE, and between ISBE and external resources; which formats, identifiers, standards and ontologies should be used, created and maintained for ISBE, and pathways to their adoption; and how interoperability between data and model resources many be achieved.

Social aspects include: how can compliance to the standards recommended by ISBE be encouraged or mandated; how can annotation and standardisation be made more straightforward and rewarding, and less time consuming, for scientists; how data, model and SOP planning and management can become embedded in Systems Biology practice and publishing; and how practices can lead to greater collaboration and openness for the research results of publicly funded research.

Cultural aspects include: how existing and new Systems Biologists in data and model management can be educated with respect to data, model and SOP stewardship; how other stakeholders such as funders, librarians and publishers should engaged in the importance of data and model management; how to drive change in the recognition of data, models and SOPs as first class, citable and creditable research outcomes; and how to establish career paths for data and model stewards.

Environment aspects include: how the community should select of the specific public resources and services to be ingested and sustained in the ISBE infrastructure; how to establish partnerships with other RIs such as ELIXIR; how to develop and implement business models for resources and services; and how to develop policies, and responses to ethical, legal, and commercial concerns.

Recommendations

1. **FAIR publishing**. All assets generated by EU researchers and projects and stewarded by ISBE recognised resources should be published FAIR - Findable, Accessible, Interoperable, Reusable/Reproducible. Data and models in the academic domain should be shared with the community as soon as possible. Linking individual researchers to their data and models, and providing persistent links to them, however, should enable scientists to gain credit for reuse of their datasets and models, encouraging an open, sharing culture. ISBE should establish FAIR guiding principles for the publishing of research data that should inform all decisions relating to ISBE's management of research data, models and SOPs. Implementation of the principles is the responsibility of all ISBE nSBCs and the cSBC.

2. **Stewardship in the service of predictive modelling**. Stewardship in systems biology requires all related research assets from a systems biology investigation (models, data, SOPs, samples, maps etc) to be aggregated and interlinked. The focus of ISBE is stewardship in the service of models. That is: model stewardship and simulation services; and data/SOP stewardship for collecting data for constructing and validating models and supporting the data results of predictive models. Legacy public archives may be transformed when possible, and dedicated archives constructed to suitably support quantitative biology. Stewardship practices focused on Systems Biology distinguishes ISBE from ELIXIR.

3. **Sustained, dedicated, public archives and repositories**. The modelling of biological systems based on integration of diverse data sets will rely on datasets being available that are suitable for integration. ISBE is responsible for the long term stewardship of strategically important research assets (data, SOPs, tools, maps and models). The research community's outcomes should, first and foremost, be placed in these sustained, dedicated, public repositories and catalogued by these sustained, public, dedicated registries. Data, models and SOPs generated by projects supported by the ISBE infrastructure/training, or publicly available and compliant with ISBE best-practice recommendations, should also be catalogued, archived and published in compliance with ISBE's FAIR principles.

   ISBE should seek to (i) *identify and sustain key established dedicated public repositories/registries* for the benefit of the community, seeking partnerships with other RI where appropriate, and *develop and sustain key missing dedicated public resources* where identified by users and stakeholders; (ii) establish, curate and sustain *a Systems Biology Tools and Resources Registry*, leveraging and aggregating pre-existing resources, in particular ELIXIR's registry; and (iii) monitor the usage, performance and quality of such resources against to be established metrics. Open and transparent processes and achievable and appropriate criteria need to be established. Selected, key, investigator-lead resources or assets will need to be migrated to become backed sustainably by nSBCs.

Compliance to the ISBE FAIR Principles will be a criteria for acceptance of a resource into the FAIR Infrastructure.

4. **A sustained Systems Biology Commons.** The modelling of biological systems based on *integration and cross linking* of diverse data sets. A Commons is a community controlled environment that brings together distributed research assets and distributed users/contributors. Systems Biology investigations are inherently integrated, cross-asset, cross-archive, cross-researcher (experimentalist, modeller), and often cross-lab. A Commons enables researchers to catalogue, pool (exchange, share, publish), cross-link, access, and analyse their own and public assets, using their own and third party tools. Benefits include: (i) aggregating repositories with contextual metadata; (ii) overcoming the fragmentation of the asset-specific repositories (iii) hosting experiment-specific, "boutique" datasets; (iv) retaining, and preserving assets of independent researchers; (v) driving compliance of standardisation practices; (vi) making project outcomes available for stakeholders and tracking their usage; and (vii) bridging research practice and research publishing.

   The key part of a Commons is the ***pan-asset, pan-repository*** catalogue that indexes and links the assets associated with a published investigation, which may well be stored in different repositories hosted by different organisations. Thus Commons are gateways to public archives to deposit outcomes, as well as access content, while retaining the connections to the investigation context and cross-links to related assets (models with data, data with SOPs etc). Commons use is governed by established regulations and policies for behaviours, for deposition and metadata standardisation, FAIR use, FAIR reuse and FAIR sharing with appropriate security, privacy and access controls regulated against a minimum set of community-accepted rules.

   ISBE should seek to (i) establish an **EU-wide Systems Biology Commons** that retains and catalogues the assets of Systems Biology projects in Europe; and (ii) monitor the usage, performance and quality of the Commons against to be established metrics.

   Compliance to the ISBE FAIR Principles will be a criteria for acceptance of a resource into the FAIR Infrastructure.

5. **Sustained stewardship services and technical services**. ISBE should provide a *set of services* to support both ISBE stewards and researchers to curate, archive and share research assets, including: data and model management planning; pathways for public publishing; and technical compliance validation of data and models against standards, policies and practices; authenticated and authorised and identified access; and data transfer. ISBE does not govern the science or scientific methodology that at undertaken using its infrastructure. That is the purview of peer review.

   The framework of services and resources must not dictate a single platform or a tightly integrated data infrastructure. Systems Biology is integrative by nature, drawing upon the ecosystem of data and model resources (legacy, emerging and provided by pre-existing or forthcoming Research Infrastructure (RIs)). In order to ensure sustainability, ISBE infrastructure, interoperability and compliance policies must be the minimal required for functionality, and devised in partnership with those RIs. The conventions for data and model services interoperability should be based on minimal "hourglass" approach, a specification of lightweight interfaces, standard protocols and standard formats.

6. **Support projects and researchers with asset management platforms.** For data, models and SOPs generated by projects supported by the ISBE infrastructure/training, ISBE should identify and support platforms that enable researchers, projects, institutions to manage their assets. Platform should to "RARE" research practices (Robust, Accountable, Intelligible, Reproducible) with workflows for "FAIR" Publishing using ISBE public resources.

7. **Support for commercially sensitive and personally sensitive data.** ISBE will support life sciences research, health research and commercial collaborations in these areas. Patient data for clinical or biomedical applications requires secure and sensitive handling. A mixture of open and commercially sensitive data/models and open and commercial services should be catered for. Commercial services may form part of the ISBE data and model framework: from the publishers and publishing services through to commercial data and knowledge bases and modelling tools and underpinning commercial cloud hosting. We anticipate potential financing as a public private partnership and the implications this may have on data visibility – its accessibility and accessibility. The operating conditions that ISMB should support private and proprietary data needs to be defined.

   Clear policies, standard operating procedures and supporting infrastructure are required to ensure that private health care information or commercial assets are kept with secure and restricted access (the "A" in FAIR stands for Accessible, not open). In some cases an Information Security Management System defined by Policies and Standard Operating procedures certified to ISO27001 will be required.

8. **Development and adoption of common practices and standards.** The ISBE data and model management framework focuses on conventions that enable data interoperability and stewardship and compliance against data and metadata standards, policies and practices. We must define, develop and adjust criteria and standards that must be met by data, maps, tools and models; support the accuracy, reliability and quality of data, models, tools and maps.; and make the re-use of data sets, models, SOPs etc. possible in future projects.

   The conventions for data and model metadata descriptions must be founded on community standards for identifiers, formats, checklists and vocabularies, developed through community engagement, to make data, models, and tools re-usable. A knowledge hub and training activities will be needed to disseminate these practices and standards, and technical development to implement them into tools.

   ISBE must be an active and engaged advocate for the development and adoption of standards. We recommend a concerted action of the European systems biology infrastructure with the respective ISO committees as ISO/TC 276 with the objective of defining a horizontal framework standard for the data and model patchwork in the field. Such a strategic alliance has to include the corresponding domain-specific grassroots standardization initiatives like COMBINE, FGED, PSI, MSI and others; wider standardisation bodies such as the Research Data Alliance and the Global Alliance; and work with journals and funders to establish practical best-practice usage of community standards for publication.

   ISBE should set in motion measures (training, services, and infrastructure) for the making and habitual use of standards for the research assets of Systems Biology, notably data, SOPs and models, and how these are related to each other and to investigations.

9. **Build stewardship capacity and capability.** When ISBE acts as a broker to bring researchers who generate data into contact with researchers who require data, standards-based and model-compliant data generation must be ensured along with data management planning. We will need *stewarding services* support to store and explore the links between data, models, protocols and results from ISBE investigations, showing the systems level details of the experiments, and to understand how separate datasets (e.g. genomics, transcriptomics and proteomics) can be interpreted together, or how they are used for construction or validation of the model, to enable a systems level understanding. *Training and education* is required across the different expertise of ISBE users, stakeholders and stewards, including members of nSBCs, ranging from in-house training to curriculum development for higher education institutions. ISBE must partner with international training initiatives such as GOBLET and Software Carpentry, and national initiatives such as SysMIC.

10. **The recognition of all assets and all stewardship activities**. Data and models must be citable and cited, with credit given to their authors and stewards, and commoditised so that they can be re-used modularly. Stewardship needs to be recognised and rewarded as a first class and habitual activity. Assets need to be recognised and rewarded as first class research outcomes with appropriate credit metrics. Dedicated stewards and Research Data Engineers, and those Research Software Engineers producing stewardship tools, should be recognised with established and rewarding career paths. ISBE should establish partnerships with stakeholders: institutions, funders, publishers, journal editorial boards, learned societies, pressure groups and networks (such as Force11) to advocate for the recognition of all assets and the recognition of the skills of asset stewards, develop supporting infrastructure and establish practical best-practice usage of community standards for creditable publication.

11. **Sustained funding and business models**. ISBE should seek avenues for sustainable funding for asset stewardship and public resources, and develop a portfolio of business models. transformative 5% tax. example: the Netherlands and NWO, DTL. Funding agencies and grant allocation could also allow funds to go directly to curation and stewardship activities, thereby facilitating the longevity of data and data accessibility in the longer term.

12. **Develop Synergies with other RIs and other partners**. Synergies should be identified across the various RIs in a systematic way; repositories that can provide data of use in ISBE should formalise agreements for data sharing and access, SOPs should be established for curation and annotation of datasets and models, with a clear policy established for responsibility for the data, in terms of where and how it is stored, and on the means it should be accessed by ISBE, and by the systems biologist.

13. **EU, national and community regulations and compliance vigilance**. ISBE must maintain awareness and vigilance with respect to EU and national regulations and compliance mandates. Regulation in ISBE is challenging as national and European regulations are at play. The most notable regulation is European Commission’s European Data Protection Regulation, which replaces the previous Data Protection Directive. The aim of the new European Data Protection Regulation is to harmonise the current data protection laws in place across the EU member states. As a “regulation it is directly applicable to all EU member states without a need for national implementing legislation. The regulation on the movement and processing of personal data is much tougher than previously. Other regulations are national or community standards for, Information Security Management Systems (ISO27001).

Implementation of Recommendations

Proposals for the process, actions and early implementations the recommendations in the Interim Phase of ISBE are outlined in Deliverable D2.4.

Table of Contents

1 Introduction

European life science research is undergoing major changes in research practice, with actions to maximise the benefit of research output for all members of the life science community. The mission of ISBE is complimentary to this and aims to give life scientists in Europe easy access to an infrastructure that supports Systems Biology approaches in research. Systems Biology enables researchers to comprehensively understand, predict, and affect dynamic behaviour of biological systems, from cells through to organisms and even ecosystems, the skills required are often difficult to maintain in a single group. ISBE will provide a clear path of access to vital tools that enable all European life science researchers, irrespective of their knowledge and skill background, to study biological systems through the inter and intra-disciplinary means that make Systems Biology successful.

Key areas of support within ISBE will comprise broadly of high-end expertise in modelling and data generation technologies, and the storage, access, and integration of data and models produced from systems approaches.

Stewardship of data, models, and processes produced within ISBE will be a vital crosscutting component of operations, and will ensure the availability, usability, longevity, and provenance of data and models. To do this ISBE must establish standardisation, curation and cataloguing tools and practices, to ensure that ISBE contributors and users can produce, retain, maintain and exchange data that is (re-)usable for ISBE modelling and interoperable with other Research Infrastructures.

The value of stewardship is universally recognised but often more in principle than action: some £3 billion of public money is invested annually in research in the UK alone, yet the research data resulting from this considerable investment are seldom as visible as they might be. The German Research Foundation (DFG) estimates that 80-90 % of all research data is never shared with other researchers. These results are never published in a scientific journal and often hidden in a drawer in the laboratories. Thus, a majority of research data is lost because of un-sustained storage and lack of sharing of these data. The preservation and sharing of digital materials so others can effectively reuse them maximises the impact of research inspires confidence among the research councils and funding bodies that invest in the work.

1.1 Objectives

- To identify the users and sector stakeholders for the ISBE Research Infrastructure data, model and SOP management.
- To establish a framework for developing the ISBE Research Infrastructure data, model and SOP management capability.
- To make recommendations to the ISBE Research Infrastructure regarding data, model and SOP management in order to meet the future needs of the community.
- To make a recommendations to ISBE Research Infrastructure to address and assess the impact of recommendations on sector stakeholders.

Deliverable 2.4 proposes actions and early implementations of the recommendations.

1.2 Methodology

The foundations of this deliverable lie within deliverable D2.1: Combined report on state of the art and horizon scanning. By identifying the wants and needs of life science/Systems Biology stakeholders that are currently met by available infrastructure, and taking the future requirements of life science/Systems Biology stakeholders within ISBE we have assembled the recommendations in this document.

More details on the methods can be found in D2.1. We list them here in brief for reference.

1. Three complimentary surveys for systematic collection of data for standards, formats and ontologies used in Systems Biology, data and model repositories used for deposition, and an audit of Systems Biology data and model management platforms. All of the results can be found in the appendix of D2.1.
2. Case studies of Systems Biologists were used to understand what typical data and model usage/transfer looked like in practice.
3. Text mining of the literature was used to compliment the surveys. We identified a total of 29477 Systems Biology papers in PubMed and extracted information regarding researchers within the community, references to any tools used, resources, standards and databases.
4. Desk research of e-infrastructure using EU and National reports, strategy documents, and briefing papers.
5. Meetings with other ISBE work packages, experts, national, EU and global initiatives.
6. A survey of institutional support for Data management undertaken by ISBE partners.

The document was also assembled through interactions with other ISBE work packages, chiefly:

- **WP3 (Overall infrastructure, eligibility and accessibility):** the organisation of the ISBE infrastructure; the provisioning and responsibility of data and model services across those centres; and the sources and sinks of data. Determines the physical interactions between distributed ISBE centres.
- **WP4 (Data Generation):** the source of raw and processed data. Work includes the readiness of data for Systems Biology and the responsibility of its preparation for interoperability, intelligibility and management through standardised and harmonised operating procedures and practices.
- **WP5 (Community Building and Synergies):** with a central portal for gathering and disseminating data and model management systems required and in use. Defines the user base and their functional requirements.
- **WP8 (Modelling infrastructure and expertise):** managing model types, multiple dimensions of space, time, chemistry and the cellular control hierarchy, multi-scale approaches and modelling formalisms, supporting the interplay between modelling and experimentation, and supporting the management of models in a pan-European modelling service.
- **WP9 (Technology and Science Watch):** data storage, compute infrastructure for model execution, data movement (data to models and models to data), data/model locality etc. Defines the user base and their functional requirements.
- **WP10 (Training and Education):** the training of modellers and experimentalists in data and model management practices, curation and archiving standards, adoption of best practices and compliance to open access and management policies. Training to enable the use of ISBE services and to promote the adoption of ISBE recommended standards and formats.

- **WP11 (Funding, Governance and Legal):** funding mechanisms and instruments for co-ordination and sustainability of data, model and SOP management infrastructure; and the implications of Intellectual Property, licensing and personal privacy (for patient data) on data and model availability.
- **WP13 (Connections):** data in particular is the commodity that is exchanged between ISBE nodes. Standard interfaces at ISBE nodes that enable computer assisted connecting and cross-node tasks include data and model interoperability and exchange standards and services. Determines the physical interactions between distributed ISBE centres.
- **WP15 (Innovation, Impact and Exploitation):** The affordability and quality delivered through the ISBE through exploitation of data and models managed by ISBE; and the management of intellectual property. Defines the user base and their functional requirements.

1.3 Proposed ISBE Infrastructure

ISBE will be a distributed infrastructure that provides services and resources to support world-class systems biology research. It will cover 5 strategic areas of services and resources required for producing successful systems biology research (Figure 1.1):

- Training and education
- Modelling
- Community activities
- Standards
- Data, model and SOP management and stewardship

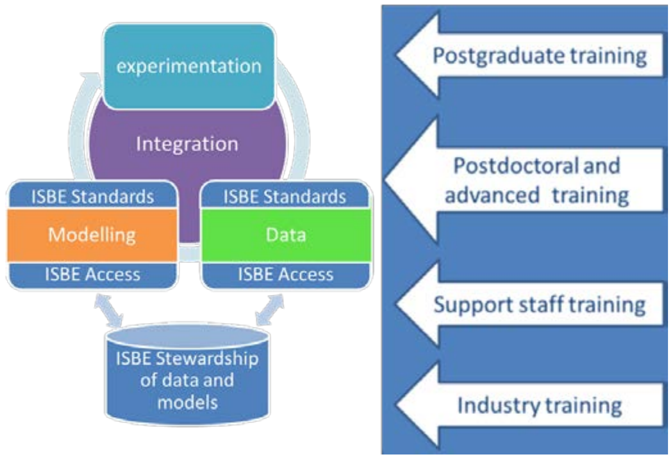


Figure 1.1: Strategic areas of service and res h.

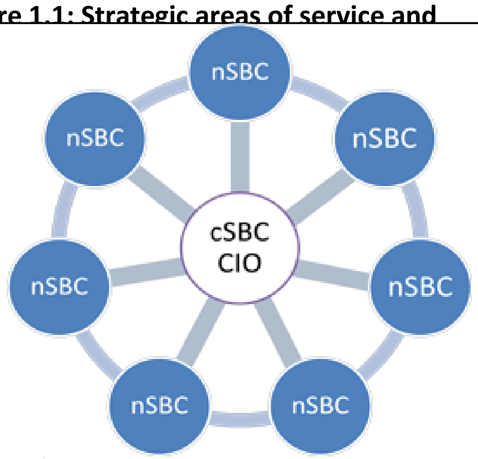


Figure 1.2: ISBE hub and spoke model.

ISBE will be structured according to a hub and spoke model (Figure 1.2). ISBE is represented on the international level by a single central Systems Biology Centre (cSBC) which is responsible for operational strategy (i.e. planning and reviewing present and future services). The cSBC will be connected to the national Systems Biology Centres (nSBCs), who will be responsible for ensuring delivery of services. At the national level the organisation is somewhat flexible depending how the country chooses to implement nSBCs. A country may consider a central institute to act as a co-ordinator with other institutes as partners, or the country may choose to implement a centralised body which co-ordinates all institutes that become members of it. Each of the nSBCs will contain component services and resources from any to all of the 5 strategic areas recognised as ISBE service priorities. The portfolio of nSBCs centrally coordinated by the cSBC will cover all strategic areas of services and resources.

The ISBE infrastructure is a complex network of physical and virtual resources designed to support a model-centric and data-centric approach to Life Sciences. Tilsley and Coveney present an infrastructure viewpoint that refers to: (i) data repositories, catalogues and libraries, and data services such as LIMS and citation tracking; (ii) software and algorithms such as modelling tools, and data/software management systems; (ii) underpinning "consumables" such as storage, compute and networks; and (iv) cross-cutting services such as access authorisation and authentication. Infrastructure also includes (v) people and their expertise: Systems Biologists who generate and use the data and models, data and model curators, systems administrators and so on.

The distributed, interconnected infrastructure envisaged by ISBE depends on the adoption of best practices, standards, technical infrastructure, and capacity for the management and distribution of data and models, and the management and sustainability of data and model management software. It is easy to overlook the fact that both data and models are entirely dependent on the software used to manage, access, search, run, exchange, regulate, validate them. In 2014 the UK House of Lords[1] went as far as to state that in fact infrastructure was software and that storage/compute facilities were consumables, a sentiment echoed in funding council's roadmaps[2]. The sustainability and maintenance of data and model management software is thus crucial to ISBE infrastructure.

Provisioning a common framework for the nSBCs and users will enable data and models arising from the ISBE infrastructure to be retained and managed. Adopting a common framework and standards will enable the FAIR exchange of data, models and SOPs between nSBCs and will allow scientists to (i) support the reproducibility of results; and (ii) discover and reuse these data and models for their own research. Adopting standards that are already in use in the wider Life Science community will additionally ensure easier exchange with external resources, such as those from ELIXIR, Euro-Bioimaging and BBMRI.

ISBE Research Infrastructure will be made up of distributed resources and services. ISBE aims to provide asset services and resources at two levels:

1. **Specialist public archives** managed for the international community by national or pan-national providers that are: (i) asset-specific datasets such as BioModels, SABIO-RK, Metabolights, BRENDA, JWS Online, COMBINEArchiveWeb etc; (ii) public tools such as COPASI for modelling and DMPOnline for data management planning; (iii) catalogues of datasets and tools such as res3data.org and ELIXIR Tools Registry, and metadata standards such as Biosharing.org. These support the Findability and Interoperability/Reusability of FAIR research outcomes. Providers may be aligned with nSBCs and those nSBCs will contribute those resources/services to the ISBE Infrastructure. Alternatively, they may be part of another RI (e.g. ELIXIR) and their provision to the ISBE infrastructure contributed through MoUs and Service Level Agreements. ISBE will also take advantage of, and partner with, general repository providers such as figshare and data infrastructure providers such as Dropbox.

[1] http://www.publications.parliament.uk/pa/ld201314/ldselect/ldsctech/76/76.pdf

[2] http://www.epsrc.ac.uk/SiteCollectionDocuments/ourportfolio/EInfrastructureRoadmap.pdf

2. **Project outcomes** with locally deployable platforms and centralised resources to support inherently integrated, cross-asset, cross-archive Systems Biology investigations. ISBE should provide a unified **Sys Bio Commons** to the outcomes of European projects (as identified by our Industry Survey). ISBE should seek to support asset management "in the field" for research projects, with platforms and services that offer a pathway for public deposition in the public archives and Commons and publisher workflows for publishing. Examples include the FAIRDOM Initiatives SEEK asset platform and FAIRDOMHub Commons[3].

Distributed nSBCs will provision a single point of access for data by users and sector stakeholders. The nSBCs implementing this ISBE Infrastructure are expected to: manage public resources; offer a **unified view** over resources generated and used, in the context of the experiments that produced them; and support the stewardship of research assets arising from Systems Biology experiments executed by users of the infrastructure.

The sources of data and models for the resources outlined above

- **nSBCs**, whereby centres are responsible for the stewardship of the models and data arising from projects or through contracts with sector stakeholders (such as funders or publishers). This content must adhere to ISBE's FAIR principles and comply to its conventions for best practice.
- **Sys Bio user community,** whereby content, independent of an nSBC, is contributed to datasets and model-sets managed by ISBE. To qualify for contribution, content must adhere to ISBE's FAIR principles and comply to its conventions for best practice.
- **Life Science community**, whereby datasets and their content are managed by RIs other than ISBE but vital to the ISBE community. Examples include Metabolights and Biomodels, managed by ELIXIR. Many such public quantitative databases provide kinetic constants for enzymes, and sometimes binding constants, but do little to help building quantitative descriptions, i.e. concentrations, sizes, diffusions etc. Limitations restrict the utilization of data for model construction and validation. ISBE will work with these providers to emphasises the importance of designing data collection against standardised SOPs for modelling experiments.

For a simple guide to data management needs, see Ten Simple Rules for the Care and Feeding of Scientific Data[4], which are:

Rule 1. Love Your Data, and Help Others Love It, Too
Rule 2. Share Your Data Online, with a Permanent Identifier
Rule 3. Conduct Science with a Particular Level of Reuse in Mind
Rule 4. Publish Workflow as Context
Rule 5. Link Your Data to Your Publications as Often as Possible
Rule 6. Publish Your Code (Even the Small Bits)
Rule 7. State How You Want to Get Credit
Rule 8. Foster and Use Data Repositories
Rule 9. Reward Colleagues Who Share Their Data Properly
Rule 10. Be a Booster for Data Science

[3] http://www.fair-dom.org, http://www.seek4science.org; http://www.fairdomhub.org

[4] Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4): e1003542. doi:10.1371/journal.pcbi.1003542

2 The Systems Biology Method

The ISBE infrastructure needs to be able to provide services which fully support each stage of the Systems Biology process. ISBE should support the management of the Systems Biology experimental life cycle: the generation, integration, validation and publishing of data and models, which will increase the reproducibility and comparability of ISBE experiments and promote reuse. In preparation for our Stakeholder Analysis and Capability Framework we set the context of assets - data, models and SOPs - in Systems Biology experimentation.

2.1 Systems Biology Lifecycle

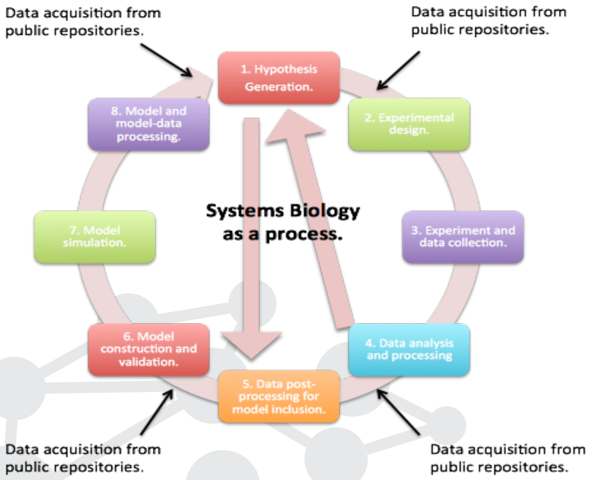

Data acquisition from public repositories.
Data acquisition from public repositories.
1. Hypothesis Generation.
8. Model and model-data processing.
2. Experimental design.
Systems Biology as a process.
7. Model simulation.
3. Experiment and data collection.
6. Model construction and validation.
4. Data analysis and processing
5. Data post-processing for model inclusion.
Data acquisition from public repositories.
Data acquisition from public repositories.

Figure 2.1. Systems biology research cycle.

Systems biology research operates as a continuous cycle where experiment informs model, and model informs experiment, as shown in Figure 2.1. The cycle contains two embedded cycles where hypothesis generation and validation can be supported with a half -turn through just an experimental, or just a computational (model) approach. Generating and validating hypothesis through these half-turns is usually reliant on the inclusion of data from public repositories.

Asset management must support the whole life-cycle of data and models through creation, consumption, storage and access for reprocessing. This would be a large burden and would be ineffective for individual research labs, leading to the generation of non-homogenous solutions that were not interoperable. The introduction of an overarching infrastructure such as ISBE, however, will ensure that these steps can be available as services, negating these issues. ISBE needs to provide a uniform and evolvable set of data and model management services to provide interoperable and integrated solutions that are available to all researchers. The geographical dispersion and inter-disciplinarity of these recent Systems Biology projects has only been made feasible by introducing bespoke platforms for inter-project data handling (e.g. SEEK4Science[5], part of the FAIRDOM Software Suite).

[5] Wruck et al, Data management strategies for multinational large-scale systems biology projects, Brief Bioinform (2012) doi: 10.1093/bib/bbs064 First published online: October 9, 2012

Stewardship is concerned with the aspects of Systems Biology related to data, reproducibility and provenance. The requirements in this respect from each stage represented in Figure 2.1 are presented below.

Table 2.1: The data and model management processes required per stage of the systems biology research cycle.

1. Hypothesis generation	• Find and Access relevant models so that current understanding of mechanisms and phenomena for a given behaviour can be identified and understood. The model must contain suitable annotations (organism, strain, modifications, included reactions/behaviour being modelled, where data is from, what conditions the model is valid under, authorship etc). These should be easily identifiable from public repositories. • Find and parse relevant data to identify support/opposition to the hypothesis, if any. • Find literature that supports the ideas (suitable linking between stored data/models and publication IDs).
2. Experimental design.	• Find what data are available, so that this can be Re-used, or the experiment can be designed as a reproducibility measure, or complementary data can be decided upon for collection. • Find relevant SOPs through public repositories. Access to the protocol should be open, or subject to request. The SOPs should be linked with data that are produced using them. • Access to an inventory of available groups and equipment which are relevant to the experiment being designed, with details of whether they perform services for certain aspects of experiments, whole experiments, rental time on equipment, and/or training in techniques/equipment.
3. Experiment & data collection.	• New SOPs are generated where new experiments need to be designed. • Raw data is typically large and difficult to handle (*see properties of data*). The data must be post-processed into sharing format for public release, and general interface management for ISBE. • Identify the raw data that will come from experiment and devise how it will be handled/stored. • Data for proteins metabolites, genes, transcripts, kinetics and microscopy data to be produced. This must be available for view in a unified-model centric format (integrated/interoperable)
4. Data analysis & processing.	• Find SOPs relating to the data analysis. Where none are available, SOPs for the new analysis need to be produced. • Systems biology data typically includes, but is not limited to, kinetic assay data and post-processed, large quantitative data sets such as genomics, RNAseq, proteomics and metabolomics. As systems biology advances the data types will broaden, already microscopic data for spatial and temporal modelling are being used. These data sets need to be structured and annotated. • Human physiology data is also a common source point for systems biology modelling. This can include measurements for heart rate, skin resistance, skin temperature, neuronal activity etc. These data sets are usually personally sensitive, non-homogenous, and non-reproducible. Therefore they require special handling for analysis and processing. • Multi-scale biology involves the interlinking, analysis and ultimately modelling of related biological phenomena that span vastly different times and scales. These can include molecular level sub-models, connected to generate a whole tissue functioning, or

	evolution of bacterial communities in non-homogenous systems. This data is notoriously difficult to analyse, interoperate, and model with current software – simulation in particular is difficult due to the scale differences in time and volumes, producing equations which tend to be “stiff” (numerically unstable unless the step size is taken to be extremely small). • Processed data needs to be compared with other data available in databases. It can be linked through data type, experimental protocol, organism, strain etc. From here it can be decided whether the data should replace older data, or whether it is complementary to other data sets. • Annotation of data sets suitable for inclusion into ISBE framework, and other RIs such as ELIXIR.
5. **Data post-processing for model inclusion.**	• Systems biology data-sets should be consistent regarding organism, strain, and experimental conditions, where possible. Modellers should be able to find complementary data-sets for inclusion within their models easily. • Data pertaining to human health may not adhere to specificity requirements owing to the inability to completely standardise conditions for collection, or samples themselves. • All data sets should contain metadata that describe the data such as organism, strain, and be directly linked to SOPs which detail the methods used for collection.
6. **Model construction & validation.**	• Models vary according to purpose and can be encoded into standardised Systems Biology formats (e.g. SBML and CellML), or be encoded within general languages (e.g. Python, Matlab, C++). Standard formats have the advantage of being able to be transferred and used within different software. • There are limitations with the standardised formats, in particular relating to spatial modelling or the need for Partial Differential Equations (PDEs). Non-standardised forms tend to be platform specific and therefore ways of sharing these models effectively must be established. Virtual machines for running non-standardised model formats would allow users to run models irrespective of their access to/ knowledge of the language used to code the model. • Other models with identical/similar cellular components (metabolites, proteins, pathways, tissue etc) should be identifiable and parameterisation differences should be cross comparable between the models. • Models can be generated using the same basic network design, but parameterised using data collected under varying conditions. This produces different instances of a model that can be released as (and indeed are) separate models and publications. There is a need to unify these models into an generic model which is capable of using/reproducing all data and findings from the collection of instances. • Models need to be tested for robustness using varying methods – and the results of this testing can help identify the validity of a model and its associated parameters. It is reasonable to expect a widely predictive, well parameterised model to be e.g. highly sensistive to parameter deviations. • The model needs to be verified and validated before moving on to steps of replication and reproducibility. • Models need to be accessible as an output for the paper/experiment/hypothesis etc, and organised so that they can be re-used. • The model needs to be tested for under-/over-fitting.
7. **Model simulation.**	• Model simulations need to be compared with the data sets that were used to construct it, as well as validated by data that was not used within its construction. This should be easily accessible in a visual way to the user. • The model should be compared to available data to identify whether/which data sets it supports/refutes.

8. Model & model-data processing.	• The model needs to be curated to ensure that is can faithfully reproduce any tables, graphs, and/or stated findings from the associated publication. The annotated storage of the curated model would be of much higher priority than data produced from model simulations, this is because researchers will re-use the model to generate data for publication, but would be unlikely to use simulation data. • Preserving the model requires for all components within the model, where possible, to have associated persistent identifiers.

2.2 Relationship between Data, Models and SOPs

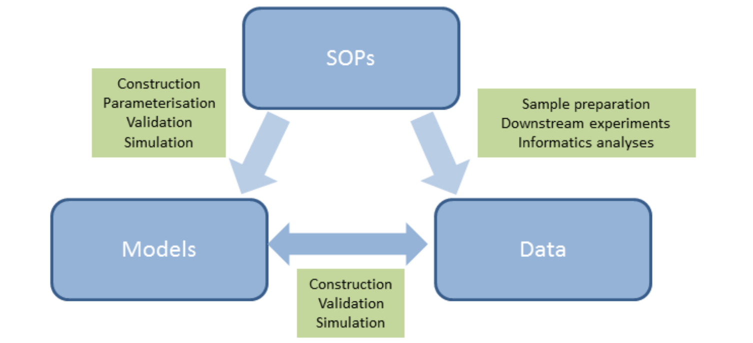


Figure 2.2: Relationship between data, models and SOPs

The Systems Biology life-cycle outlined above integrates data generation, analysis and modelling activities, but the relationships between data and models can take a variety of forms.

Data can be used for either constructing or validating models, which means that data generated in the laboratory can be directly fed into models as parameter values. Equally, data from the literature can be used in the initial model and laboratory data can then be compared with model simulations in order to validate the results. Model simulations themselves, however could also be considered as a type of data.

SOPs are related to both data and models. For example, there are SOPs and protocols governing the creation of samples, in order to ensure that all subsequent experiments are carried out on standard, comparable samples. There are also SOPs for the downstream experiments and the informatics analyses of the results obtained. In ISBE, SOPs will be essential for quality assurance across the data generation and stewardship centres and will assist in the understanding and therefore reuse of data.
SOPs for modelling are still rare. It is not yet common practice in the modelling community, even in large consortia. In ISBE, however, SOPs for different modelling techniques and procedures (for example parameterising a model) will be necessary for the same quality assurance reasons and to allow scientists to understand and reuse models. Figure 2.2 shows the relationships between data models and SOPs in systems biology investigations.

2.3 The Systems Biology Asset Lifecycle

The nature of scientific research means that hypotheses, and the data and models supporting them, evolve over time. This includes expanding data-sets, new findings which refute old ones, higher resolution/quality data from more advanced protocols/machinery, changes in the type of data collected, and new methods and mediums for modelling phenomena. In addition to this, the data and models that are created typically have a longer life-span than the projects that created them: the projects 'added value' come from being able to use these data and models in follow-up projects. In order to ensure that data and models stored within ISBE remain available, useful and relevant over the long-term, it is important that they are stored, updated and replaced at suitable times. This can be handled by the life-cycle model. Figure 2.3 shows three such cycles from different perspectives, with details of how data and model management should be approached for each of the steps detailed in Table 2.2. Lifecycle (a) is from an institutional/researcher viewpoint (based on Oxford RDM). Lifecycle (b) is from a European-scale Research Infrastructure view (ELIXIR). Lifecycle (c) is from a librarian or information science perspective (the UK's Digital Curation Centre[6]). These models for asset lifecycles (and there are many others) need to be aligned to support the Systems Biology method described above. We cannot separate data stewardship from software stewardship. Models, algorithms to analyse data, infrastructure, standards and software to deal with management, authentication, authorisation, security and privacy cannot be seen and developed in isolation from 'what we want with the data'.

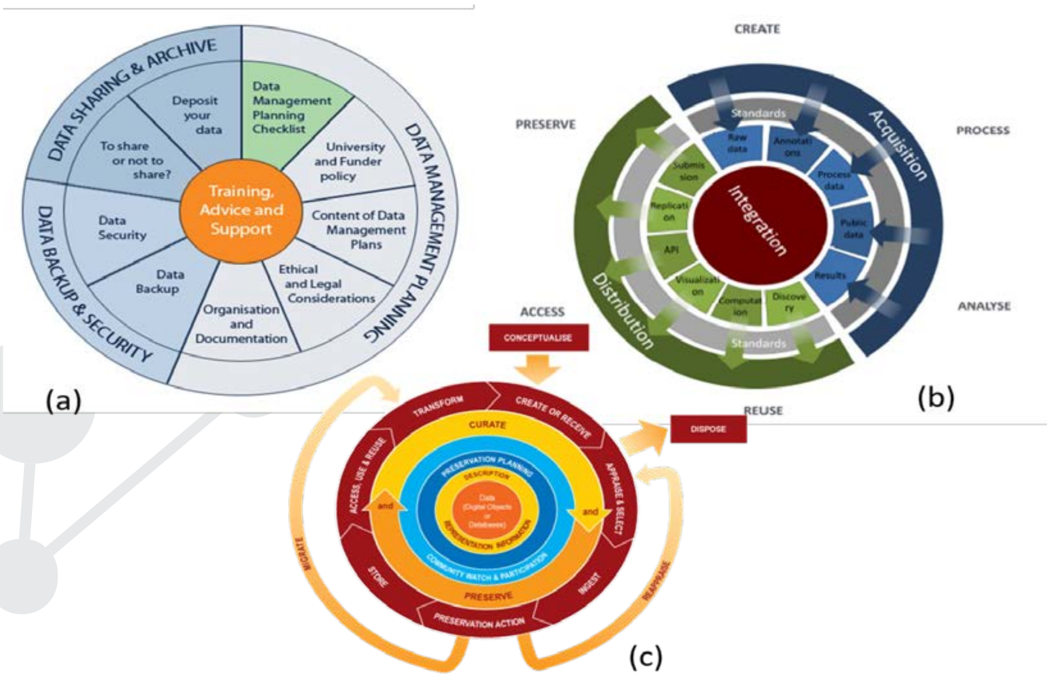


Figure 2.3: Life-cycle models for managing models and data.

Table 2.2: How ISBE should approach each step that form the data and model life-cycle management.

Step	Implementation
Plan	• Research asset (e.g data, models, maps, tools) protection requirements will be agreed in the planning phase.

[6] http://www.dcc.ac.uk

	• Available ISBE resources (standards, tools, equipment, databases etc) and in house resources should be identified at this point.
Create or Receive	*Model Generation* • ISBE nSBCs and users will produce preliminary models from available resources, in standard formats, in consultation/collaboration with ISBE users. • The community produce their own models using data available from ISBE repositories and/or data commissioned from an ISBE nSBC, and/or in-house data. • Models should be made available formatted and annotated according community best practice, where possible . Where this is not possible, original scripts can be maintained and the model can be set up to run in a virtual machine when accessed by the general public. • All models should contain metadata and should be curated upon publication to ensure that it reproduces the behaviour detailed in associated publications. *Data Generation* • Once users have identified available resources for their project, and flagged "missing" datasets, the missing data can be brokered by ISBE cSBC or nSBCs and/or generated by commission from ISBE nSBCs. • nSBCs ensure that data and models they produce use appropriate SOPs. Where no suitable SOPs are established new ones should be developed and shared. • nSBCs drive the community to produce their own data according to appropriate, established SOPs, and that SOPs be openly shared. • All data available to ISBE should be stored as post-processed, and conforming to ISBE standards for context, syntax, and structure. • Data should be made available to other ISBE centres and/or public repositories in compliance to appropriate data protection obligations established in planning. • ISBE cSBC, through partnership agreements with other RIs e.g. ELIXIR can commission the collection of systems biology ready datasets for community use.
Appraise and Select	• Evaluate data and select for long-term curation (where appropriate) and preservation. Adhere to documented guidance, policies, agreements, or legal requirements. • All data/models from projects explicitly supported by ISBE resources/services (defined by some form of contract), and data/models produced by ISBE nSBCs (through internal work, or paid for services) should be stored and shared according to the associated data policy – it is expected a minimum term of data/model storage/availability will be 10 years. • All available high-value data that complete current data sets for modelling, or form parts of new required sets will be obtained and made accessible through ISBE resources. • Post-processed, final data sets, suitable for inclusion into models will be of high value to ISBE users, so should made widely available in commons interfaces. The raw or pre-processed data should be stored appropriately (e.g. in specific archives such as Pride, or locally but accessibly) and linked to via the commons. • All stored data and models must meet a minimum requirement for ISBE quality which includes suitable meta-data mark-up. These requirements, and appropriate resources to support them will be identified by ISBE and training guides/courses available. • nSBCs with data integration expertise can incorporate newly generated data into models (adhering to ISBE best practice) and share these models through ISBE resources, according to the agreed data sharing policies. • Data and models that do not meet the minimum best practice/quality requirements of ISBE may be restricted in which ISBES resources can be used to share the data/models. • The inclusion of data and models into ISBE supported resources will depend upon the data and model management agreement between ISBE and the supported project.

Ingest	Ingesting can include: • Data or models submitted to ISBE nSBCs irrespective of privacy settings. • ISBE data from point of collection through to final data sets. • Other ESFRI data at the point of functional grouping for modelling purposes. • Models generated within ISBE nSBCs at point of first construction. • Legacy models seen as valuable for the community e.g models useful for teaching and training, appropriate groundbreaking models, or even test models that have refactored into standardized formats. • Submission of non-ISBE data will be primarily user controlled, through personal research spaces. • Submission of ISBE data will be made by the data handlers and where extra information is required can be flagged and sent to relevant other ISBE data handlers. • Any data or models generated by an ISBE nSBC will automatically be ingested into ISBE. • Appropriate datasets from partnered infrastructures will also be annotated to ISBE best practice and stored in ISBE specific, or ISBE shared resources. Individual nSBCs are responsible for curating any data they broker/commission/assimilate/generate in accordance with ISBE best practice. • The responsibility for curating data/models in accordance with best practice will lie with the data/model generators, unless projects have specifically requested services in curation of their model and data.
Preservation	• All ISBE managed data/models will be continually transformed in accordance with the latest community best practice where appropriate(e.g as standards update). • All modifications made to data/models for preservation purposes will be documented with the data. • All data brokered through ISBE cSBC and nSBCs should be checked and maintained for best practice compliance by a relevant nSBC. • Data/models submitted to ISBE resources from users will be graded for usefulness based on formatting and suitable meta-data markup (guidelines for best practice will be available). • nSBCs will be responsible for producing a unified view of ISBE activities (i.e. linking 'sets' of data, models, SOPs and experimental descriptions). This will ensure associations are preserved and different versions are recorded. • Modularisation of model libraries, first on model structures. Modularisation of parameterized models is difficult - mixing them together may be possible but not scientifically valid. This problem becomes pronounced for multi-scale modelling - particularly how they interact. • Certain models can be refactored into standardised formats as part of ongoing work from nSBCs. This is an informative activity for building the standards required to support complex modelling. • Models unsuitable for standardised formatting can be preserved and made accessible using a virtual machine. • All data should be post-processed - little to no raw data should be included from any experiments - however full data handling procedures up to point of delivery need to be documented, and linked to SOPs that demonstrate validity of method.
Store	• Research assets stored in ISBE resources will be stored in accordance with the associated data sharing policy agreement which includes. • Length of storage requirement. • Security requirements (e.g. for personally/commercially sensitive data). • Back-up and replication; for example using the EUDAT B2SAFE service or LOCKSS.

	• Selection of data to be long term archived. Not all data needs to be immediately available. A tiered storage (immediate access, short term archive, long term archive) will need to be defined; immediate access is the only one that needs expensive spinning disk solutions.
Access, Use and Reuse	• Completed nSBC projects will share the associated research assets with all ISBE nSBCs, and the wider scientific community (according to data sharing policies). • All research assets should be accessible on a day-to-day basis. This requires upkeep and monitoring of access platforms, and quick responses to any access issues that arise. • Privacy policies on ISBE data and models, and restricted access rights on data and models uploaded to ISBE, should be stringently adhered to. Exceptions should only be made where personalised access rights are counter to the official data sharing policy the data was created under (e.g. must be public after so many years - assuming no additional issues such as patent applications or pending publications). • All published data and models stored within ISBE should be made publicly available and linked to the corresponding publication. • ISBE users will be able to access research assets programmatically and through user interfaces: catalogues, APIs, portals. • Persistent identifiers (e.g. DOIs) will be available which can be linked to from a number of different platforms (publications/ blogs/personal web pages/LinkedIn/ResearchGate etc). • Semantically linked research assets e.g. data, models, SOPs, and software that were used to produce a single publication, can be packaged as research objects for download. • All ISBE models should be distributed with detailed instructions on use and re-use, this could be in the format of 'read me' files. • All models added to ISBE or ISBE related databases are encouraged to provide suitable detailed instructions to aid users in use and re-use of the model. Instructions of suitable 'best practice' will be available. • Promote data: through RSS feeds, Altmetrics, links into publication repositories such as OpenAIRE; through publisher platforms like F1000. • Sharing protocols will be established on any data stored within ISBE, covering immediate, short-term (whilst the project is still running), long-term (after the project ends up to a maximum of 10 years), post-requirement (what happens to the data after 10 years). • Links to existing shared data need to be made.
Transform	• nSBCs will be responsible for transforming relevant existing data sets to formats that can be used in Systems Biology projects, where appropriate.
Dispose	• Research assets that are not suitable for long-term inclusion into ISBE resources, based on specific policies, guidance or legal requirements must be: • migrated to a different storage source, where appropriate. • deleted permanently, if invalid. • destroyed securely, depending on nature of data.
Reappraise	• Research assets generated by ISBE nSBCs or brokered by ISBE nSBCs should be appraised for appropriateness of formatting and annotation. Where it does not meet ISBE standards, it should be improved. Adhering to ISBE best practice is a prerequisite for ISBE research assets to be made available, and is mandated for all ISBE nSBCs. • ISBE does not have responsibility for the scientific quality of research assets available in ISBE resources, however they do have a responsibility to flag research assets that are not well formatted or annotated such that they are reusable. These should be flagged to users.
Migrate	• ISBE data and models may be migrated to different formats to ensure that data does not become unusable due to hardware or software obsolescence.

3 User and Sector Stakeholder Analysis

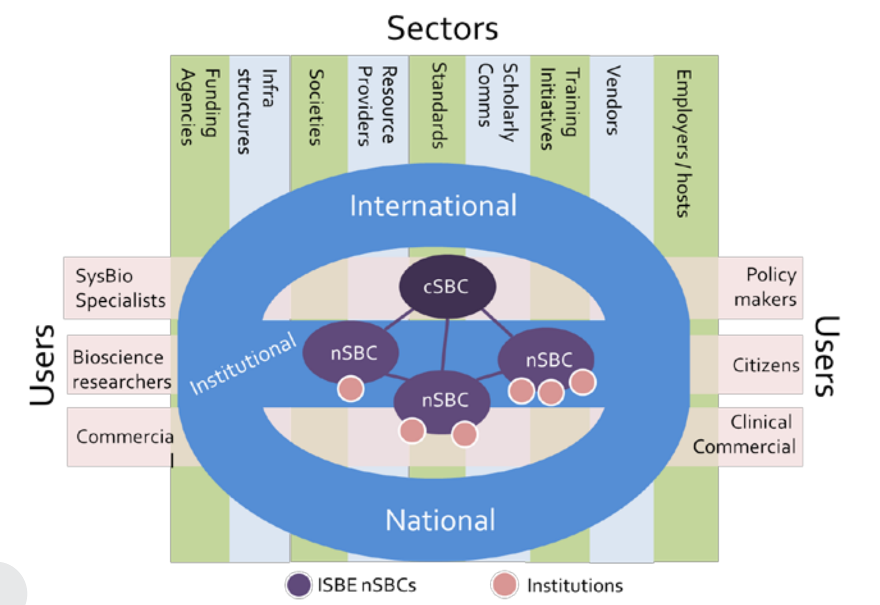


Figure 3.1: ISBE users and stakeholders.

Our stakeholder analysis is organised into six user categories and nine sector categories operating across three levels: institutional, national and international (Figure 3.1). This elaborates the users identifier in WP3 (researchers from academia (novices and experts); industry (SMEs and large); and non-scientists (funders, policy makers, politicians, publishers, digital libraries, patient organisations, press etc.)). Stakeholders may be consumers of the ISBE Infrastructure and/or providers contributing to the Infrastructure.

Users are identified as: researchers that are systems biology specialists or general bioscientists; application users from clinical/healthcare and/or commercial; and end user policy makers and citizens.

ISBE users are foreseen to be from a large breadth of the research community**.** They may be independent or large, multi-site and multi-partner collaborations. Researchers operate on a day to day basis at their host institution level impacted by local policies on Research Data Management, the availability of curator expertise and the role scholarly outcomes have in promotion criteria.

The ISBE infrastructure also aims to support small, specialised research groups, independent research fellows, and the usual large groups and known collaborators to be working towards a common goal, with access to high quality systems biology knowledge and data. Investments in individual fellows and small research groups could still lead to large community impact, along with larger more intensive research projects such as the Virtual Liver Network. Currently this unlikely due to the range of skills a small group or individual would need access to but currently does not have or cannot acquire.

Sector stakeholders are identified as: funding agencies; vendors/commercial interests; employer/host institutions; scientific societies/community groups/networks; standards bodies/groups; research infrastructures; training initiatives; resource/service providers; and public and commercial scholarly communication bodies (notably publishers and libraries). Examples can be seen in Table 3.1. Drivers for data and model management can be seen in Tables 3.2 and 3.3.

Table 3.1: User and Sector Stakeholder Analysis. Several categories may apply to one organisation.

Users	**Examples**
SysBio Specialist researchers	International projects (e.g. http://www.nano3bio.eu, VPH); EraNets (ERASysAPP, SysMO); national programmes (German VLN)
Bioscience researcher	Generalist
Commercial	Industrial biotech, pharmaceuticals, agritech; SMEs and international concerns.
Clinical and health care	Hospitals; biobanks.
Policy makers	Projects (e.g. http://www.synenergene.eu/), Initiatives, Policy institutes, government departments; NGOs.
Citizens	The general public; patient advocate groups; NGOs.
Sector Stakeholders	
Funding agencies	National research council (e.g. BBSRC, BMBF, NWO); Charities and foundations (e.g. Wellcome Trust); European Commission; commercial contractors and sponsors.
Vendors/ commercial interests	Research data management services (figshare);
Employer/host Institutions	Universities. National Capability Institutes (e.g. TGAC (UK), DTL (NL). International Institutes (e.g. EBI). National research centres (SynthSys (UK)).
Research infrastructures	European RI (ELIXIR, EUDAT, EGI, OpenAIRE, Zenodo, IMIs); national (SyBIT, de.NBI); commercial (e.g. figshare, dropbox, googledrive); public (dataverse, dryad); collaboration tools (e.g. ResearchGate)
Resource/service providers	European labs (e.g. EMBL-EBI), national labs (e.g. TGAC), independent institutions/projects (e.g. SABIO-RK).
Training initiatives	European (e.g. RITrain, CORBEL); National (e.g. SysMIC)
Scientific societies/ community groups/networks	Learned Societies (ISSB); Community groups (e.g. Force11, OBF, BioCurators, RSE); National Networks (e.g. NIBB, MSBN (UK)); international networks (e.g. METSYN).
Standards bodies/groups	Community grassroots (e.g. COMBINE, PSI, FGED); pan-national general initiatives (e.g. RDA); established bodies (e.g. W3C, ISO); non for profits (e.g. Digital Preservation Coalition);
Scholarly comms bodies **Public and commercial**	Institutional and national libraries; journal editorial boards; publishers (e.g. FEBS, Elsevier); lobby groups (e.g. Force11); venders (e.g. figshare, impactstory, Mendeley); infrastructure providers (e.g. datacite, orcid); national centres (e.g. UK's Digital Curation Centre, Software Sustainability Institute)

Table 3.2: Sector stakeholder drivers for asset management

Driver	Reasons	Prime Stakeholder
Capitalising	• Reuse of existing platforms • Pool capacities • Sustained asset management • Knowledge Transfer	Funding agencies, Employer/host Institutions
Skills	• Build knowledge networks • Trained researchers	Funding agencies
Justification & Compliance	• Audit investment outcomes • Metrics for renewals and reinvestments • Public funds accountability	Funding agencies, Employer/host Institutions
Reproducibility	• Reproducible publications • Peer review	Funding agencies, Scholarly Comms bodies
New publishable assets	• Data, model, SOP publishing • Research Object publishing • Software publishing • Publications companion sites	Scholarly Comms bodies
New business models and services	• Citation analytics for new assets, • Cloud repositories (figshare), • Vendors (ImpactStory), • Open Repositories (OpenAIRE)	Scholarly Comms bodies

Table 3.3: User drivers for asset management

Driver	Reasons
Doing Great Science	• Storing, analysing, modelling, collections • Sharing: dynamic memberships, modellers+experimentalists
Resources	• Local management, Sustainability • Exploiting local+public resources and tools • Managing curation cost vs benefit • Skills
Publication	• Reproducible publications • Release paradigm: model evolution • Deposit into public archives • Credit and citation for assets • Releasing results
Compliance	• Institutions, funders and publishers

3.1 Stakeholder Case Studies

ISBE is an infrastructure that aims to support researchers in all types of Systems Biology. Although it will connect researchers and promote collaborations it is not a network. As an infrastructure it has limited command over the science that uses it: for example, the data that it manages should be secure, well documented, and accessible and of suitable quality for use by Systems Biology; but how it is used by a Systems Biology researcher is not ISBE's concern. Case studies include individuals or projects consuming

ISBE services (potentially producing new assets); ISBE nSBCs performing or supporting the routine work of model and data production; and ISBE nSBCs supporting public resources that will be the sources and destinations of data, models and SOPs.

3.1.1 Stakeholder case study 1: ISBE and the researcher.

Sarah is the leader of a Computational Biomedicine group based in the UK. She is looking to model the changes in iron metabolism within cancerous cells. The project requires the generation of 6 different data sets (a mixture of high throughput and single cell analysis) which Sarah does not have the expertise for in her group. The expertise for producing the data is distributed across 3 different European centres, and the data is legally sensitive. Sarah also wants to couple her model with an already available ISBE cell cycle model.

ISBE Scenario: The raw data is collected, structured, and annotated according to available and agreed SOPs in two of the ISBE nSBC. The raw data is then stored in an "Embassy cloud[7]", to be accessed and post-processed by a third nSBC, according to relevant SOPs, into sharable formats (structured and annotated according to community and ISBE defined minimal standards). The share-format data is loaded into ISBE specific databases, and made available privately (length defined by client/ISBE/legal requirements) to Sarah in a data-unified interface. The model is constructed by Sarah's group through consultation models from an nSBC to ensure that its structure and format is compatible with the cell cycle model Sarah wants to integrate it with. After the full model is constructed and integrated with the cell cycle model, it is uploaded into a relevant ISBE model database where it can be kept private, or shared with collaborators until publication. At the point of publication the model and data are made available to the public subject to legal restrictions governing the data. The model is curated such that all data can be directly linked and identified with model components.

Impact: 5 sets of high quality data are released into the public domain, and are available for other projects to use, subject to legal restrictions. Provenance of the data and model are available and will be tractable through the lifetime of the data and model. The public can access the model and simulate it using ISBE simulation services. Other researchers can (re-)use the data and model for their own research, and satellite work based on this work will be tractable by the community. Sarah's group can be credited for their input into new projects.

3.1.2 Stakeholder case study 2: ISBE and the journal.

Systems Biology at Multi-Scale is an open-access journal dedicated to publishing the growing number of multi-scale models developed within the Systems Biology community. They have strict policies for publishing models: (i) all data used to construct the model must be available in the public domain, fully annotated to ensure reproducibility, and directly traceable to and from the model; (ii) All models must be publicly available, structured and annotated according to community standards, and "simulatable" for (re-)use by the community. (iii) The model must be able to reproduce all the finding in the paper; (iv) The data and model must be guaranteed to be available, and (re-)usable, in the public domain for at least 10 years post-publication.

[7] http://www.embl-em.de/downloads/5/EMBL-EBI_Embassy_Cloud.pdf

ISBE Scenario: The Journal can work directly with nSBCs in order to turn the requirements into a functional set of formats and annotations for authors to follow. nSBC train staff from the journal in data and model curation, submission and interlinking. ISBE provides temporary data and model areas that are private for reviewers to access. Upon publication the data and models will be referred to the trained journal staff who ensure the formats, and metadata standards of the data and model are suitable, that acceptable cross linking is present, and that the model produces the findings in the paper correctly. This is then submitted to permanent, publicly accessible (subject to any legal restrictions) storage facilities, where the model and data can be viewed in a unified interface. The data will be stored there for at minimum the lifetime of 10 years required by the Journal.

Impact: Journals want to publish high impact, highly cited research. A barrier to this is often the lack of availability of the datasets, models and SOPs included in journal papers. Poor availability of these assets prevents other researchers assessing the quality of the research, and also being able to use the research to build on within their own work. This will reduce the impact of the research on the community to the detriment of the journal, and the researchers who submitted the work. The standards imposed by the journal, and guided by ISBE mean that articles within the journal are more accessible by readers and therefore also more re-usable. This will lead to higher citations for the journal, and improved research reuse in the community.

3.1.3 Stakeholder case study 3: ISBE and the national research council.

A National Research Council (NRC) wants to ensure that the Systems Biology research it funds has the highest impact possible both in Europe and globally. They have identified that one of the key weaknesses in long-term asset storage from their funded projects is accessibility and (re-)usability. They want to devise a strategy to be implemented on all future funded projects that will overcome these issues.

ISBE Scenario: The NRC consults with ISBE about its requirements for future Systems Biology projects. Data handling frameworks will be established between NRC and ISBE, and a full set of recommendations for data and model formatting, annotation, and storage will be defined and made available for reference by holders of future successful grants. Training courses can be designed by ISBE and made available voluntarily or mandatorily to future grant holders.

Impact: When funding projects with public money, especially those with large budgets, it is vital that all assets of suitable quality are made available to the public. By establishing data management and stewardship practices early, and making this a requirement to researchers it improves the likelihood that funded research will achieve higher impact. The development of suitable training made available to grant holders increases the likelihood of the practices being followed correctly. A centrally managed framework means that groups do not have to waste time and resources developing their own formatting, annotation and storage procedures, and therefore reduces the burden and the cost to the researchers whilst allowing the NRC to achieve their goals.

3.1.4 Stakeholder case study 4: ISBE and the citizen.

Joe is diabetic and as an avid DIY-biologist is interested in how his blood sugar level impacts the metabolic behaviour of his organs.

ISBE Scenario: The Consensus Human Diabetes Model is stored in a standardised format in an ISBE managed model database. The database is searchable using key-words allowing Joe to find the model quickly. The model has several associated links including the open-access paper it was published in - with a public summary, the patient data that was used to build the model, and services for simulating the model. After reading the paper Joe can understand the basics about what the model does. After launching the simulation, he alters the blood glucose levels through many different ranges. After spotting some clear changes in behaviour, he uses identifiers in the model that link to external resources, in order to understand their function. Joe soon discovers the wide-reaching impact that deviations in his blood sugar levels can have over the short and long-term. He signs up to receive automatic notifications for when the model is updated.

Impact: An open, well managed, and easily accessible infrastructure is not just useful for research scientists; it is also a powerful resource for the enquiring public. The careful storage, annotation, and linking of resources within ISBE has allowed someone with little expert knowledge to gain access to information that impacts their understanding of a common disease.

3.1.5 Stakeholder case study 5: ISBE and the commercial user.

DSM have heard that a large EU consortia project has identified a full set of plug-and-play genes that can be inserted in yeast in order to produce a huge range of commercially important chemicals.

ISBE Scenario: All plug and play genes, yeast chassis, and associated models are stored in ISBE, and semantically linked through a commons interface. DSM are able to access the data and models, and reproduce the experiments in order to determine which yeast chassis, and plug and play genes are suitable for scale-up with DSMs current capabilities.

Impact: Businesses have fewer resources for high-risk exploratory research, however promising findings such as that from the EU consortia can be tested and ramped up into testing and production quickly with the resources industry has available to it. DSM are able to quickly identify which products they produce cost effectively at a large-scale, and commission the production. This leads to large-scale cheaper and sustainable production of very expensive antiviral drugs, used to treat HIV, producing an immediate impact on public health in the EU.

3.2 Selected Sector Stakeholder reviews

3.2.1 European Research Infrastructures

ISBE will need to work closely with other EU RIs, most notably ELIXIR. Consultations between ISBE and ELIXIR aiming at synergising their activities have started. From the data and model management perspective, and that of the SCs, consultations are also needed with a number of other EU-RIs.

- ISBE must clarify its expectations, responsibilities, approaches, resources and services, both technical and social, with key domain related RI, notably the ERANets, ELIXIR, Euro-Bioimaging, BBMRI, and the IMIs. It is not possible to define what will be provided by ELIXIR because it has yet to define this. We must agree to common data and service interoperability standards; decide who is responsible for which key resources and agree to sustainability and access Service Level Agreements. In order to be tractable, ISBE must carefully select the resources it plans to make core to its infrastructure. We should engage where possible with the ELIXIR Programme of Work where it is related to ISBE, and do the same for other programmes.

• ISBE must clarify its expectations, responsibilities, approaches, resources and services, both technical and social, with key cross-domain RIs, notably the EUDAT and OpenAIRE. The ISBE SCs need to take advantage of the services available (some of which may be mandated by the EU), and proactively ensure that the services are appropriate for ISBE data stewardship and sustained, through Service Level Agreements and Joint Ventures.

Key EU-RIs are as follows:

ELIXIR (http://www.elixir-europe.org/)

ELIXIR is an ESFRI that has now entered the implementation stage, with seven signed up nations at the time of writing and many more preparing to sign. Its aim is to *orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments*. Some of these datasets are highly specialised and would previously only have been available to researchers within the country in which they were generated. ELIXIR is creating an infrastructure that will integrate European research data, ensure a seamless service provision and make access easy and open. Its scope extends to data access, data stewardship, high-performance computing, the interoperability of public biological and biomedical data resources, and scalability. ELIXIR is organised as a coordinating Hub, hosted at the EMBL-EBI, and nodes. The nodes are national, hosted and (sometimes) funded by their nation states, with the exception of the EMBL-EBI which is also a node.
ELIXIR's role is primarily the compliance and governance of infrastructure chiefly provided by the nodes. ELIXIR has recently won a €20million implementation project. Infrastructure activities include:

- The (re)organising and interoperating resources that are contributed by the nodes; for example, Sweden is contributing HPA; Switzerland is contributing SwissProt. Other nodes are "mini-ELIXIRs" in their own nations or have a single cross-node focus. All nodes balance the need to serve their nation and the requirement to contribute to the overall ELIXIR community.
- Pilot projects funded by contributions from nodes and the hub, between nodes and between the node and the hub. Pilots of interest to ISBE's data and model management include:
    - *Safeguarding resources*. Establishing the EGA as a Joint Venture between ELIXIR nodes.
    - *Private, virtual workspaces in the ELIXIR data infrastructure*: creating a virtual working environment next to the reference data, with seamless access through their host institute.
    - *Seamless, uninterrupted transfer of major datasets across Europe*. The transfer of major European Genome-phenome Archive (EGA) datasets between the UK and Finland.
    - *Secure access to genomic data through distributed authentication* to support Data Access Committees with electronic application tools and is endorsed by Geant3Plus.
    - *Interoperability of protein resources for drug discovery*; Swedish and EMBL-EBI nodes to work together to make the Human Protein Atlas interoperable with PRIDE, the proteomics resource; InterPro, the database of protein families and motifs; and the Gene Expression Atlas.
    - *Software and Data Carpentry* rollout across ELIXIR Nodes

    - *Research Data Alliance ELIXIR Bridge Interest Group*
- Five Platforms define the work programme, to be delivered by cooperating nodes in partnership with international initiatives and other RIs: Data, Tools, Interoperability, Compute and Training. Cross cutting activities include: data management planning and use cases.

EUDAT (http://eudat.eu/)

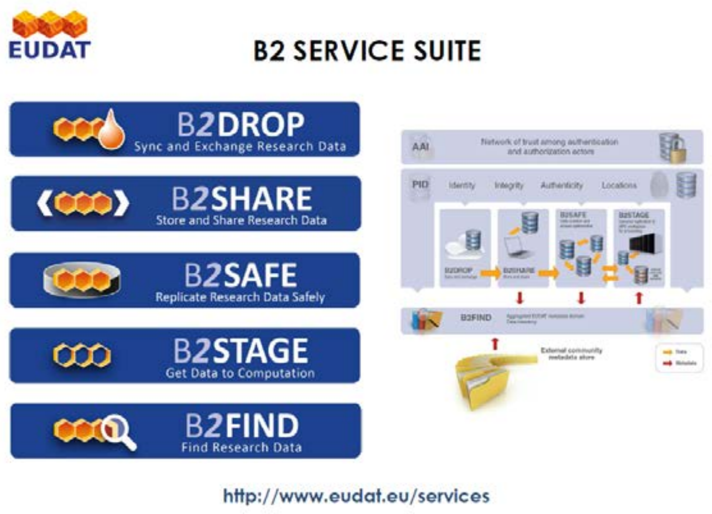

EUDAT
B2 SERVICE SUITE
B2DROP
Sync and Exchange Research Data
B2SHARE
Store and Share Research Data
B2SAFE
Replicate Research Data Safely
B2STAGE
Get Data to Computation
B2FIND
Find Research Data
AAI
Network of trust among authentication and authorization actors
PID
Identity
Integrity
Authenticity
Locations
B2SAFE
B2STAGE
B2DROP
B2SHARE
B2FIND
External community metadata store
http://www.eudat.eu/services

Figure 3.2: Services offered by EUDAT.

EUDAT is a pan-European data initiative which brings together a unique consortium of 26 partners, including research communities, national data and HPC centres, technology providers, and funding agencies from 13 countries. EUDAT's mission is to design, develop, implement and offer "Common Data Services" as they have been introduced in the "Riding the Wave report" to all interested researchers and research communities. A Collaborative Data Infrastructure (CDI) is being planned by many data different initiatives at community, research organisation and cross-border level (disciplines and countries). Common data services must be relevant to several communities and be available at European level and they need to be characterised by a high degree of openness: (i) Open Access should be the default principle; (ii) Independent of specific technologies since these will change frequently and (iii) Flexible to allow new communities to be integrated which is not a trivial requirement given the heterogeneity and fragmentation of the data landscape.

EUDAT thus aims to provide an *integrated solution for finding, sharing, storing, replicating, staging and performing computations with primary and secondary research data*. EUDAT is currently rolling out its first set of data services (see Figure 3.2) which are:

- *B2SHARE:* a "user-friendly, reliable and trustworthy way" for researchers and communities to store and share small-scale research data coming from diverse contexts. This service is open to all researchers and EUDAT is looking for special collaboration with communities to develop customized solutions.
- *B2SAFE:* a "robust, safe and highly-available replication service" allowing community and departmental repositories to replicate and preserve their research data. Different access and deployment options are offered which range from tailored solutions for Fedora and DSpace repository systems via simplified utilization options to a full integration of repositories with the network of EUDAT data nodes.
- *B2STAGE:* a "reliable, efficient, easy-to-use service to ship large amounts of research data" between EUDAT data nodes and workspace areas of high-performance computing systems.

- *B2FIND:* a "simple and user-friendly portal to find research data collections stored in EUDAT data centres and other repositories". B2FIND harvests metadata from diverse sources, maps it, and makes it publically available through its cross-disciplinary catalogue.
- *B2DROP:* a "secure and trusted data exchange service for researchers and scientists" to keep their research data synchronized and up-to-date and to exchange with other researchers. It is not clear why this is better or more sustained than commercial services (like Google drive or Dropbox).

EUDAT has begun work on B2NOTE offering a "Semantic referencing and annotation" service. Semantic annotation can be applied to derived and typical long tail data, in addition to regular raw data created by machines. A typical scenario human-generated data with errors, where scientists will want to annotate the errors and create references to accepted ontologies. Semantic annotation can be seen as a common service that can be applied to processes of data enrichment in many scientific disciplines. Such an annotation module is proposed as a plug-in for EUDAT core services, and as a plug-in for community services. EUDAT's Semantic Annotation Working group has recently established the *European Ontology Network* to share and coordinate expertise and experience in the European ontology community, with a view to re-using existing ontologies and tooling solutions and reducing waste in reproducing ontologies that already exist.

OpenAIRE (https://www.openaire.eu/)

OpenAIRE is an e-Infrastructure to support the implementation of Open Access in Europe. Open Access is a strong theme in H2020, extending to support for Open Data Publishing and Data-backed publishing. Linking the aggregated research publications to the accompanying research and project information, datasets and author information, and providing access to publications, datasets or project information, is specifically called for in H2020 including *data* publication (EINFRA-2-2014 and EINFRA-3-2014). OpenAIRE also offers support services for researchers, coordinators and project managers such as statistics and reporting tools. It relies heavily on a decentralized structure and operates a federated or "Aggregrated Data Infrastructure" approach, drawing data from free-standing national, community and international infrastructures. OpenAIRE has:

- *support structures* for researchers in depositing research publications through a European Helpdesk and the outreach to all European member states through the operation and collaboration of 27 National Open Access Offices.
- *an e-infrastructure* for handling peer-reviewed articles as well as other important forms of publications (pre-prints or conference publications), through a portal that is the *gateway* to user-level services, including access (search and browse) to scientific publications and other value-added functionality (post authoring tools, monitoring tools through analysis of document and usage statistics);
- specific work with subject communities to *explore* the requirements, practices, incentives, workflows, data models, and technologies to deposit, access, and *combine research datasets* of various forms in combination with research publications.

BioMedBridges (http://www.biomedbridges.eu/)

BioMedBridges is a joint effort of twelve biomedical sciences research infrastructures on the ESFRI roadmap. Together, the project partners planned to develop the shared e-infrastructure—the technical bridges—to allow data integration in the biological, medical, translational and clinical domains and thus strengthen biomedical resources in Europe. It aims to bridge data:

- across different spatial scales, from molecules through cells and organs to humans and the environment

- between different species, from bacteria through model organisms to humans
- between different technologies and the heterogeneous data they generate, from the nanotechnology of sequencing through the spectroscopy of cellular and whole organism imaging to the powerful synchrotrons for structure determination
- across different research communities, from basic molecular biologists to clinicians and environmental researchers.

Notable outcomes of BioMedBridges are with regard to interoperability: the EBI Linked Data Platform; recommendations for rules on standardising identifiers[8] and infrastructure for identifiers (identifiers.org), and a tool registry that forms the basis of the ELIXIR tools registry. BioMedBridges has a follow-on project, CORBEL, which starts in Sept 2015 and includes ISBE.

VPH Institute (http://www.vph-institute.org/)

The VPH is an international non-profit organisation whose mission is to ensure that the Virtual Physiological Human is fully realised, universally adopted, and effectively used both in research and clinic. It has members throughout Europe. Systems Biology embraces modelling across scales: physiological modelling as well as molecular pathway modelling.

ERANets

The EU ERA-NET scheme aims to develop and strengthen the coordination of national and regional research programmes. Under the ERA-NET scheme, national and regional authorities identify research programmes they wish to coordinate or open up mutually. The participants in these actions are programme 'owners' (typically ministries or regional authorities defining research programmes) or programme 'managers' (such as research councils or other research funding agencies managing research programmes). Key ERANets relevant to ISBE are:

- *SysMO "Systems Biology of Microorganisms":* to record and describe the dynamic molecular processes going on in unicellular microorganisms in a comprehensive way and to present these processes in the form of computerized mathematical models. SysMO sponsored the development of the SEEK4Science Data and Model management platform (http://www.seek4science.org) and associated tools, as a special associated data management project. This forms the basis of the FAIRDOM Initiative (http://www.fair-dom.org).
- *EraSysAPP Systems Biology Applications* to coordinate and enhance research opportunities in the emerging scientific field of Systems Biology. It predominantly aims at funding transnational Applied Systems Biology research. EraSysAPP sponsors the FAIRDOM Initiative as an explicitly funded data and model management action for the funded projects.
- *EraSynBio* promotes the robust development of Synthetic Biology and to structure and coordinate national efforts and funding programs. No specific data management action is sponsored.
- *CASyM* Coordinating Action Systems Medicine is a multidisciplinary European consortium that joined forces to develop an implementation strategy (road map) for Systems Medicine. It aims to identify areas where a systems approach will address clinical questions and solve clinical problems. No specific data management action is sponsored.

The Innovative Medicines Initiative

The Innovative Medicines Initiative (IMI) is Europe's largest public-private initiative aiming to speed up the development of better and safer medicines for patients. IMI supports collaborative research projects and builds networks of industrial and academic experts in order to boost pharmaceutical innovation in

[8] 10 Simple rules for design, provision, and reuse of persistent identifiers for life science data doi:10.5281/zenodo.18003

Europe. IMI is a joint undertaking between the European Union and the pharmaceutical industry association EFPIA. IMI was launched in 2008. Of its 50+ projects, some focus on specific health issues such as neurological conditions (Alzheimer's disease, schizophrenia, depression, chronic pain, and autism), diabetes, lung disease, oncology, inflammation & infection, tuberculosis, and obesity. Others focus on broader challenges in drug development like drug and vaccine safety, knowledge management, the sustainability of chemical drug production, the use of stem cells for drug discovery, drug behaviour in the body, the creation of a European platform to discover novel medicines, and antimicrobial resistance. In addition to research projects, IMI supports education and training projects.

Notable data/knowledge management projects include:

- *The Open PHACTS Discovery Platform* (https://www.openphacts.org/): brings together pharmacological data resources in an integrated, interoperable infrastructure using a Linked Data platform.
- *eTRIKS* (http://www.etriks.org/): provides a platform and services for data staging, exploration and translational research, focusing on Driving the adoption of a common open source platform; Promoting multi-study data harmonisation; Developing best practice guidelines and resources for the re-use of research data; and Providing advice and support for translational research projects. It develops the tranSMART platform (http://transmartfoundation.org/), an increasingly widely used warehouse for clinical data.

Cross EU-RI Synergies

ISBE WP5 outlines the synergies anticipated with other EU RIs, we have replicated these in Table 3.4.

Table 3.4. Summary of synergies anticipated with other EU RIs. Services expected to be operated by ISBE are shown in green. Overlaps between other ESFRIs are shown in red.

	Data Generation / Technologies	**Data Stewardship Data Discovery / Access/ Management / Curation / Preservation**	**Data Integration Analysis / Modelling**
BioMedBridges		Data access Data standardization, harmonization and interoperability between RIs; Registry of resources and software and services	Data integration between RIs
BBMRI	Systems biology technologies	Management of biological data and resources Data access,	Data integration and modelling
EuroBioImaging	High-throughput imaging for systems approaches	Data storage and integration	Streamlining data generation-integration processes for SB modelling purposes
EATRIS	Systems approaches for translational research	Storage of data and models	Modelling for compound/drug selection

ELIXIR		Data access to, and stewardship of, "kite-marked" data resources and services;
		Data standardization, harmonization and interoperability;
		Tools interoperability;
		Data storage for ELIXIR data resources, high-capacity computing facilities;
			Tools interoperability;
			Tools interoperability training
			Kite-marked tools;
			Data mining and analytics services.

		Secure handling and access to data. Data carpentry training.	
ECRIN	-omics approaches for translational research	Management of data and models	Prediction of drug safety/toxicity and efficiency of treatment
EU Open Screen	Combining high-throughput compound screening facilities with Systems biology technologies	Access, storage and integration of screening-, -omics-, and modelling data/ models	Integration of screening data for systems biology modelling
EMBRC	-omics and high-throughput sequencing of uncharacterized organisms, natural products ect.	Data access, storage and integration	Coupling of physical, chemical and biological metadata to SB analysis of communities, ecosystems, and processes
ERINHA	-omics analysis of patient data and host pathogens	Data access, storage and integration	Modelling for ID of compounds against high-risk pathogens
Infrafrontier	High-throughput systems analysis of mouse phenotypes	Storage and integration of phenotypic data (together with BioMedBridges)	Systems-wide analysis of the mouse, phenotypic data integration and modelling
Instruct	Combining 3D structure technologies and Systems biology facilities	Management and integration of structural data and models	Streamlining the integration of structural data into systems wide modelling analysis, e.g. for the prediction of compound-target interactions
MIRRI	-omics, high-throughput sequencing of microorganisms	Data mining, data access, and integration, SOPs,	Integrated analysis of uncharacterized organisms/bio. Material
EUDAT		Research data discovery (B2FIND), Data replication services (B2SAFE), Storage and sharing (B2SHARE), Moving data to computation (B2STAGE), Semantic annotation.	Moving data to computation (B2STAGE)
OpenAIRE		Open data access.	

3.2.2 National Infrastructures

ISBE will work within the context of National infrastructures as well as European. Some examples follow.

de.NBI (http://www.denbi.de/)

The 'German Network for Bioinformatics Infrastructure' provides comprehensive first-class bioinformatics services to users in life sciences research, industry and medicine. The de.NBI program coordinates bioinformatics training and education in Germany and the cooperation of the German bioinformatics community with international bioinformatics network structures.

Dutch Techcentre for Life Sciences (DTL) (http://www.dtls.nl)

DTL focuses on the great potential of high-end technologies in pioneering life science research, and on the skills and solutions to professionally use computers to deal with the ever-growing data streams in research. The mission of DTL is to establish an expert network and a federated research infrastructure that enables integrated life science research in academia and industry in a cost-effective manner. DTL offers the platform where DTL Partners can bundle their capacities and make use of each other's strengths. DTL takes responsibility for data management of projects funded by NOW (5% "tax" is levied on each project) and is a driver of the "FAIR" data paradigm.

UK Multi-scale Biology Network (http://www.multiscalebiology.org.uk/)
The MSB Network aims to facilitate the exploitation of areas of synergy between researchers across the UK in the broad area of multi-scale biology (MSB); raise awareness of the benefits of MSB; identify areas that are most ready for MSB and horizon scan for emerging areas; and explore opportunities for MSB collaboration within Europe and further afield.

3.2.3 Non-European Research Infrastructure

Non-EU RIs are also relevant to for understanding ISBEs place as in infrastructure world-wide. ISBE must compliment the work of other infrastructures on the global scale in order to be relevant to the Systems Biology community.

ISBE must clarify its expectations, responsibilities, approaches, resources and services, both technical and social, with key international RIs and resources, for example NIH BD2K, iPlant Collaborative, KBase, PMR and KEGG.

KBase (https://kbase.us/)
KBase is a US initiative that aims to combine a broad base of knowledge, with easy to use analysis tools in order to generate a platform for generation, and sharing of hypothesis within Systems Biology. It contains an open development environment where users can develop new tools that be accessed by the wider community. Its aim is to make data analysis more efficient by removing the need to install and learn a multitude of methods, to run on one data set, or difficulties in running one tool on multiple datasets. It looks to merge different datasets into a single integrated data model. This data presentation is similar to what ISBE would intend to do. It is not a data repository, but relies on existing databases. This is what we would expect one component of ISBE to do, probably to datasets managed by ELIXIR. They do not control this data. Aim is to prevent replication of data.
KBase has:

- Access to tools for annotation and simulation of heterogenous datasets.
- Access to data sets for a wide range of organism types held in diverse databases. It then interfaces them as a single data model.
- Community sharing of tools and data, with a view to standardisation and interoperability.
- Training material for resources they have.

KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)

KEGG is a Japanese initiative that collects and integrates molecular level information such as genes, proteins, and metabolites, in one place, to facilitate the high-level understanding of organism behaviour. Its display of information is primarily visual mapping, with access to descriptions and functional linking of genes through to proteins through to small molecule behaviours (Figure 3.3).

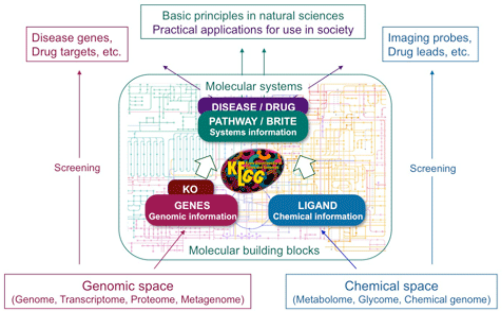

Basic principles in natural sciences
Practical applications for use in society
Disease genes, Drug targets, etc.
Imaging probes, Drug leads, etc.
Molecular systems
DISEASE / DRUG
PATHWAY / BRITE
Systems information
Screening
Screening
KO
GENES
Genomic information
LIGAND
Chemical information
Molecular building blocks
Genomic space
(Genome, Transcriptome, Proteome, Metagenome)
Chemical space
(Metabolome, Glycome, Chemical genome)

Figure 3.3: KEGG services.

It allows access to information:

- Searching and mapping of pathways to allow analysis and reconstruction.
- Hierarchical organisation of pathways and object components.
- Searching of single components (modules) and associated information (gene alias, EC number etc.)

Physiome Model Repository (https://www.cellml.org/tools/pmr)

Physiome Model Repository is a New Zealand based content management system for models. It allows models to be stored in any format, and the models can be modified by users with a full version history tracking the changes. It supports running of CellML models but not other formats, they are working to introduce this. The aim is to have a community repository for all systems biology models. Annotations are encouraged so that users can re-use models correctly.

- Facilitate model transfer between researchers without a reliance on a central repository.
- Maintenance of detailed editing history.
- Maintain privacy on models where necessary.
- Embed workspaces so that models can be used and reused successfully, and can be developed in a modular way.

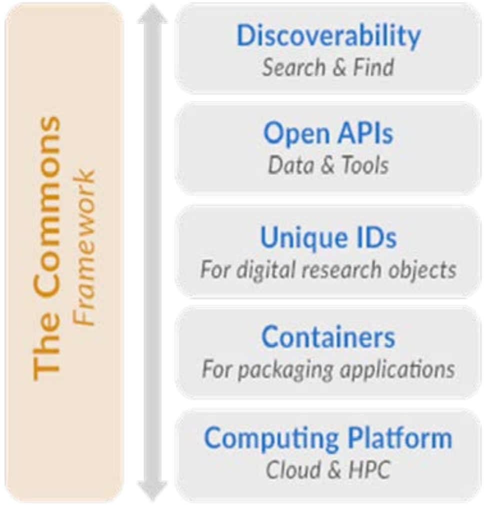


Figure 3.4: NIH Commons.

NIH BD2K **(https://datascience.nih.gov/bd2k)**

NIH BD2K is a major US initiative aimed at making big data from the biomedical community more standardised and accessible to the majority. This move to big data management is a natural progression from previous database initiatives from NIH. The data in these databases are growing in size and complexity, and therefore their management and handling are becoming difficult for traditional silos. The initiative aims to develop capacities similar to that which we see from smaller more manageable datasets currently:

-Discoverability, management, curation, and meaningful re-use a priority for all big data.

-Tool development for processing, analysis, integration, and visualization.

- Development of researchers skilled in big data analysis.

NIH BD2K has funded 12 Centres, including CEDAR (Center for Expanded Data Annotation and Retrieval) to facilitate automated annotation of data with high quality metadata by generating community-based metadata standards and a metadata repository for training learning algorithms to develop metadata templates. These templates will initially be evaluated, validated, and adapted with the NIAID ImmPort multi-assay data repository and other data repositories.

The BD2K is based on the notion of a Research Object Commons - an ISBE asset recommendation is the establishment of a Sys Bio Asset Commons (see Figure 3.4). BD2K has funded the bioCADDIE project (Biological and HealthCare Data Discovery and Indexing Ecosystem http://www.biocaddie.org) to develop a prototype Data Discovery Index Infrastructure that will enable finding, accessing and citing biomedical big data. bioCADDIE has a Community Engagement mandate that seeks to work with the broader biomedical community to better identify data, and other digital objects, so that they may find shared data in ways that allow for extracting maximal knowledge. ELIXIR and ISBE representatives serve on the bioCADDIE working groups for metadata and identifiers. A Software Discovery Index is expected to be funded.

iPlant Collaborative (http://www.iplantcollaborative.org/)

iPlant Collaborative was established by the U.S. National Science Foundation (NSF) in 2008 to develop the national cyberinfrastructure for data-intensive biology driven by high-throughput sequencing, phenotypic and environmental datasets, provide powerful extensible platforms for data storage, bioinformatics, image analyses, cloud services, APIs, and more, and make broadly applicable cyberinfrastructure resources available across the life science disciplines (e.g., plants, animals, and microbes). iPlant Collaborative's infrastructure is gaining widespread acceptance in the Plant community.

- *Atmosphere* is iPlant's cloud service that lets you launch your own isolated virtual working environment and software. You use Atmosphere's web interface to get computing resources, such as iPlant-provided software suites, in a virtual machine (VM). If you want to publish your own software suites, you can create your own work environment in Atmosphere and run your own software for community use. You can use Atmosphere as a gateway to access the iPlant's core infrastructure resources, including high performance computing (HPC), grid computing environments, and the iPlant Data Store. Atmosphere provides easy access to three levels of service to match your needs and compute skills: Infrastructure as a Service (IaaS), with advanced APIs; Platform as a Service (PaaS), for developing and deploying software to the science community; and Software as a Service (SaaS), with preconfigured, frequently used analysis routines, relevant algorithms, and datasets as a software service in an on-demand environment designed to accommodate computationally and data-intense bioinformatics tasks.

- *The Discovery Environment* (DE) is the primary web interface and platform to access the powerful computing, storage, and analysis application resources of iPlant's cyberinfrastructure. The DE is designed to facilitate data exploration and scientific discovery by providing: analytical tools that can be used individually or in workflows; seamless access to the iPlant Data Store; flexibility to run tools on local or high-performance computing nodes, as appropriate; collaboration tools for sharing data, workflows, analysis results, and data visualizations with collaborators or with the community at large. The DE is integrated with iPlant's data management system and compute resources, so researchers can access tools and data with scalability. In the future, the DE will employ provenance tracking of both primary and derived files to track and reproduce experiments.

- The *iPlant Data Store* is where store all the data related to your research. The Data Store is cloud-based, provides reliable and redundant storage, and is the central repository from which data is accessed by all of iPlant's technologies, including the Discovery Environment, Atmosphere, or

> application programming interfaces. With a unified place to store and analyze your data, sharing data between your personal computer and iPlant's tools, among your collaborators, and even with other web-based systems is easy. The iPlant Data Store is built using iRODS, a well-established separate cyberinfrastructure project for managing all kinds of data, including very large datasets similar to those generated by high-throughput technologies, such as NextGen Sequencing. It supports different interfaces, from web services to mountable file systems to high-speed command line transfers; permissions for sharing or privacy; fast upload times and partial transfers and automatic storage allocation of 100GB for each user account, with increase up to 1TB or more possible.

ISBE researchers working in Plants may well use iPlant Collaborative, and many of its features echo those to be provided by the recommended ISBE Sys Bio Asset Commons.

Galaxy (https://galaxyproject.org/)
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses. It is arguably the most widely used open source platform for genomics and increasingly post-genomics analysis. It links command line tools. Many ISBE researchers will use it. It does not explicitly support models.

3.2.4 Funding agencies

Funding agencies across the world are increasingly adopting policies with regard to open access for research articles, and increasingly the open availability of data (and in future software). Drivers vary (as seen in Table 3.1), but are in part focused on economic and ethical arguments for the wider availability of publicly funded research for reuse and public confidence in the reproducibility of investments (for example, a US study claims $28 billion a year spent on irreproducible biomedical research[9]). The EU has established open access policies in line with the EU objectives on Responsible Research & Innovation. Facilitating open access to research data is one of the priorities flagged in the European Commission’s Horizon 2020 funding framework[10]. Open data is a principle endorsed by an increasing number of countries and international organizations, and several OECD countries are adopting policies to promote open research data, for example by requiring the archiving of research outputs in a digital formats, also requiring the development of international open standards.

In the UK The Expert Advisory Group on Data Access (est. by 4 UK funders, Wellcome Trust, Cancer Research UK, the Economic and Social Research Council, and the Medical Research Council) provides strategic advice to these funders on the emerging scientific, legal and ethical issues associated with data access for human genetics research and cohort studies – identifying best practice and encouraging harmonisation in governance and decision-making. In their recommendations to funders[11] a key point is that data-access plans should be integral to the grant-application process, that the review process should advise on this and the data-access plan should be an integral, auditable part of the funded grant. We have summarised UK funder data policies in Table 3.5.

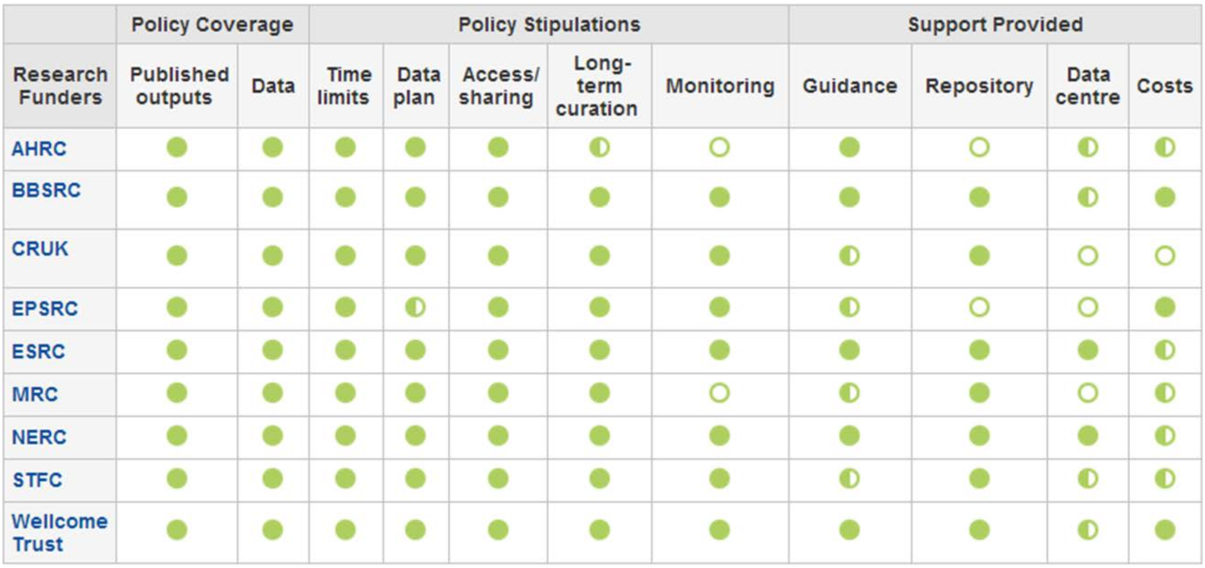
[9] http://news.sciencemag.org/biology/2015/06/study-claims-28-billion-year-spent-irreproducible-biomedical-research?utm_campaign=email-news-latest&utm_src=email

[10] European Commission: “Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020”; G8 Science Ministers Statement, 13 June 2013, https://www.gov.uk/government/news/g8-science-ministers-statement

[11] http://www.wellcome.ac.uk/EAGDA

In the UK on 1 May 2015 the EPSRC funding council implemented its compliance expectations[12] concerning the management and provision of access to EPSRC-funded research data. It had a significant impact on UK research institutes. The expectations present seven core principles which align with the core RCUK principles on data sharing. Two of the principles are of particular importance: firstly, that publicly funded research data should generally be made as widely and freely available as possible in a timely and responsible manner; and, secondly, that the research process should not be damaged by the inappropriate release of such data.

Table 3.5: An overview of UK Funder Data Policies[13] (from http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies).



	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	○	●	○	◐	◐
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
CRUK	●	●	●	●	●	●	●	◐	●	○	○
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	◐
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

Table 3.6: Terminology for table 3.5.

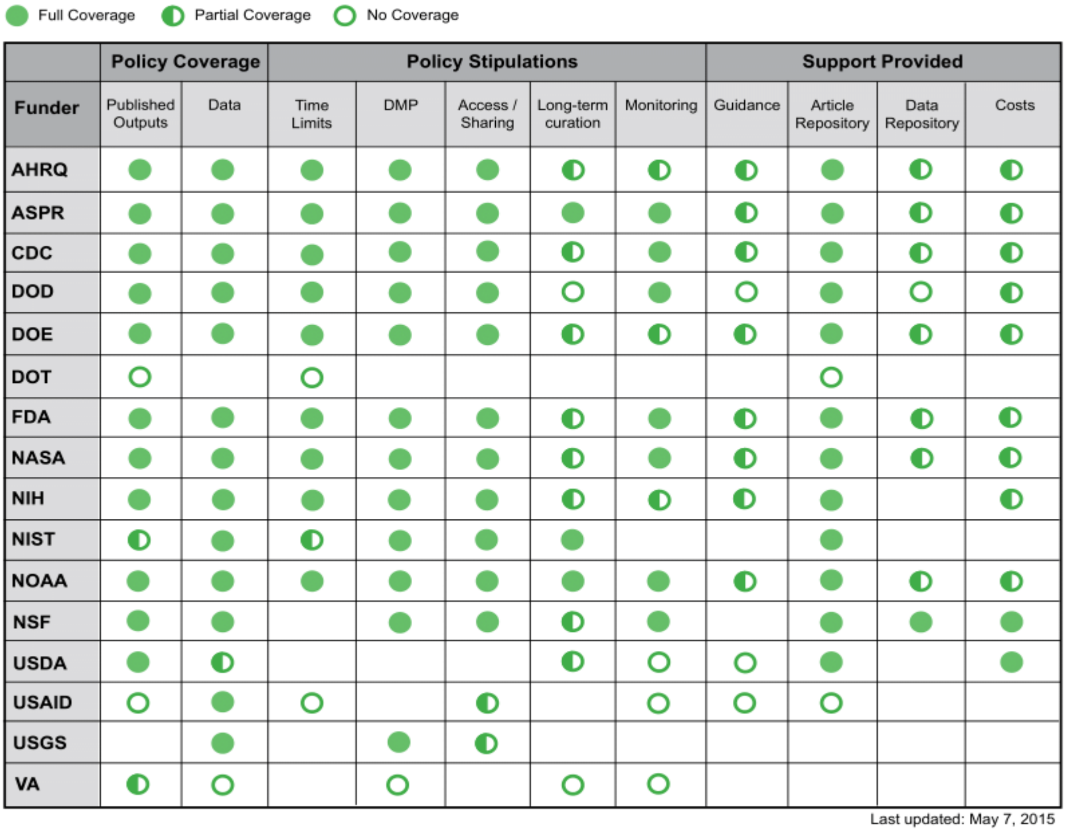
Published outputs	a policy on published outputs e.g. journal articles and conference papers
Data	a datasets policy or statement on access to and maintenance of electronic resources
Time limits	set timeframes for making content accessible or preserving research outputs
Data plan	requirement to consider data creation, management or sharing in the grant application
Access/sharing	promotion of OA journals, deposit in repositories, data sharing or reuse
Long-term curation	stipulations on long-term maintenance and preservation of research outputs
Monitoring	whether compliance is monitored or action taken such as withholding funds
Guidance	provision of FAQs, best practice guides, toolkits, and support staff
Repository	provision of a repository to make published research outputs accessible
Data centre	provision of a data centre to curate unpublished electronic resources or data
Costs	a willingness to meet publication fees and data management / sharing costs

[12] https://www.epsrc.ac.uk/about/standards/researchdata/

[13] : http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies#sthash.Ud09h6KJ.dpuf

Table 3.7 summarises a crowd-sourced Google Spreadsheet that consolidates guidelines from US funding agencies stemming from the Whitehouse's Office of Science and Technology Policy's (OSTP) 2013 memo to expand public access to the results of federally funded research. The chart provides an overview of each agency's compliance with policies that are intended to "[open] government data resources" by working towards public access for all research outputs supported by federal funding. (Process Toward Opening Government Data Resources. The White House, 16 Aug. 2013. Web.)

Table 3.7: Overview of USA Funder Data Policies[14]



● Full Coverage ◐ Partial Coverage ○ No Coverage

Funder	Policy Coverage		Policy Stipulations					Support Provided			
	Published Outputs	Data	Time Limits	DMP	Access / Sharing	Long-term curation	Monitoring	Guidance	Article Repository	Data Repository	Costs
AHRQ	●	●	●	●	●	◐	◐	◐	●	◐	◐
ASPR	●	●	●	●	●	●	●	◐	●	◐	◐
CDC	●	●	●	●	●	◐	●	◐	●	◐	◐
DOD	●	●	●	●	●	○	●	○	●	○	◐
DOE	●	●	●	●	●	◐	◐	◐	●	◐	◐
DOT	○		○						○		
FDA	●	●	●	●	●	◐	●	◐	●	◐	◐
NASA	●	●	●	●	●	◐	●	◐	●	◐	◐
NIH	●	●	●	●	●	◐	◐	◐	●		◐
NIST	◐	●	◐	●	●	●			●		
NOAA	●	●	●	●	●	●	●	◐	●	◐	◐
NSF	●	●		●	●	◐	●		●	●	●
USDA	●	◐				◐	○	○	●		●
USAID	○	●	○		◐		○	○	○		
USGS		●		●	◐						
VA	◐	○		○		○	○				

Last updated: May 7, 2015

3.2.5 Publishers

In the face of concerns about the irreproducibility of results (in no small part due to their own editorial policies) journals are starting to finally mobilise to demand that data and software are available to back the claims made in articles (see Nature Irreproducible Research supplement[15]). Nature Biotechnology announced a plan[16] ask peer reviewers to assess the availability of documentation and algorithms used in computational analyses, not just the description of the work and test complex code using services such as Docker (software for shareable representations of computing environments). Nature and other

[14] from http://figshare.com/articles/Overview_of_OSTP_Responses/1367165 by Dan Valen (figshare)

[15] http://www.nature.com/news/reproducibility-1.17552

[16] Nature Biotechnol. 33, 319; 2015

journals introduced have produced guidelines for reporting preclinical research[17], and introducing editorial measures such as reporting guideline checklists[18], and services such as NPG's Scientific Data. Nevertheless, many journals have weak data or software policies and do not stick to them when they do[19].

Twenty-five journals in which Systems Biology models are published regularly were checked on model format guidelines in the instructions for authors. Eighteen of the journals had no specific information for the mathematical model description format or submission. Some of the journals have a very generic formulation: (a) Nature publishing group: "Nature Journals consider it best practice to release custom computer code in a way that allows readers to repeat the published results." (b) Molecular Systems Biology, EMBO ".... computational models should be deposited in one of the relevant public databases prior to submission (provided private access is available at the database) and authors should include accession codes in the Materials & Methods section." Only five journals (FEBS, Gene Regulation and Systems Biology, IET Systems Biology, Metabolomics, Microbiology) were specific in model description format (SBML) and model database to which the models should be submitted (JWS Online and Biomodels).

Data and software publishing broadly falls into several categories:

1. *Data or Software articles* Data Journals[20] apply the traditional articles model to data through the proxy of a narrative description of the data. Examples include the NPG Scientific Data, Elsevier "data in brief", F1000, GigaScience and the Data Journal. The data and article are actively reviewed, curated, formatted, indexed, given a DOI and publicly available to all upon publication. Software articles are similarly emerging.
2. *Linking into specialist public archives* Elsevier[21] encourages authors to connect articles with external databases, giving readers access to relevant databases using standardised identifier formats (e.g. e.g., TAIR: AT1G01020). This encourages authors to deposit their data in sustained public archives. NPG's Scientific Data[22] similarly recommend public repositories and go further with a (fee-based) curation and support service.
3. *Supplementary partner repositories* such as PLoS arrangements with the figshare and github so that authors publishing in PLOS journals host their data on figshare and their software in github. This has the advantage of providing specialist data/software publishing services (citation tracking, DOI assignment, version management etc) for the journal. In Systems Biology, JWS Online and Biomodels act as a supplementary partner repository for models for FEBS.
4. *Supplementary materials* whereby data (typically CSV) are supplementary files in the journal's repository without special support and without independent DOIs or tracking.
5. *Author sites* whereby data is available from links on investigator web sites, their institutional repository or an investigator/project data repository. Sustainability is a significant issue as web sites are not maintained and URLs decay and projects run out of funds.

[17] http://www.nih.gov/about/reporting-preclinical-research.htm

[18] http://www.nature.com/authors/policies/checklist.pdf

[19] Stodden V, Guo P, Ma Z (2013) Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS ONE 8(6): e67111. doi:10.1371/journal.pone.0067111

[20] Candela et al Data Journals: A Survey, J of Association for Information Science and Technology, 2014. DOI: 10.1002/asi.23358

[21] See http://www.elsevier.com/databaselinking for more information and a full list of supported databases.

[22] http://www.nature.com/sdata/

ISBE Asset management partnerships with publishers should encourage and enable 1-3 and discourage 4-5.

All data publishing, bar a few notable exceptions, follows a "fossil snapshot" approach where published data is fixed. An exception is F1000Research Living Figures[23]. It is anticipated, however, that future publishing will migrate to a "release based" model reflecting the evolutionary nature of science[24],[25]. A release based approach demands a systematic management and standardisation of identification versioning, version-based citation and tracking.

The current scholarly communication landscape is in churn with a plethora of platforms emerging for discovery, analysis, writing, publication, outreach and assessment (see Figure 3.5, Boseman and Kramer 2015[26]). Notable new players include data citation services such as DataCite, and bibliometric service providers like ImpactStory, Altmetrics, and the Thomson Data Citation Index. Mozilla Science Labs have pioneered computer code as a citable research object[27] in partnership with Github and figshare.

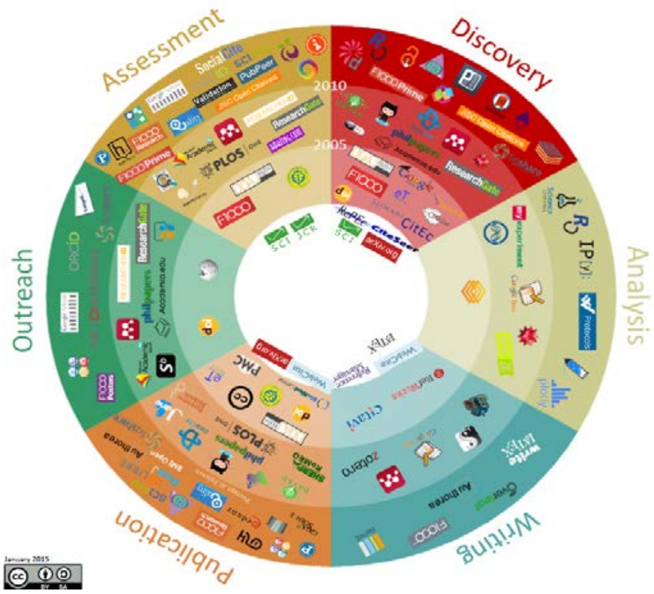


Figure 3.5: The scholarly communication landscape.

The Research Data Alliance Publishing Data Interest Group aims to bring together all stakeholders involved in publishing research data including researchers, discipline specific and institutional data repositories, academic publishers, funders and service providers. ISBE should engage with this group, which aims to:

- Establish the workflows for publishing data: the role of QA/QC, peer-review and certification; implications for the editorial of data and articles and journal editors, publishers, peer-reviewers and the costs/benefits of supplementary materials compared to published data related to an article.
- Establish data publishing services as part of scholarly publishing: identifying the services needed to embed data publications into the current framework of scholarly publishing (data centres / science publishers / bibliometric service providers) and the commensurate organizational and technical requirements and standards for content and interoperability.

[23] http://www.nature.com/news/living-figures-make-their-debut-1.17382

[24] Evolving manuscripts': the future of scientific communication? http://www.timeshighereducation.co.uk/news/evolving-manuscripts-the-future-of-scientific-communication/2020200.article

[25] Schopf, Treating data like software: a case for production quality data. JCDL 2012: 153-156

[26] 101 Innovations in Scholarly Communication - the Changing Research Workflow, Boseman and Kramer, 2015, http://figshare.com/articles/101_Innovations_in_Scholarly_Communication_the_Changing_Research_Workflow/1286826

[27] https://mozillascience.github.io/code-research-object/

- Promote and establish the data publication concept among science publishers and bibliometric service providers.

3.2.6 Industry

The WPX Industry survey[28] revealed that 63% lacked knowledge of existing resources. Table 3.8 summaries outcomes with respect to asset management.

Table 3.8: Summary of WPX Industry survey.

Activities undertaken by businesses	Challenges identified by businesses	Services anticipated from ISBE
91% Data mining &integration 77% Modelling and simulations 77% Data processing & analysis 70% Tools development 56% Curation of systems 42% Experimentation/data generation	84% Standardization of models, data 56% Stewardship, curation and reusability of data 42% Access to relevant education &disciplines 42% Access to funding 35% Access to modelling and data 14% Competition with US and Asia	74% standardize and ensure interoperability of data 59% provide access to resources: tools, data, maps, models and standards, i.e. human biological data. 46% provide stewardship and curation of model-compliant data and models making results re-usable 43% provide training 41% develop modelling of biological systems based on integration of diverse data sets (by big pharma)

Organisations using a Systems Biology approach (84% of respondents) envisage, in the next 10 years, the standardisation of models, data and tools as major challenges. This refers to establishing common standards when developing models that can be implemented across the various sectors. Stewardship, curation and reusability were also important as it allows data from experimentation not to being in formats that are later out of date and inaccessible.

Industry suggested ISBE services include:

- **Open access to data**: accessing and coordination of existing public libraries. On-demand and cost-free access to data not affected by IP issues.
- **Access to models:** making available systems biology models for oncology as this area could lead to a wider variety of drug discovery.
- **Data provision**: creation and maintenance of i.e. compound database, experimental database, etc.
- **Data standardization:** file formats, metadata, etc.
- **Data quality** Higher quality of data, consistently recorded
- **Develop of a user friendly interface** to existing data repositories
- **Access to resources** to identify available resources and make them visible
- **Stewardship**: manage data, set up rules

[28] From the ISBE industry breakfast.

- **Developing predictive models** and making them available. Knowledge-driven modelling versus data-driven modelling e.g. large gap in chem-informatics knowledge comparing with bioinformatics knowledge.
- **Database to host results from EU funded projects**: ISBE provides sustainability and continuity – including failures and successes
- **Groups according to the disease**: lung cancer, diabetes, etc. to facilitate and coordinate knowledge in this area.

In addition SMEs saw ISBE as

- **A marketplace of expertise** to link industry, academia and public institutions.
- **Providing scientific advice**: how to approach a new research in systems biology before starting the project. Best practices (for asset management) and learning from failures
- **Support to collaborate in EU research**: information on EU programs, projects and funding.
- **Networking –** improve communications between industry, academia and researchers

Commercial organizations see the ISBE infrastructure as a possible source of research acceleration. By providing standardization and guidance to users on how to use systems biology and the infrastructure, it is believed that time and money will be saved. In addition, providing common models (understood as generally accepted models) as a starting point would be useful particularly for SMEs.

From the asset management perspective ISBE must have a clear business offer of services, differentiated from and synergistic with, other RIs (e.g. ELIXIR).

4 Asset Management Capability Framework

The Asset Management Capability Framework (Figure 4.1) is a tool to: profile the current readiness / capability of ISBE; highlight priority areas for change and investment; and develop roadmaps. This Framework will serve as a systematic device for planning the Interim Phase of ISBE.

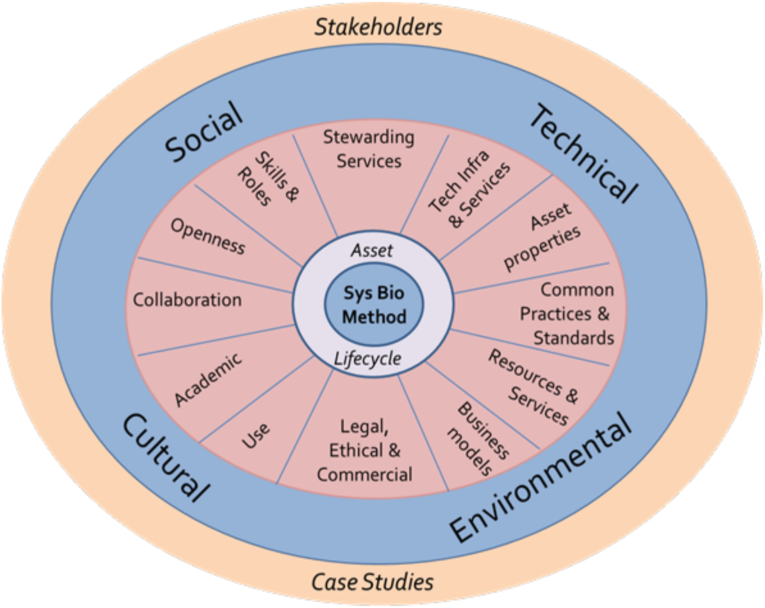


Figure 4.1: The Asset Management Capability Framework.

The Framework is inspired by the Community Capability Model Framework (CCMF)[29] developed by University of Bath, and Microsoft Research to assist institutions, research funders and researchers in analysing and planning the capability of their communities to perform data-intensive research. The CCMF was designed for institutions and research groups. Therefore, we extended it to include the influence of users/sector stakeholders and their case studies, and to reflect Systems Biology research method and outcomes (as sketched in Section 2) and the lifecycle of Sys Bio asset management. We added capabilities to support ISBE as an international infrastructure providing resources and services. Twelve capabilities are broadly arranged into four implementation aspects - technical, social, cultural and environment – each influenced by the others. We briefly introduce them here:

- **Technical aspects include**: how data, models and SOPs should be managed and exchanged within ISBE, and between ISBE and external resources; which formats, identifiers, standards and ontologies should be used, created and maintained for ISBE, and pathways to their adoption; and how interoperability between data and model resources many be achieved.

- **Social aspects include**: how can compliance to the standards recommended by ISBE be encouraged or mandated; how can annotation and standardisation be made more straightforward and rewarding, and less time consuming, for scientists; how data, model and SOP planning and management can become embedded in Systems Biology practice and publishing; and how practices can lead to greater collaboration and openness for the research results of publicly funded research.

[29] https://communitymodel.sharepoint.com/Pages/default.aspx

- **Cultural aspects include**: how existing and new Systems Biologists in data and model management can be educated with respect to data, model and SOP stewardship; how other stakeholders such as funders, librarians and publishers should engaged in the importance of data and model management; how to drive change in the recognition of data, models and SOPs as first class, citable and creditable research outcomes; and how to establish career paths for data and model stewards.

- **Environment aspects include:** how the community should select of the specific public resources and services to be ingested and sustained in the ISBE infrastructure; how to establish partnerships with other RIs such as ELIXIR; how to develop and implement business models for resources and services; and how to develop policies, and responses to ethical, legal, and commercial concerns.

Each capability factor has a series of community characteristics that are relevant for determining the capability or readiness of ISBE's users, stakeholders and Centres to steward the community's assets, sketched in Table 4.1. **The FAIR principle pervades these capabilities.**

Table 4.1. Capability factors

(1) Digital Asset properties • *Data* • *Models* • *SOPs* • *Samples*	**(2) FAIR Resources and services "fabric"** • *Specialist public archives and knowledge bases* • *Projects' Commons* • *Discovery catalogues* • *Metadata registries and services* • *Tools* • *Generic cloud repositories and services*
(3) Common practices and standards • *Standard Operating Procedures* • *Data/model/SOP formats* • *Research Object level cross- and multi-asset packaging* • *Investigation Study and Assay/Analysis* • *Data collection methods* • *Processing workflows* • *Data packaging and transfer protocols* • *Data/model/SOP descriptions & checklists* • *Vocabularies, semantics, ontologies* • *Data/model identifiers* • *Stable, documented APIs* • *Data/Model/SOP citation and credit* • *Reproducible models* • *Metadata catalogues*	**(4) Technical infrastructure and services** • *Repository and commons platforms* • *Cataloguing platforms* • *Computational tools and algorithms* • *Tool support for data capture, processing and annotation* • *Tool support for model versioning and annotation* • *Data storage, data transfer* • *Support for curation and preservation* • *Data/model/SOP discovery and access* • *Integration and collaboration platforms* • *Model simulation platforms* • *Visualisations and representations* • *Platforms for citizen science* • *Virtual machines*
(5) Openness / Accessibility • *in the course of research* • *of published literature* • *of data, models, SOPs* • *of method, workflows, scripts* • *of software* • *of software platforms, proprietary platforms (e.g. Matlab)* • *Re-use of existing data, models, SOPs*	**(6) Collaboration / Interaction** • *within the discipline/sector* • *within and across projects and programmes* • *within and across laboratories and institutes* • *across disciplines (notably experimental and modelling)* • *across sectors* *with the public*

Secure/controlled access to sensitive assets (commercial/personal) *Licensing*	
(7) Academic researcher culture • *Entrepreneurship, innovation and risk* • *Reward models for researchers* • *Credit and citation* • *Quality and validation frameworks* • *Re-use of existing data, models, SOPs* • *RSE and Steward careers*	**(8) Use culture** • *Entrepreneurship, innovation and risk* • *Quality and validation frameworks* • *Re-use of existing data, models, SOPs*
(9) Skills, training and roles • *Skill sets* • *Pervasion of training* • *Data roles (curator, steward, manager, producer, consumer, experimentalist, modeller)*	**(10) Stewarding services** • *Technical services* • *Support services*
(11) Legal, ethical and commercial issues • *Legal and regulatory frameworks* • *Management of ethical responsibilities and norms*	**(12) Economic and business models** • *Sustainability/geographic scale/size of funding for infrastructure/services/resources* • *Public–private partnerships* • *Productivity and return on investment*

We now explore the capabilities, starting from the asset landscape.

4.1 The Asset landscape

As described in section 2.2, the relationships between data and models can take a variety of forms. Data can be used for either constructing or validating models, which means that data generated in the laboratory, or mined from the literature or public archives, can be directly fed into models as parameter values. SOPs are related to both data and models, governing the standardised procedures that make multi-experiment investigations consistent, experimental outcomes comparable and assuring data and model quality.

The asset landscape strongly relates to two capabilities:

- the properties of the assets themselves; and
- the "fabric" of resources used to find and manage them publicly and within projects.

4.1.1 Asset properties

SOPs

SOPs are related to both data and models. SOPs and protocols govern the creation of samples, in order to ensure that all subsequent experiments are carried out on standard, comparable samples and downstream experiments and the informatics analyses of the results obtained. SOPs are essential for quality assurance across the data generation and stewardship centres and will assist in the understanding and therefore reuse of data.

SOPs for modelling are still rare. It is not yet common practice in the modelling community, even in large consortia. In ISBE, the development and support of SOPs for different modelling techniques and procedures (for example parameterising a model) will be necessary for the same quality assurance reasons and to allow scientists to understand and reuse models.

SOPs are still largely described as free text. Some efforts have been made to establish templates against Nature Protocols templates, with moderate success.

Data

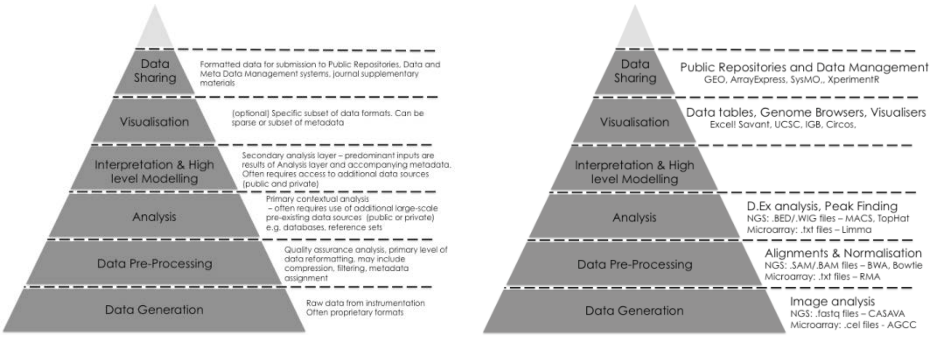

Data Sharing
Formatted data for submission to Public Repositories, Data and Meta Data Management systems, journal supplementary materials
Visualisation
(optional) Specific subset of data formats. Can be sparse or subset of metadata
Interpretation & High level Modelling
Secondary analysis layer – predominant inputs are results of Analysis layer and accompanying metadata. Often requires access to additional data sources (public and private)
Analysis
Primary contextual analysis – often requires use of additional large-scale pre-existing data sources (public or private) e.g. databases, reference sets
Data Pre-Processing
Quality assurance analysis, primary level of data reformatting, may include compression, filtering, metadata assignment
Data Generation
Raw data from instrumentation Often proprietary formats
Data Sharing
Public Repositories and Data Management
GEO, ArrayExpress, SysMO,, XperimentR
Visualisation
Data tables, Genome Browsers, Visualisers
Excel! Savant, UCSC, IGB, Circos,
Interpretation & High level Modelling
Analysis
D.Ex analysis, Peak Finding
NGS: .BED/.WIG files – MACS, TopHat
Microarray: .txt files – Limma
Data Pre-Processing
Alignments & Normalisation
NGS: .SAM/.BAM files – BWA, Bowtie
Microarray: .txt files – RMA
Data Generation
Image analysis
NGS: .fastq files – CASAVA
Microarray: .cel files - AGCC

Figure 4.2: Data property triangle. From data generation through to data sharing.

Raw data and its associated derived data used in systems biology arises from many technological areas. A summary of data properties can be seen in Figure 4.2. The ISBE-wide survey identifies systems biology researchers as already using, or expecting to use data from multiple technologies (D. 9.1) including microarrays, next generation sequencing, proteomics, metabolomics, single cell-based technologies, and imaging. Primary data types associated with different technologies can be seen in Table 4.2. Within each technology area, multiple different types of experimentation are possible and each is characterised by a matrix of possible file types, file sizes and overall data volumes.

A major characteristic of 'raw data' is that volumes, speed of acquisition and file types/formats are subject to technology-driven changes (instrumentation and software changes, relative costs, new methodologies). This variation can be far more rapid than that seen during the later stages of data lifecycle – for instance in the specific data repositories and stages of data-sharing.
The earlier stages of the data lifecycle are frequently characterised by a number of additional factors that influence their storage requirements.

- Proprietary or vendor-specific data formats that require reformatting before further analyses (e.g. MRI vendor-specific formats before conversion to DICOM); may require storage of both versions.
- Many of the file formats used are not particularly predictive of individual size – since many accommodate a whole dataset.
- Temporary storage of multiple interim data files and reformatted versions to allow transfer of data from one to another stage in the analysis pipeline. This frequently transiently increases the total data holdings for an experiment by many multiples of the original raw data volume.
- A requirement for local copies of previously acquired data. These may be private or data from public repositories/databases and can be larger than the primary dataset. Some analysis software requires specifically reformatted or indexed versions, which may themselves require specific versioning and update schedules.

- Since data volumes may be large, and reformatting computationally intensive, primary data formats are most usefully kept local to the original data source and local compute, and network traffic minimised (e.g. between remote sites).

Table 4.2: Primary data file types

Experiment type	Description	Filename	Type/use
Microarray	Affymetrix	cel	Tab-delimited text
	Other microarray data formats	mev, Stanford	Can contain data from single or many chips. tab-delimited text, but different column orders, degree of commenting
	Simple Omnibus Format in Text	SOFT	GEO microarray data exchange format – line based plain text
Next generation sequencing - including genome sequencing, re-sequencing and variant detection, RNA-Seq, ChIP-Seq	Binary alignment	BAM	Compressed (binary) version of SAM
	Sequence alignment/map	SAM	Created by alignment programs
	Defining annotation lines on a reference sequence	BED	For visualising annotations in genome browser
	'wiggle' format for continuous-valued data in a track format, also binary compressed version (BigWIG)	WIG BigWIG	e.g. visualisation of GC percent, probability scores, and transcriptome data on genome sequence
	Contains sequence and quality scores	FASTQ	Fasta format sequence and quality data
	Variant calling format (variant positions in genome)	VCF	Text - Often binary format
	Reference-based compression	CRAM	Tuneable binary format for multiple sequences
	General feature format	GFF	Placing features on a genome (reference) sequence
Medical imaging	Open file format for medical imaging	DICOM	
Confocal microscopy	Tagged image file format (Generic)	TIFF	Information not changed when format created
	Joint Photographers Experts Group image format	jpeg	Uses lossy image compression – different compression ratios available
	Multipage TIFF with OME XML data block	OME-TIFF	Encodes additional metadata
	Proprietary image formats containing microscope-specific metadata	Zeiss LSM Leica LEI	Instrument or software-specific
Super-resolution microscopy	Tagged spot file format	tsf	Binary format for that methods that generate images by locating the position of single fluorescent emitters
Metabolomics - Mass Spectroscopy	Network Common Data Format	netCDF	Machine independent array-oriented binary data format
	ms and ms/ms proteomics data	mznld	open data format for storage and exchange of mass spectroscopy data
	Proprietary examples – Thermo	RAW Baf	

	Bruker, ABI/Agilent	wiff	
Metabolomics - NMR	Self-defining Text Archival and Retrieval format	NMR-STAR	Chemical shift file

We give some illustrations of changing data volumes during the experimentation analysis lifecycle

1: Large-scale genome re-sequencing/variant detection

Sequencing 40 human genomes on the Illumina platform, each to a forty-fold coverage produced the following files at different stages in the analysis pipeline [cite]:

- gzipped fastq files from the sequencer: 772Gb
- bam files: containing reads aligned to human reference genome: 910 Gb
- Vcf format variants: 3.8 Gb

Sequencing 170 human genomes to four-fold coverage on the Illumina platform yields:

- gzipped fastq files: 2039 Gb
- bam files: 3.6 Tb
- Variants (Vcf format): 48Gb

However, multiple intermediate copies of BAM files may be retained for practical reasons until all stage of analysis are complete. For the analyses, indexed versions of the reference human genome are required locally, together with formatted versions of dbSNP, 100Genomes data[30] and Ensembl[31].
Data submitted to the ENA (European Nucleotide Archive Short Read Archive[32] from this experiment included cleaned BAM files and VCF variant calls, in the region of 5 Terabytes of data. Maximum local data volume however was over 30 TB.

2. Transcriptomics Experiments

Here we look at representative files for 1 sample for an RNA-Seq platform and a microarray platform experiment, and the data volumes representing a model experiment studying 2 biological conditions over 4 time points with 3 biological replicates – i.e. a multiplication factor of 24 for each platform type.

RNA-Seq	1 sample*	24 samples (2 conditions x 4 time points x 3replicates)	File type
raw image data	1TB	24TB	
raw data	10GB – 2 x 5GB for paired end run	240GB	fastq.gz
processed data	1.5GB	32GB	.BAM
analysis file	12MB	260MB	.xlsx

*assumes that one sample is one of 4 multiplexed samples in one lane of a HiSeq2000 run (i.e. one of 32 samples).

Microarray	1 sample	24 samples (2 conditions x 4 time points x 3replicates)	File type
raw image data	60MB	1.4GB	.DAT
raw data	15MB (.cel)	360MB	.CEL

[30] http://www.1000genomes.org/

[31] http://www.ensembl.org/Homo_sapiens/Info/Index

[32] http://www.ensembl.org/Homo_sapiens/Info/Index

processed data	0.5 MB (.txt)	12MB	.txt
analysis file	3MB	72MB	.xlsx

3. High throughput Metabolomics - targeted profiling on serum or urine samples

- For NMR, 0.5 to 2 MB per sample assay
- For Mass Spectroscopy (MS), volume is more variable but in region of 7GB per sample assay
- Targeted Mass spectroscopy assays yield less data per sample, in the 100's of MB

MS data collected in proprietary format is reformatted to .netCDF (between 50% and 100% of original data size, dependent on sample) or .mznld (c20% of original size).

An assay may be run as frequently as every 15 mins on all platforms (but turnaround time is method dependent).

A facility consisting of multiple MS instruments is currently able to generate 206GB of raw data per day per instrument, which need to be moved to network storage and backed-up immediately. Feature extraction analysis yields approximately 50MB data for each assay. Raw data are retained in archive for at least 5 years currently.

Secondary data resources

Secondary databases collect content from the literature or from other published primary repositories. They provide detailed, highly curated collections of data for specific research areas. They are used primarily as referenced background knowledge. It is therefore common to annotate raw or derived datasets with information from secondary resources. For example, differentially expressed genes from an RNA-Seq experiment may be annotated with Gene Ontology terms, in order to functionally cluster results, or the same gene-set may be annotated with pathway information from KEGG and/or REACTOME, in order to identify over-represented pathways.

These data files are typically no larger than 100's of MB and this kind of annotation is not applied to every sample assay, but to summary results. Secondary data is employed to interpret results in the context of community knowledge and to generate new hypotheses for experimentation. Secondary resource annotation is also used as input for models. Model reactions, for example, could be associated with KEGG reaction information. This makes these resources invaluable for making connections between data and models.

Models

In Systems Biology projects mathematical models are made for a large variety of systems, e.g. gene regulatory networks, signal transduction, metabolism; and often work at several hierarchical scales, e.g. pathways, cells, organs, whole body, and population level. In multi-scale models several scales in time and/or space are combined. Dependent on the research question, different model types are used, ranging from core models containing just a few components to illustrate a principle, to very detailed models, potentially including thousands of components and interactions. Models can be dynamic if they contain information on the time dependency of the system, e.g. differential equation models, or structural if they only deal with network structure such as flux balance analysis models. Other classifications can be made with respect to models being discreet or continuous, include spatial information, are deterministic or stochastic.

Different model strategies have resulted in a wide range of model formalisms being used in Systems Biology projects. The formalisms most often used are: Boolean network models, Bayesian networks, Petri nets, Constraint based models, Differential equations, Rule based models, Cellular automata, Agent based models. Often such formalisms consists of different sub-classes, for instance differential equations can be ordinary or partial, and can also contain algebraic equations and time delays. Hybrid models use more than one model formalism, for instance combining ordinary and partial differential equations, or combining stochastic and deterministic methods. Multi-scale models can integrate different modeling formalisms, for instance agent based models for cells with internal dynamics modelled with ODEs. These hybrid multi-scale models are often directly coded in a programming environment to improve simulation time, making it difficult to separate model description from simulation instructions.

To integrate different model formalisms it will be important to have a standard model description **format across the different formalisms. Thus far two model formats have been widely to describe** mathematical formats: SBML and CellML. SBML covers most of the model formalisms that are used in systems biology, and extensions of the model format to include the few remaining formalisms have been proposed. CellML describes the structure and underlying mathematics of cellular models in a very general way and is mostly used for describing higher organisational systems. Both SBML and CellML are XML based model description formats that can easily be stored in model databases, such as Biomodels and JWS Online for SBML models, and the CellML model repository respectively. Models coded in simulation environments such as Matlab, Mathematica or in generic programming environments such as C++, are hard to use outside the environment and model management is mostly limited to model storage in the native format.

Samples

Samples have traditionally referred to biological samples, and identify a specific aliquot of an instance of a biological strain e.g. E.coli K12 growing in a shaking flask with a defined media, sampled at 2 hours into exponential growth. This is well suited to traditional micro-biology experiments but they do not reflect the wide variety of samples that can be taken as part of a systems biology experiment. There is often also not a comprehensive ontology that can be used to describe samples. This has been noted as problematic in Biosamples database[33]. In Biosamples they pick appropriate annotations from separate ontologies, even if they are directly applicable (e.g. a liver tissue could use naming conventions from a mouse ontology).

Systems biology experimentation can involve *in silico* samples, biological samples, ecological samples, amongst many others. Therefore the identity of samples in a metadata framework must be much more flexible for the community use. It would be useful for samples to refer to a class that can involve a large range of specialised sample types, and a much more flexible approach to using ontology terms from mixed ontologies.

[33] https://www.ebi.ac.uk/biosamples/

4.1.2 Resources and services fabric

All scientific e-infrastructures require similar structural elements. Figure 4.3 is adapted from[34] and is the roadmap used by EUDAT. It is a layered model of services and interfaces, with the cross-cutting concerns of trust and curation.

Data and Model Generation services such as quantitative data collected with physiological conditions, and processes for constructing, parameterization and validating models. Data generated that is not suitable for Systems Biology needs to be avoided or made so (if possible) by nSBCs.

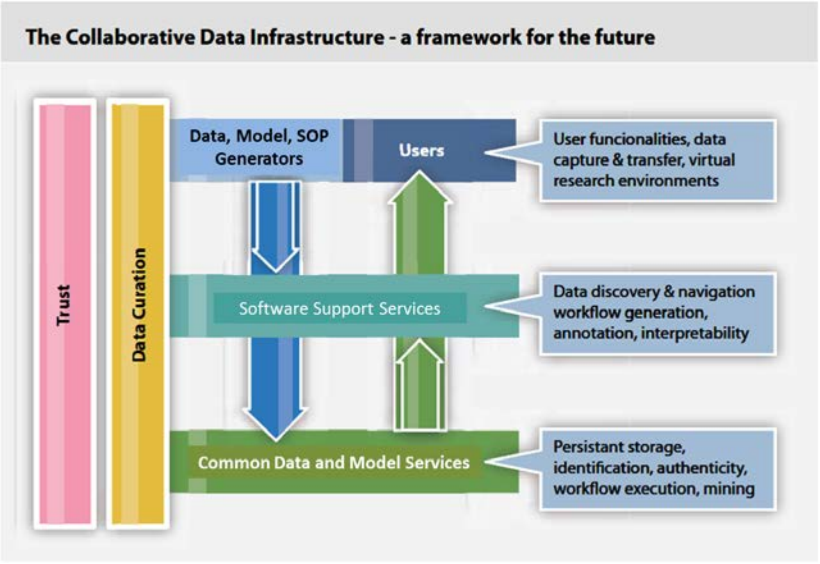

The Collaborative Data Infrastructure - a framework for the future
Trust
Data Curation
Data, Model, SOP Generators
Users
User funcionalities, data capture & transfer, virtual research environments
Software Support Services
Data discovery & navigation workflow generation, annotation, interpretability
Common Data and Model Services
Persistant storage, identification, authenticity, workflow execution, mining

Figure 4.3: Structural elements required in e-infrastructures.

Users interact with the content, allowing the identification and use of resources from within the infrastructure and from external sources.

Curation of the processes and of the data, models and SOPs themselves with compliance to community metadata standards (checklists, minimum information models, identity schemes, format and ontologies), context of the experiments and links between experiments.

Trust through management of policies and procedures, management of access and authorisation, association of data and models with their creators (to ensure credit and attribution), provenance of data and models.

Common Data and Model Services describe the physical infrastructure and the services required to interact with it. It defines where and how data and models will be stored, how they are identified, the security protocol required to access them, their versioning, backups and federation (in the case of distributed architecture). Underpinning services for data storage, access & authorisation, data shipping, data citation, cloud compute, identity resolution, preservation etc will be provided by European RIs

[34] Riding the wave How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data A submission to the European Commission, 2010, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

already in place, such as EUDAT and ELIXIR (see Section 3.2.1) or already widely used in Systems Biology. These are discussed further in Section 4.2.2.

Community Software Support Services describe the services required to discover, navigate and annotate contents (i.e. indexes, catalogues, registries and tools for interpretation). This includes tools to explore and run models in addition to data analysis and catalogues. A large ecosystem of data and model management services and platforms, integration platforms and knowledge bases and gateways already exist (surveyed in Deliverable D2.1). The "fabric"[35] of resources used to find and manage models, data, SOP, samples is the rich landscape of infrastructures that ISBE will be expected to manage, make available and sustain. Assets should, first and foremost, be placed in sustained, stable, dedicated, public repositories and catalogued by sustained, maintained, public, registries. Sustainability of repositories is critical: Nature's Scientific Data initiative[36] evaluates repositories to ensure that they meet requirements for data access, preservation and stability, with a repository evaluation questionnaire[37].

Different types of resources of the public fabric include:

- **Specialist public archives**: repositories for data, models and SOPs, and public knowledge-bases and reference databases. These repositories tend to be 'silos' tailored for specific assets (Biomodels for SBML models, Metabolights for metabolomics experiments and derived information, SABIO-RK for biochemical reactions etc). These should be the final submission location for outcomes produced using ISBE.
- **Public Projects' Commons** for project to prepare, record and share their data, models and SOPs within projects and consortia, allow them to perform analyses over their content, and to gateway to and bridge across the public archives while retaining project and investigation context. ISBE should provision and sustain a Systems Biology Commons for EU projects.
- **Catalogues and registries** for recording and finding assets, and search tools.
- **Metadata services and resources** (ontologies, templates, format management) for annotating and interoperating assets.
- **Tools** for analysis, pipelines and modelling.
- **Generic cloud repositories, institutional and national repositories** used for "black box" data deposition often to meet funder or publisher compliance, or are the best specialist resource for domain-neutral asset types, notably software (e.g. Github).

ISBE Research Infrastructure will be made up of these distributed resources and services. A single point of access for ISBE assets is needed by users and sector stakeholders. ISBE will require services to access and search across distributed resources in order to enable researchers to discover relevant information for their experiments, and mechanisms to submit new data and models back into public repositories, with policies governing when and how this occurs.

The nSBCs implementing this ISBE Infrastructure are expected to: manage public resources; offer a Commons, unified view over resources generated and used, in the context of the experiments that produced them; and support the stewardship of research assets arising from Systems Biology experiments executed by users of the infrastructure.

[35] This refers to the term "Data Fabric" coined by the Research Data Alliance https://rd-alliance.org/group/data-fabric-ig

[36] http://www.nature.com/sdata/data-policies/repositories

[37] http://www.nature.com/uploads/ckeditor/attachments/1535/SciData_respository_evaluation_April2015.docx

Many external resources will be ELIXIR resources, or will be co-ordinated by ELIXIR. *The adoption of community standards in the ISBE Infrastructure, for storing and sharing data, will allow more straight-forward interchange with ELIXIR and other resources.*

The sources of data and models for the resources:

- **nSBCs**, whereby centres are responsible for the stewardship of the models and data arising from projects or through contracts with sector stakeholders (such as funders or publishers). This content must adhere to ISBE's FAIR principles and comply to its conventions for best practice.
- **Sys Bio user community**, whereby content, independent of an nSBC, is contributed to datasets and model-sets managed by ISBE. To qualify for contribution, content must adhere to ISBE's FAIR principles and comply to its conventions for best practice.
- **Life Science community**, whereby datasets and their content are managed by RIs other than ISBE but vital to the ISBE community. Examples include Metabolights and Biomodels, managed by ELIXIR.

Specialist public archives

Archives are managed for the international community by investigator-based, national or pan-national providers. In Deliverable D2.1 we surveyed 122 Systems Biology researchers who identified over 80 repositories in use. Figure 4.4 gives the most popular public archives.

Special public archives or repositories fall into two classes:

- **Experimental reports** specialising in one type of data collection; effectively making asset silos. For example: proteomics (PRIDE), metabolomics (Metabolights), models (BioModels, JWS Online, PMR), microarray (ArrayExpress), samples (BioSamples). Some SOP registries are available as community efforts (e.g. MolMeth, OpenWetWare) or associated with publishers (e.g. Nature Protocols). In the D2.1 survey, a variety of model repositories are used for model deposition, with BioModels being the most popular (33% of survey respondents), and JWS Online the second (7.5%. 26% of respondents did not submit their models to any public repository.

- **Biological knowledgebases and reference databases**, that gather and collate information from other databases and the literature in order to give a curated, comprehensive overview of a domain. These include resources such as the pathways databases KEGG and Wikipathways, the biochemical reaction databases SABIO-RK and BRENDA, and protein sequence annotation (UNIPROT).

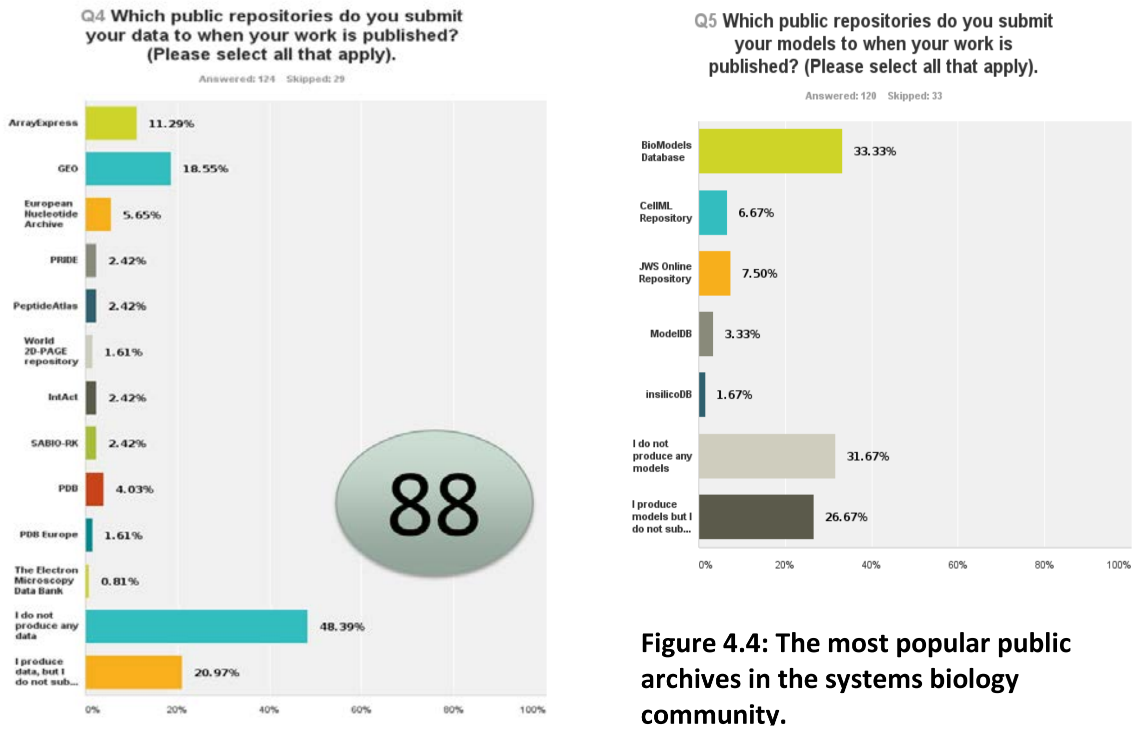


Figure 4.4: The most popular public archives in the systems biology community.

Challenges for ISBE regarding public repositories are:

- **Suitability for Systems Biology.** Most general public quantitative databases (particularly those manage by other RIs, notably ELIXIR) provide kinetic constants for enzymes, and sometimes binding constants, but do little to help building quantitative descriptions, i.e. concentrations, sizes, diffusions etc. Exceptions include gene expression data, proteomics, and metabolomics. Even so, a significant issue is one of *localisation*; for example the average concentration of a protein in a piece of brain is of limited use (due to the mix of tissues and subcellular compartments). Some limitations restrict the utilization of data for model construction and validation. This emphasises the importance of designing data collection against standardised SOPs for modelling experiments.

- **Selecting contributions.** The identification of which resources will contribute to the ISBE Infrastructure. An open and transparent process, with achievable and appropriate criteria, is needed for selecting, monitoring and sustaining the archives that should be managed by ISBE. Processes must be established to monitor the usage, performance and quality of such resources against to be established metrics.

- **Gap analysis.** The identification of gaps in repository provision for particular data, models, SOPs, samples.

- **The provisioning and sustaining of the resources.** Providers may be aligned with nSBCs and those nSBCs will contribute those resources/services to the ISBE Infrastructure. Alternatvely, they may be part of another RI (e.g. ELIXIR) and their provision to the ISBE infrastructure contributed through MoUs. Selected, key investigator-lead resources will need to be migrated to become backed sustainably by nSBCs. The curation costs of both classes of archives are significant. The burden of curation can, to some extent, be pushed down to contributors to the experimental report collections (for example using a Commons or using upload tools like ISATools). Highly curated knowledge bases, however, must bear curation costs.

- **Interoperability** ISBE is an Open Infrastructure: it does not "own" all data and model resources in Europe. ISBE will need to set up and provide a coordinated Interoperability Backbone that allows partners (e.g. other ESFRI Research Infrastructures such as ELIXIR , national resources, institutional archives) to make use of existing resources and connect and interoperate their own resources. Providing a sustainable infrastructure that manages data identifiers, secures data archiving and access, and ensures mappings between resources will enable long-term, cost-effective, data management and drive "standards as the default" across Systems Biology.

- **Managed release and service operational practices.** Public archives need transparent operational protocols, for: releasing, versioning, provenance, updates, deprecating and contributing content etc. For ISBE supported resources we emphasise self-described datasets and explicitly described and published life-cycle metadata, using machine processable representations and common APIs for access to dataset descriptors as well as access to content. We tackle experiment report repositories and biological knowledge bases differently, reflecting their different content and their content lifecycles.

- **The ISBE Solution** must be agnostic of source repository (BioModels, JWS Online, etc), of submission format (SBML, CellML, etc), and of original metadata content.

A Systems Biology Projects' Commons

A Commons is a community controlled environment that brings together distributed research assets and distributed users/contributors. Systems Biology investigations are inherently integrated, cross-asset, cross-archive, cross-researcher (experimentalist, modeller), and often cross-lab. A Commons enables researchers to catalogue, pool (exchange, share, publish), cross-link, access, and analyse their own and public assets, using their own and third party tools. Commons use is governed by established regulations and policies for behaviours, for deposition and metadata standardisation, fair use, fair reuse and fair sharing with appropriate security, privacy and access controls regulated against a minimum set of community-accepted rules (what is available at little or no cost, what can be altered and reused with few restrictions, and what is in the public domain but restricted by licences).

Commons are gateways to public archives to deposit outcomes, as well as access content, while retaining the connections to the investigation context and cross-links to related assets (models with data, data with SOPs etc). Figure 4.5 shows a schematic of the FAIRDOMHub[38], a Commons that supports projects from several national and European Sys Bio initiatives. The key part of a Commons is the ***pan-asset, pan-repository*** catalogue. Figure 4.6 shows a screenshot of FAIRDOMHub which catalogues and links the assets associated with a published investigation, which may well be stored in different repositories hosted by different organisations. Software suites and platforms to build Commons include Hubzero[39] and SEEK4Science[40]; the latter is used to deploy the FAIRDOMHub Commons.

[38] http://www.fairdomhub.org
[39] http://www.hubzero.org
[40] http://www.seek4science.org

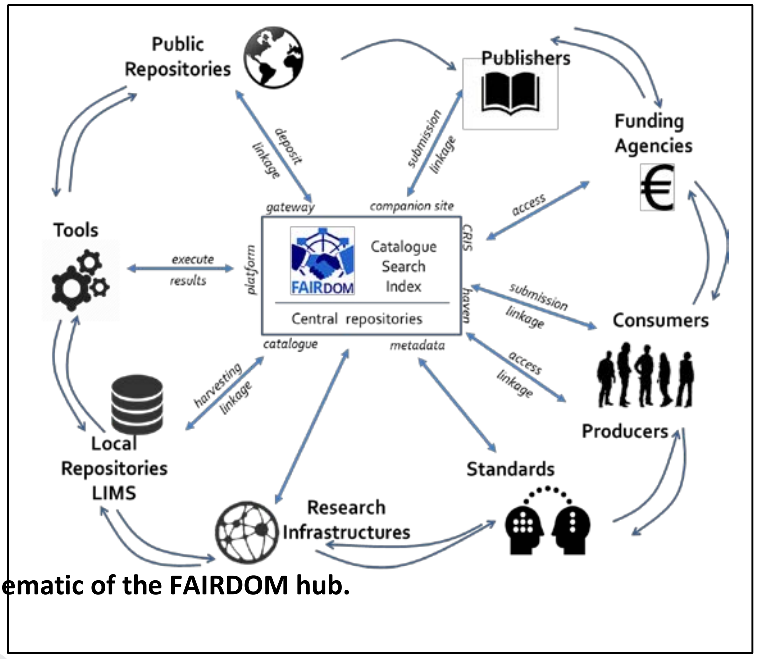


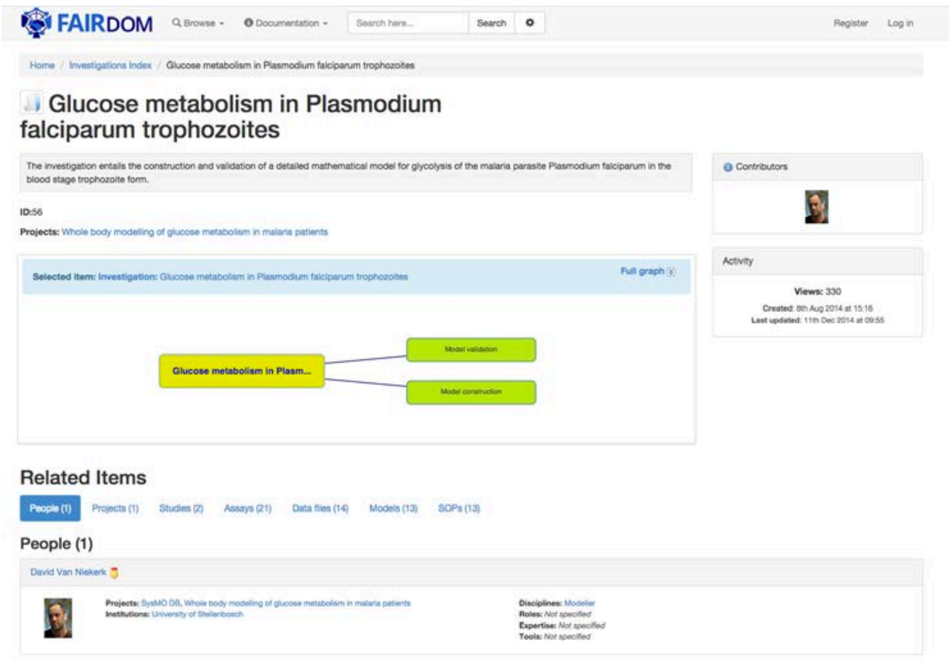
Figure 4.5: Schematic of the FAIRDOM hub.

Commons contribute to overcoming several significant asset management issues:

- **Aggregating repositories with contextual metadata**. As described above, specialist public archives for specific assets are managed by the community. There is no one database catering for all assets types and neither should or could there be. A Commons leverages and cross-links public repositories, so that the right database is used for the asset-type. The Commons retains the experimental context through the metadata used to describe and structure the assets held in different places[44]. The FAIRDOMHub uses the ISA (Investigations, Studies and Assays) framework for such structuring[38]. The aggregated content indexing/multi-repository spanning nature of the Commons is critical. It also and highlights the importance to the Commons of interoperability standards and shared identifiers. The shared stewardship of the data is both a benefit and a challenge.

- **Overcoming fragmentation**. The need for maintaining an overarching integrative context to outcomes has become widely recognised. Depositing research outcomes in type-specific silos or local/institutional/publisher repositories leads to reporting fragmentation[41]. Incrementally publishing results to feed promotion prospects, further adds to fragmentation, which exasperates efforts to find assets, present investigation context and reproduce results. This is recognised by such initiatives as and ISATools[42] and "overlay" resources or Aggregated Content Infrastructures such as the FAIRDOM Initiative[43], BioStudies[44], and the NIH BD2K bioCADDIE Data Discovery Index[45]. A Commons such as FAIRDOMHub or the D4Science platform offers the capability of cataloguing and indexing across platforms, respecting that content held in local or public repositories remains in situ. Such Commons have a sophisticated, pan-platform metadata framework, ETL pipelines and access (license) arrangements to support FAIR principles (recall: Findable, Accessible, Interoperable, Reusable). The FAIRDOMHub uses a Research Objects[46] model structured by using the ISATab format, for capturing experimental context[47].

[41] Sansone et al Toward interoperable bioscience data, Nature Genetics 44(2) 2012, doi:10.1038/ng.1054
[42] http://isatools.org
[43] http://www.fair-dom.org
[44] http://www.ebi.ac.uk/biostudies/
[45] http://figshare.com/articles/bioCADDIE_white_paper_Data_Discovery_Index/1362572
[46] http://www.researchobject.org
[47] Jones et al Capturing the Experimental Context via Research Objects, ERCIM News 100, Jan 2015



Home / Investigations Index / Glucose metabolism in Plasmodium falciparum trophozoites

Glucose metabolism in Plasmodium falciparum trophozoites

The investigation entails the construction and validation of a detailed mathematical model for glycolysis of the malaria parasite Plasmodium falciparum in the blood stage trophozoite form.

ID:56

Projects: Whole body modelling of glucose metabolism in malaria patients

Selected item: Investigation: Glucose metabolism in Plasmodium falciparum trophozoites

Contributors

Activity

Views: 330

Created: 8th Aug 2014 at 15:16

Last updated: 11th Dec 2014 at 09:55

Related Items

People (1)

David Van Niekerk

Projects: SysMO DB, Whole body modelling of glucose metabolism in malaria patients

Institutions: University of Stellenbosch

Disciplines: Modeller

Roles: *Not specified*

Expertise: *Not specified*

Tools: *Not specified*

Figure 4.6: FAIRDOMHub Commons. An investigation of many parts (data, models, SOPs) held in different places, packaged together and given a DOI (https://doi.org/10.15490/seek.1.investigation.56) for the whole package.

Such a package, or Research Object, can be bundled onto zip or Docker Image and exported as a snapshot, or evolved in situ as new research unfolds.

- **Experiment-specific, "boutique" datasets** are unique to an experiment and do not fit into general public archives. A Commons is a home for such datasets. It can also catalogue datasets held in project-specific local repositories, enforcing compliance to community standards.

- **Project asset retention, rescue and preservation,** and collaboration for investigators through managed and controlled cloud "spaces" on a commons. These spaces enable individual investigators to handle their data, methods, collaborations, and support project management. Retaining results beyond a project funding cycle or PhD student tenure, while organising investigations and consistently reporting using standardised metadata practices, improves project collaborations and productivity but must be aware of the fluid nature of project "Virtual Organisations". Import, export and search bridges between the Commons and local project stores create federation pathways between the projects and the commons catalogue, and ultimately to the public archives. Rescue can occur at end of projects, when investigators leave or move between institutes, and when funds finish. Spaces provide an asset management place for researchers unable to support their own resource.

- **Standardisation practices.** A Commons drives compliance for a Structuring, organizing, managing, and interlinking the commons components of an investigator's research experiments.

- **Availability and usage tracking of project outcomes**. A commons can act as an asset publishing repository for asset publishing and bibliometric services (see Section 4.4.1) incorporated into publishing workflows; as a record of project outcomes for funders (Section 3.2.4) incorporated into compliance workflows; and as a source of project results for industry (Section 3.2.6) as well as academics. A Commons is a knowledge portal single point of access.

- **A Science 2.0 Repository environment**. Assante et al[48] propose the notion of Science 2.0 Repository (SciRepos) that blur the distinction between research life-cycle and research publishing. They argue that research product creation and publishing should occur within the Research Infrastructure used for research, not elsewhere, and during the research activity not "on date" (when the scientists think they are sufficiently mature). SciRepos are characterised by integration with RI ICT services to intercept the generation of products; repository tools; and social networking-like scientific communication ("posting" rather than deposition; "open commenting" rather than dissemination).

Technical challenges for a Commons include:

- Managing, exchanging and processing standardised, machine actionable metadata descriptions, including support for provenance and versioning of all kinds of commons objects (assets).
- Locating, indexing and cataloguing the data, software, services, models, workflows, SOPs, samples, documents, with results scattered across the resource fabric.
- Navigating the commons to use and reuse the objects within it.
- Interoperability and analysis services: developing and executing methods for integrating and analyzing data and constructing and validating models.

Discovery catalogues

As well as the Commons services described above, catalogues of datasets and tools such as res3data.org, omictools.com and BioCatalogue.org support the Findability and Accessibility of FAIR research outcomes and applications. ELIXIR is developing a ELIXIR Tools and Data Services Registry[49] which ISBE should contribute to, as well as contributing to res3data and other public catalogues.

Metadata services and resources

Catalogues of metadata standards such as Biosharing.org, support the Interperability and Reuse of FAIR research assets. As described in section 4.2.1:

- Reporting guidelines, ontologies, formats, templates, identification schemes are used to standardise metadata and support robust, consistent and comparable reporting as well as interoperable and unambiguously interpretable outcomes.
- API practice and compliance such as Common APIs like the Global Alliance GA4GH API for the secure, privacy respecting and interoperable sharing of Genomic data, and managed APIs for public repositories to report versioning, licensing, adhere to release cycles etc. The conventions for service interoperability should be based on the minimal "hourglass" approach, a specification of lightweight interfaces, standard protocols and standard formats.

[48] Assante et al Providing Research Infrastructures with Data Publishing, ERCIM News Number 100, Jan 2015, 20-22, http://ercim-news.ercim.eu/images/stories/EN100/EN100-web.pdf

[49] https://elixir-registry.cbs.dtu.dk

Such metadata infrastructure needs tools and services to support its operation and compliance. ELIXIR proposes an "Interoperability Backbone" of services to support common practices and standards (see Sect 4.2.1) :

- Identity management, mapping and tracking services (identity authorities for specific data types and concept categories; identity resolution, identity mapping, and entity resolution.
- Reporting guidelines, formats, controlled vocabulary (ontology) and template services for creation and maintenance, mapping, compliance validation, model and data annotation etc.
- Regulated dataset publishing for API interoperability including distributed revision control, dependency management. Regulated APIs with standardised common capabilities (versioning, provenance) to platforms (databases, portals, storage), software (workflow engines, script engines) and infrastructure (cloud resources). Well described, validated and maintained APIs registered in catalogues.
- Biological knowledgebase publishing for Linked Data interoperability with 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.

Generic cloud repositories and services, institutional and national repositories

These are often used for "black box" data deposition often to meet funder or publisher compliance, or are the best specialist resource for domain-neutral asset types, notably software (e.g. Github). ISBE will also need to take advantage of, and partner with, general public and commercial repository providers such as figshare, Dataverse, and Dryad. Increasingly institutions are establishing their own repositories to comply with local and national policies, and in some cases national repositories are being established for disciplines and asset types.

In addition cloud data infrastructure providers such as dropbox, onedrive and google drive are universal, easy to use and powerful. Many institutions have developed their own cloud service, often based on ownCloud[50] (examples include SwitchDrive and polybox in Switzerland). Encryption services such as boxcryptor, truecrypt and ncryptedcloud, in conjunction with dropbox, offer secure, spacious, good uptime and sustained solution (more so than local home-grown solutions).

Local laboratory and investigator project specific repositories and filestores are often used for working data and intermediate data. All this creates a complex asset management landscape and complex asset publishing workflows.

Tools and services

Data, models and SOPs are assets produced and consumed by community, proprietary or investigator tools. Resources allow researchers to run simulations over models (e.g. COPASI, JWS Online), compare models and track their evolution (e.g. BiVes) or perform informatics analyses over data resources, and so on. Services for data management include data management planning such as DMPOnline[51] and software sustainability planning. Tools and workflows for the construction and annotation of models need to be developed and made available either through individual national Systems Biology Centres (nSBCs), or accessed directly from the central Systems Biology Centre (cSBC).

[50] owncloud.org

[51] https://dmponline.dcc.ac.uk

The conventions for tools and services interoperability should be based on minimal "hourglass" approach, a specification of lightweight interfaces, standard protocols and standard formats.

4.2 Underpinning the Asset Landscape

4.2.1 Common practices and standards

The heterogeneity and diversity of systems biology data and models gives rise to large challenges for stewardship in systems biology. For effective sharing and reuse in systems biology, common metadata, formats and ontologies need to be used and common practices for annotating, aggregating and publishing research assets need to be established. These standards should be mandated within ISBE and promoted throughout the infrastructure and the broader systems biology community by providing incentives for adoption.

The highly heterogeneous and integrative nature of systems biology means that:

- Standards are critical in systems biology because they describe how datasets are related to one another and can be used to identify identical biological objects in multiple datasets.
- Systems biologists have to work with a wider number of standards than researchers in other fields because of the broad range of data and modelling types that they work with and produce.

Standards exist to describe (i) the metadata elements that should be recorded for a given research asset , (ii) the syntactic formats they should be stored and exchanged in (iii) the semantic vocabularies or ontologies used for describing entities in systems biology data and models (iv) the relationships between research assets.

Metadata - is the description of the data, its content and its origin. Minimum Information guidelines are prevalent in the Life Sciences. Minimum Information is the least amount of metadata required in order to understand and interpret data. MIAME (Minimum Information about a Microarray Experiment) was one of the first to be developed. Biosharing.org hosts a large collection of Minimum Information guidelines, covering both omics technologies and modelling approaches (e.g. Minimum Information About a Proteomics Experiment (MIAPE), Minimal Information Required In the Annotation of biochemical Models (MIRIAM), Minimum Information About a Simulation Experiment (MIASE)).

Formats - mark-up languages specify the syntax, content and style of metadata elements. Like minimum information guidelines, there are a large number of biological mark-up languages, predominantly described in XML format, although some now also have tabular versions. Examples include the Systems Biology Markup Language (SBML) and mzML for mass spectrometry data.

Ontologies - an ontology is a set of concepts that describes a domain and describes the relationships between those concepts. The comparison and exchange of biological data and models can be facilitated by the use of common ontologies as annotation vocabularies. In systems biology, the Gene Ontology, the Systems Biology Ontology (SBO) and ChEBI (Chemical Entities of Biological Interest) are widely used examples.

Relationships – in order to interpret a systems biology experiment, it is necessary to understand how multiple datasets relate to one another and how those datasets relate to any models and model predictions. Standards which allow the aggregation and interlinking of research assets are required for this. Examples include the Research Objects model (RO) and the COMBINE Archive.

Despite the clear necessity for standards use in systems biology, some standards are poorly adopted. The community survey designed in WP2 (WP2.1 Appendix C) established not only the current standards (formats, MI checklists, vocabularies) that are utilised within the Systems Biology community, but also identified key obstacles which currently prevent more efficient working practices. These must be addressed, and must be a paramount consideration in the design and implementation of proposed solutions going forward.

The finding of this survey suggest that:

- SBML was the most popular submission format (59%)
- MIRIAM was the most popular checklist used (34%)
- 44% of respondents did not use and minimum information guidelines (Figure 4.7)
- The most popular ontologies used were Gene Ontology (47%), which is not specific to the Systems Biology domain, and Systems Biology Ontology (28%)

There is more uptake of modelling standards, which could be largely due to the successful running and community engagement of COMBINE. COMBINE is an international *grass-roots* initiative to develop standards for systems biology modelling. These include SBML, CellML, SED-ML and SBGN and they already have stable specifications, implemented software support, a user-base and community governance.

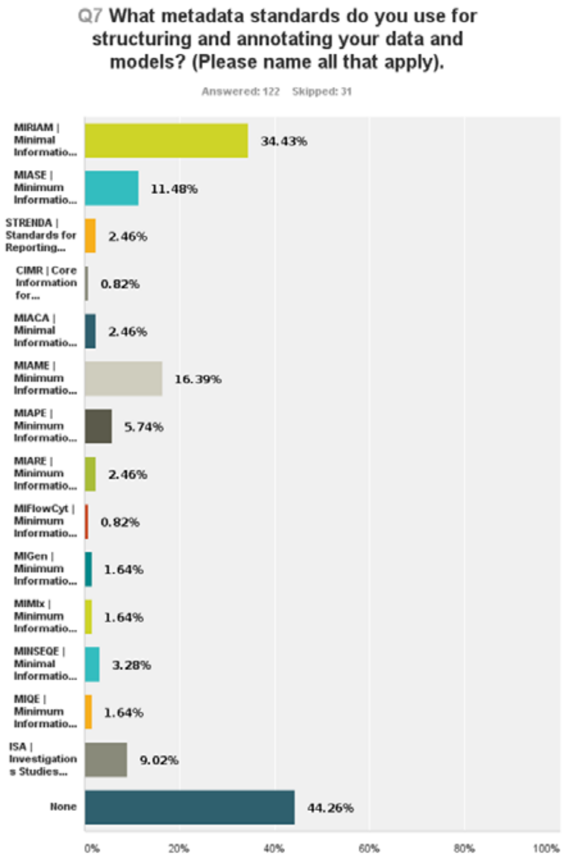


Figure 4.7: prevelance of minimum information guidelines within the community.

Systems biology stewardship requires the use of multiple standards. Recommendations for metadata, formats, ontologies and best-practice will have to be defined by ISBE, to ensure that research assets are consistent and exchangeable across ISBE centres and are immediately reusable by the systems biology community. All components that form part of the ISBE infrastructure and all components that integrate with it (e.g. specialist public archives, public project commons, catalogues and registries, analysis and modelling tools, and generic repositories) should enable standard data exchange, by providing tools and incentives. Training on the use of standards is also important for widespread uptake.

ISBE standards recommendations should encompass the whole stewardship life-cycle, from pre-proposal planning, to project collaboration, to publication, reuse and archiving. They should also cover stewardship activities for every type of research asset. Deliverable 2.1 provides a comprehensive survey of available standards and further usage summaries from survey respondents. The survey also identified important deficiencies and inefficiencies in current Systems Biology working procedures:

- Respondents found a lack of construction and validation tools (63%)
- A majority of users found it difficult to reuse models (72%)
- Annotation (72%) and curation results (47%) were insufficient
- Specific lack of annotation for parameters (66%), reactions (60%) and species (58%) was identified
- The use of SOPs was limited (39%)
- 26% of respondents did not submit their models to any public repository

Areas that require the recommendation and application of standards span the technical, social, cultural and environmental aspects of the capability framework and are as follows:

- **Data and model identifiers.** All biological objects and model components should be annotated with common, persistent identifiers from public repositories, in order to enable cross-linking and comparison. Identifiers.org is a system which provides resolvable, persistent URIs (Uniform Resource Identifiers) and therefore provides a uniform interface for accessing identified data from specified reference locations.

- **Vocabularies, ontologies and semantics**. Biological objects, model components and their interactions should be described with community ontologies wherever possible. The adoption and use of community ontologies (such as GO and SBO) enable semantic cross-linking and comparison. Arguably, the most critical step in the systems biology research asset life cycle is the integration of data and models. This also involves the use of ontologies to describe entity relations and corresponding qualifiers, as well as the evidence and quality and provenance of data and model entities.

- **Data/model/SOP formats.** Standard formats for research assets of all kinds assist with exchange and comparison. If standard formats are in common use in a community, software developers can use these for import/export and interoperability. ISBE should develop a clear set of guidelines for standards use across the different types of research asset. ISBE centres should implement the use of these standards across ISBE centres and promote their use in the wider community.

- **Stable, documented APIs.** ISBE will provide tools and services for the systems biology community and will steward other community resources. By stipulating the requirement for common, stable APIs, to access and search research assets, ISBE will be able to increase interoperability.

- **Standard Operating Procedures.** Research results are often only comparable if the experimental conditions and methods used in the procedure are standardised. ISBE Commons publishing will enable the sharing of Standard Operating Procedures (SOPs), with common formats and content, within individual projects and potentially across the community, fostering reuse and potentially refinement of successful SOPs.

- **Standard Processing workflows**. As with SOPs, standard processing workflows that capture *in silico* methods, for preparing and analysing data, should be captured and shared. Unlike experimental SOPs, *in silico* SOPs could be made executable. Standard workflows for processing, analysing, formatting and curating data/models will enable the validation and reuse of methods and better comparison of results.

- **Reproducible models.** Models are knowledge-rich structures. They are only comparable and reproducible if the components, interactions, parameters and parameter value provenance are captured and understood. An ISBE Commons could provide an environment for sharing and versioning standardised models and simulation results, drawing on existing model standards, such as SBML and MIRIAM, and work from the COMBINE archive and SED-ML.

- **Provenance.** The ability to trace the history of information relating to an experiment, from the data or models perspective, is an important part of experimental validation and reproducibility. The COMBINE Archive and the Research Object (RO) model both address this issue, by enabling the packaging and interlinking of multiple research assets and their relation to one another, allowing scientists to record how data or models were produced and with which methods. ISBE Commons should support the import, export and generation of COMBINE Archives and Research Objects. Work to align the COMBINE Archive and the RO model is already underway.

- **Research asset aggregation, packaging and exchange.** Standards for recording the provenance of experiments contribute to reproducibility and reusability, but in systems biology, standards for recording how datasets are related, and which data was used to construct or validate a model are also essential. The COMBINE Archive, ROs and ISA (Investigations, Studies and Assays) standards allow the description of data or models *in context*.

- **Research asset citation.** Research assets that are frequently reused by the community are valuable commodities. The producers of such resources should be credited in a similar way to researchers who produce a paper that is highly cited. This offers incentives for sharing and publishing all research assets, and not only those that are included in a publication. Standard, persistent identifiers for research assets should be minted on publication, with metadata detailing the researcher and the project they originate from. Digital Object Identifiers (DOIs) are an established standard that can be used for this purpose.

- **Construction and annotation of models** need tools and workflows to be developed and made available either through individual nSBCs, or accessed directly from the cSBC. On submission, there should be automated processes in place to ascertain the annotational content of the model, with respect to species and parameters, and with respect to provenance metadata. Where annotation levels are sufficient, the models would pass to the nSBC responsible for curation; where insufficient, it should pass to the nSBC responsible for annotation. It is anticipated that the nSBCs for annotation/curation would be encoding-format specific, with an individual nSBC taking responsibility, for example, for SBML-encoded models. To increase the reusability of models, accepted models should undergo a transformation process, which would generate a variety of alternative formats. For example, an SBML encoded model can already be transformed into an SBGN (graphical) format. There are various transformations that can be undertaken for the various formats, but some of these may be 'lossy'. Hence, it is necessary to: document the original format (and model); provide a list of alternative formats in which the model can be made available ; provide a list of the alternative tools that can produce the desired transformation; document the performance of the model in the transformed format (additional curation); document and cross reference to any data that is used for the model, with the aim to identify any similar data that could be used to corroborate the model in the future.

Standards as a sustainable foundation

Standards are essential in scientific data management across all domains. Certain standards are systems biology-specific (e.g. SBML for systems biology models), but other standards underpin work across broader areas of science (e.g. MIAME for transcriptomics, or the Research Objects Model, for packaging and exporting scientific investigations in any field).

Standards recommended and implemented by the ISBE infrastructure should have the following common properties:

- Interoperability with other standards, to allow exchange and synergy with related resources and infrastructures

- Support multi-scale data integration and modelling
- Take into account different national regulations (e.g. for sensitive clinical data)
- Allow reproducibility of experiments and model simulations
- Allow unambiguous tracking of datasets and entities (unique identifiers)
- Be supported by publishers (authors of submitted manuscripts for publication should submit their data and models in standardised formats along with the manuscript)
- Be supported by funders (standards could be recommended or mandated in common funder sharing policies, or funders could specify that researchers use current ISBE recommendations)

The assessment of the standards landscape in deliverable D2.1 revealed a wealth of resources that ISBE could exploit and build upon:

- **Reusing and driving adoption of established standards**. Despite the complexities created by data and model heterogeneity and by the necessity to link data and models, most recommendations made by ISBE and best-practices mandated for ISBE centres can be based on existing community efforts. ISBE can also play a key role in driving standards and common practices for citing research assets. The ISBE Commons infrastructure provides an interface for persistently identifying and accessing assets. It could also expose information on the number of citations and uses that research assets have, promoting those that have large uptake.

- **Developing standards where necessary.** ISBE should take the lead in developing new community efforts in areas where standards are lacking. One notable example is in the area of multi-scale modelling. More research is required to determine how to semantically interlink existing model information relating to different anatomical, spatial and temporal models; and how to manage the different technologies and mathematical modelling methods already in use. Another current problem is the recording of parameter value provenance in models. This is a major barrier to the reuse of existing models because researchers cannot assess whether those values have been experimentally determined or estimated. If they have been experimentally determined, they cannot critically assess the methodologies used if the link with experimental work has already been broken. ISBE can drive these new kinds of standards initiatives due to its unique position as a community nexus.

Engaging with standards setting

For a sustainable approach to standardisation in ISBE, standards that are systems biology-specific, and internationally accepted, should be adopted and ISBE should contribute to on-going development activities to ensure their continuation. Many systems biology-specific standards have emerged from the activities of COMBINE.

COMBINE ('COmputational Modeling in BIology' NEtwork) is a grassroots initiative to coordinate the development of the various community standards and formats for computational models. By doing so, it is expected that the federated projects will develop a set of interoperable and non-overlapping standards covering all aspects of modelling in biology. Building on the experience of mature projects, which already have stable specifications, software support, user-base and community governance, COMBINE helps foster or support fledgling efforts aimed at filling gaps or new needs. As those efforts mature, they may become part of the core set of COMBINE standards. One of the initial activities of COMBINE is to coordinate the organization of scientific and technical events common to several standards. Those events, as others related to our field of research are gathered in a calendar. ISBE stewards should become active members of COMBINE and ISBE should develop strategies to ensure continued funding for organisations like COMBINE. Standards must be freely available for the whole community, but maintenance and development can be costly. Grassroots initiatives rely heavily on time and resources from the research community, but they still require funding for workshops and *hackathons*.

For cross domain standards, ISBE should contribute to broader, international standards initiatives wherever possible to ensure cross-pollination of ideas and synergy with the broader scientific community. This should be both at the grass-roots level (e.g. *HUPO Proteomics Standards Initiative (PSI), the Functional Genomics Data Society (FGED), the Metabolomics Standards Initiative (MSI) and other resources from biosharing.org)*, and in international initiatives, such as the Research Data Alliance, the International Organisation for Standardisation (ISO) and other ESFRI initiatives, like ELIXIR.

The Research Data Alliance Publishing Data Interest Group[52] brings together all stakeholders involved in publishing research data including researchers, discipline specific and institutional data repositories, academic publishers, funders and service providers. The working group focuses on:

- Reporting guidelines, ontologies, formats and identification schemes to standardise metadata and support robust, consistent and comparable reporting as well as interoperable and unambiguously interpretable outcomes. These tasks align completely with the aims of the ISBE stewardship standardisation requirements.
- API practice and compliance, for example, common APIs like the Global Alliance GA4GH API for the secure, privacy respecting and interoperable sharing of Genomic data, and managed APIs for public repositories to report versioning, licensing and adherence to release cycles. By adopting the use of common APIs, ISBE can ensure broader interoperability and compliance with emerging regulations.

ELIXIR is developing a sustainable European infrastructure for biological information, which encompasses all available types of published biological information across the life sciences. As described in Section 4.2.1, ELIXIR proposes an “Interoperability Backbone” of services, to support:

- Practices of data management and data publishing; managed APIs and message formats, with agreed APIs for access to dataset descriptors.
- Common exchange formats.
- Common reporting guidelines: submission, curation and validation tools using data templates, focusing on interoperability of standards via common data element mappings.
- Common ontologies and terminologies
- Common APIs: for common data types.

For biological knowledge bases common conventions are to be developed 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.

[52] https://rd-alliance.org/groups/rdawds-publishing-data-ig.html

- Good practices for publishing data as Linked Data, leveraging the EMBL-EBI's RDF Platform and resources of other platforms (e.g. IMI Open PHACTS Discovery Platform), as a semantic interoperability platform in addition to the use of APIs.

It is essential that ISBE aligns with these activities, as large amounts of ELIXIR-managed data will be of interest to the systems biology community and should be accessible through ISBE. ISBE will connect to multiple ELIXIR-managed resources, as a consumer and contributor. Standard formats, exchange protocols and ontologies are key to achieving this synergy.

The International Organization for Standardization (ISO) and the European Committee for Standardization (CEN) are international standardisation initiatives across a broad range of sectors in science, engineering, industry and beyond. They have dedicated efforts in the Life Sciences domain, which are directly relevant to ISBE activities. For example, the technical committee for biotechnology standards (ISO/TC 276) has initiated a sub-committee and working groups to develop a standard framework for the interoperability of existing life science standards. Rather than integrating existing community standards, such a standard framework has to refer to existing standards requirements and standard formats for the generation, formatting, description, processing, visualization, validation and downstream-integration of research data, as well as for the creation, formatting, description, visualization, validation and simulation of such computer models and their simulation results. In doing so, the domain-specific standards can be constantly adapted to novel technological developments and kept up-to-date by the domain-specific experts, whereas the framework standard would be kept steady as a hub.

The initial work of defining such a domain-independent horizontal framework that allows the assembly of complex data and model conglomerates has already begun. The 'Data Processing and Integration' committee includes experts from different European countries, as well as from the United States, Canada, China, Japan, South Korea and other countries and focuses on biotechnology processes, including many fields that are crucial for systems biology, such as terms and definitions, biobanks and bioresources, analytical methods, bioprocessing, data processing and integration (including annotation, analysis, validation and metrology. An alliance with a world-renowned standardisation body like ISO is beneficial for the dissemination and uptake of the defined standards (especially for the non-academic stakeholders of a pan-European systems biology infrastructure).

4.2.2 Technical infrastructure and services

The technical infrastructure that supports Systems Biology research comprises tools and services that are used at different stages of the research lifecycle. The technical infrastructure and services enable the stakeholders to manage experimental data, SOPs, models, as well as other actionable procedures such as workflows and virtual machines. Platforms range from LIMS to Electronic Lab Notebooks; from annotation tools to model simulation and validation. Tools range from custom tools and scripts, to commercial offerings (e.g. Matlab) and open source (e.g. R). Section 4.1.2 outlined many of the tools and services of the Community's Data Fabric.

Common Data and Model Services in Figure 4.3 described the physical infrastructure and the services required to interact with it, defining where and how data and models will be stored, how they are identified, the security protocol required to access them, their versioning, backups and their federation (in the case of distributed architecture). Underpinning services for are needed for data storage, access & authorisation, data shipping, data citation, cloud compute, identity resolution, preservation etc. The relationship between technical infrastructure, security concerns and institutional/national barriers for sharing data and compute is complex. Compute power to process data may be secured behind firewall. External web servers handling data access may not be set up to handle the loads or have the necessary bandwidth. If shipping data is prohibitive (due to size or security) than compute may ship to data (via VMs, for example). Services may be open source or proprietary; free or commercial.

Many of these services are already provisioned by European e-Infrastructures or by well-established commercial or open source platforms. ISBE needs to build on or establish a basic infrastructure layer that provides basic computational services by providing a gateway to European e-Infrastructure services (GÉANT, EGI, EUDAT, PRACE) in partnership with ELIXIR. ISBE needs to define Systems Biology service requirements and identify areas and activities that could be sourced by the European e-Infrastructures.

Services include:

- **Computational tools:** e.g. computational workflow systems (Taverna, VisTrails, Galaxy, KNIME etc); scripting environments (R, Matlab); and programming environments (Python, C++).
- **Versioning and release management systems**: e.g. Github
- **Data transfer and synchronisation**: Data transfer using transport mechanisms (e.g. GridFTP, http, Aspera, UDPipe, iRods); collaboration with GÉANT (e.g. bandwidth-on services) for dedicated network links (e.g. lightpaths) for regular or large data transfer activities between the nSBCs. Commercial systems like dropbox and EU dropbox-like platforms (B2DROP) support synchronisation.
- **Data storage**: ISBE nSBCs will need to provision the data storage (plus backup and backup protocols) for the Sys Bio Commons and the supported public archives.
- **Data replication**: Data replication (an updated dataset being moved to multiple remote locations) and data submission (where a dataset is made available for subsequent retrieval and remote analysis). Replication policies around the data and updates any relevant data catalogues (e.g. B2FIND) using triggers; The LOCKSS Program is another platform - an open-source, library-led digital preservation system built on the principle that “lots of copies keep stuff safe.”
- **Dataset pull for detailed analysis** (e.g. Galaxy running on an ISBE-affiliated cloud resource during training event) which may be discarded after processing and just the results retained.
- **Data location services**. To manage and discover data replicas within ISBE sites (e.g. B2FIND or the EGI Data Catalogue). AAI mechanisms and workflows (e.g. REMS) will be needed for gaining approved access entitlements.
- **Computation**: e.g. cloud (Amazon, Microsoft, Google), cluster (CONDOR) and grid computing platforms, both commercial and public. The EBI’s Embassy Cloud is of particular interest for secure computing. ISBE needs to integrate cloud and compute resources available to ISBE nSBC compute centres, and access to open-source cloud technologies, to support scientific software workflow, scripting and simulation platforms for pipelines and predictive modelling.
- **Virtual machines and packaging**: for remote compute, reproducible modelling and platform exchange: e.g. DOCKER (for packaging files) or VMWare for Virtual Machines, with VM farms for model simulation.
- **Authentication and Authorisation**: for access management. ISBE could use the services of European federated identity that establishes an ISBE Identity and provides additional AAI services, attribute self-management, authorisation management, and credential translation.

ISBE needs to build on or establish a basic infrastructure layer that provides basic computational services, such as data storage and data transfer (local, or cloud-based, plain or encrypted). Building on these, the repository and commons platforms (for FAIR storage of data), cataloguing platforms (for FAIR cataloguing of metadata), integration and collaboration platforms maybe built. These platforms support and interface with computational tools such as tools for model versioning and annotation or model simulation platforms (e.g. JWS Online, Biomodels).

The full usefulness of these technical services can only be established by establishing the appropriate training and support, both informing about the technical challenges on how to make best use of the systems involved (how to store and preserve large data sets, how to use the Commons system), as well as support centred around the assets, such as support for curation (enriching and interlinking data), including technical model curation, as offered by e.g. JWS Online or Biomodels.

4.3 Access, Sharing and Collaboration

4.3.1 Openness, Accessibility and Availability

Open Science[53] encompasses the ideals of transparency in experimental methodology, observation, and collection of data coupled with the public availability and reusability of scientific data and public accessibility and transparency of scientific communication. Openness is a means of achieving accelerated knowledge transfer and networked science[54]. Underlying this open ideal is a notion of voluntary sharing of methods, results, and scholarship, and the objects of scholarship belonging not to the individual scientist but to the larger community.

The Open Science movement has gained traction in the past five years. Open access to research data, models and, more recently, software, is increasingly seen as a principle in many research communities. The Royal Society produced the influential Science as an Open Enterprise report[55] outlining six areas for action:

- Scientists need to be more open among themselves and with the public and media
- Greater recognition needs to be given to the value of data gathering, analysis and communication
- Common standards for sharing information are required to make it widely usable
- Publishing data in a reusable form to support findings must be mandatory
- More experts in managing and supporting the use of digital data are required
- New software tools need to be developed to analyse the growing amount of data being gathered

Publishers have changed editorial policies to mandate the availability of data and software for peer review (see section 3.2.4). Funders such as the NIH, H2020 and national funding agencies such as the UK EPSRC have developed open science principles and established open science mandates (see section 3.2.4). Table 4.3 gives a typical public funder agency's open data principles.

[53] The Research Information Network. Open science case studies. http://www.rin.ac.uk/ (2010)

[54] Goble, De Roure, Bechhofer, Accelerating Scientists' Knowledge Turns, Knowledge Discovery, Knowledge Engineering and Knowledge Management Communications in Computer and Information Science 348, 2013: 3-25

[55] https://royalsociety.org/policy/projects/science-public-enterprise/Report/

Table 4.3: UK's EPSRC Data sharing principles

EPSRC-funded research data is a public good produced in the public interest and should be made freely and openly available with as few restrictions as possible in a timely and responsible manner
EPSRC recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process (including the collaborative research process) is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these constraints are considered at all stages in the research process.
Sharing research data is an important contributor to the impact of publicly funded research. To recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
EPSRC-funded researchers should be entitled to a limited period of privileged access to the data they collect to allow them to work on and publish their results. The length of this period will depend on the scientific discipline and the nature of the research.
Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice and should exist for all data. Data with acknowledged long term value should be preserved and remain accessible and useable for future research.
Sufficient metadata should be recorded and made openly available to enable other researchers to understand the potential for further research and re-use of the data. Published results should always include information on how to access the supporting data.
It is appropriate to use public funds to support the preservation and management of publicly-funded research data. To maximise the scientific benefit which can be gained from limited budgets, the mechanisms for managing and providing access to research data should be both efficient and cost-effective in the use of such funds.

Preliminary figures on the first wave of open data pilot projects in Horizon 2020, the opt-out rate among proposals submitted to the "open by default" categories was below 30%, and the opt-in rate amongst other proposals was around about the same[56]. This suggests that at least in principle some scientists are happy to share data.

However, our own findings and experiences[57] in Systems Biology are more nuanced: briefly, modellers are more likely to share than experimentalists, and even if products are available they may not be sufficiently well described to be independently reusable or even interpretable (interestingly, experimentalists are more comfortable reusing models than modellers). Mismatched motivations, value placed on knowledge and social capital, reward schemes, poor reciprocity and distrust together conspire to block the circulation of knowledge. Issues cited by researchers include:

- **Copyright and licensing** A lack of knowledge about the legal aspects of data sharing and data reuse, in particular around intellectual property rights, copyright and licensing, is a barrier for opening data and re-using someone else's data.
- **Cautious sharing** Researchers invest significant time and effort in collecting hard won data to be used as a competitive advantage over others. Incentives for researchers to share data are comparatively weak. Credit and career progression concerns drive researchers to cautious sharing practices (see section 4.4 for more discussion on credit).

[56] Sandberg et al Open Data – What do Research Communities Really Think About It?, ERCIM News 100 Jan 2015

[57] WP2 systems biology survey that formed D2.2.

- **Infrastructure.** Access and use tracking requires adequate information and communication technology infrastructure.
- **Sector policies.** Policies and practices at universities that prefer patenting over publishing (also a barrier to the replication and validation of scientific experiments). The policies and practices of scientific publishers that limit web-based access to research results hinder openness. Priorities of lab directors who do not prioritise time, resource and training to curate for publishing other than the minimum needed to jump through compliance and publishing hoops.
- **Curation costs.** The time and resources needed to prepare; the inconvenience and/or difficulty in preparing to share or sharing; potential long-term sustainability obligations of shared results (including answering questions) are further dis-incentives.
- **Available rather than open data.** Access to scientific data is often subject to administrative, legal and privacy regulations. Personal data, and commercially sensitive data, is particular challenging, esp in the light of large-scale combination and data mining and when data are re-purposed for processing outside the original ethical collection rules. The "A" in FAIR stands for Accessible, not open.

Privacy aware management of data is one of the key challenges to be met by systems approaches such as systems medicine, synthetic biology, and -at the core- systems biology. Even now, systems biology projects need to be carefully designed to minimize interfaces between privacy sensitive clinical and less privacy sensitive systems biology data. In particular, human genome data can be perfectly identified, as it is unique and does not change over a lifetime. At the same time, it carries sensitive information, not only about the person who has a given DNA, but also about its relatives. This in turn causes ethical and legal problems of clinical day-to-day work as well as data management, as addressed for example by the Eurat and the Global Alliance 4 Health (GA4GH). At the same time, these problems spawn new fields of research, e.g. research about Genome Privacy, i.e. ways to combine the advances in privacy enhancing techniques with the field of 'Omics analyses. Along with such technical research, there is research into consent models that respect the patients' privacy as well as making the most of study participation in the interest of both science and patients.

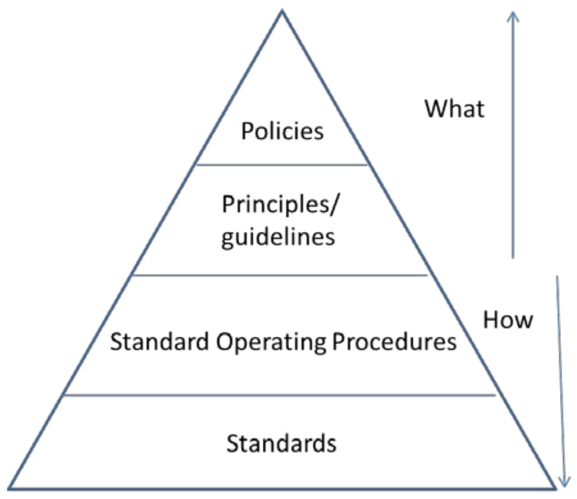


Figure 4.8: Data and model security.

For patient sensitive data ISBE may well have to establish "Data Safe Havens" to mediate between research access and patient privacy. A Safe Haven provides a range of services that minimise the risk of data leakage. The EBI's Embassy Cloud is a step towards a Safe Haven; the Farr Institute of the UK are establishing their own Safe Havens to policies and SOPs certified to ISO27001[58] (ISO/IEC 27001 - Information security management).

Figure 4.8 is a variation of the Gartner triangle for enterprise security, which provides a roadmap for ISBE.

[58] http://www.iso.org/iso/home/standards/management-standards/iso27001.html

ISBE's FAIR principles promote that data and models in the academic domain should be shared with the community as soon as possible, and within the commercial domain subject to IPR and licensing.

This requires ISBE to:

- Provide practical policies, guidelines, SOPs and standards for making assets available;
- Provide training and support on the legal issues and licensing;
- Provide stewarding services to ease the burden of curating (see section 4.5.2);
- Provide tools and services to enable these policies and, ultimately, the sharing and reuse of data, particularly for handling sensitive data.
- Monitor and influence the legal frameworks around sharing of science-relevant clinical data, providing services for projects that encompass data with multiple levels of privacy sensitivity, providing data integration and aggregation across privacy levels.
- Work with other RIs notably ELIXIR and BBMRI) in the CORBEL project to provide secure cloud, secure transfer and Authentication and Authorisation services.

4.3.2 Community Collaboration, Cooperation and Interaction

There is a growing interest among policy makers and scientists in open collaborative work. This implies identifying and reducing barriers to inter-institutional, inter-disciplinary and international collaboration among researchers, research institutions, industry and citizen groups. For example, science-industry initiatives are increasingly used to reduce the costs of and barriers to drug discovery by applying semantic technologies to available data resources (e.g. Open PHACTS Discovery Platform). Entrepreneurial initiatives are also emerging, such as ResearchGate, a social networking site for scientists to connect, raise and answer questions, and share papers and data.

Systems Biology, by its very nature, is multi-disciplinary, requiring cooperation and collaboration between modellers, experimentalists and bioinformaticians: within the discipline/sector; within and across projects and programmes; within and across laboratories and institutes; across disciplines (notably experimental and modelling); across sector and with the public. The SysMO-DB project, which developed the SEEK4Science asset sharing environment for 15 multi-institution projects in the ERANet Systems Biology for MicroOrganisms, specifically tackled the challenges of community collaboration. The behavior of the members of the SysMO-SEEK Commons highlight the value placed on knowledge capital and the distrust that lies between rivals, manifested as incremental sharing that widens the availability of content as its local value proposition changes[46]. At first (or perhaps only) an individual or laboratory uses the e-Laboratory as a private, preserved repository. This is useful when scientists are mobile, moving from grant to grant and institution to institution. Next, trusted collaborators within each project may exchange pre-published content. Results shared outside a trusted group prior to publication are rare. When a scientific article is finally published publicly we could expect its associated data/method/model to be deposited publicly. However, if the investigator thinks they can wring another paper out of some data they will not share it even if it is the basis of an announced result. Data are only made widely available when their local capital is exhausted. We also observe that (i) models, procedures and workflows are more likely to be openly shared than data, suggesting that the scientific community places greater value on data than experimental method; (ii) formal consortia are less likely to publicly share than individuals; and (iii) young researchers and very senior well established researchers are more willing to share than mid-career researchers in the midst of establishing their reputations.

These collaboration behaviours are well known and the subject of a whole field of research in the Information Sciences (including a dedicated conference series “Science of Team Science”). Table 4.4, for example, situates collaboration behaviours around data in terms of goods, rivalry (subtractability) and whether individuals can be excluded.

Table 4.4: types of data in a knowledge commons, adapted from (Hess and Ostrom 2007)[59]

		Subtractability	
		Low	High
Exclusion	Difficult	Public Goods Open data	Common-pool resources Data repositories
	Easy	Toll or club goods Data by subscription	Private goods Competitive data “raw” data

Further complications include:

- the joint creation of “goods” – data, models and SOPs - and the assignment of credit.
- the changing nature of group membership: staff and students leave, perhaps to join rivals; new project consortiums form.

ISBE’s FAIR principles promote the joint making, sharing and reuse of data, models and SOPs.
This requires ISBE to:

- Provide a Commons and promote platforms capable of supporting the sharing permissions of collaboration behaviours, including social collaboration features;
- Provide advice and infrastructure to support the crediting of joint goods;
- Provide advice and infrastructure to track the reuse of goods.

4.4 A Culture for FAIR Asset Management and Sharing

4.4.1 Academic Research Culture

Scientists are just people working within their social norms and as self-interested as any other group of people. Their prime motivations include funding, building reputation, and getting sufficient time, space, and resources do their research. Sharing results is not a motivation in itself, so has to be placed within a context of maximizing reward, minimizing risk and optimizing costs:

- **Reward for Sharing**: to gain competitive advantage over rivals by establishing a claim on priority of a result; to establish public reputation and recognition through credit; to accelerate the widespread adoption/acceptance of a result; to gain access to otherwise unavailable instruments, data, techniques or expertise.
- **Risk of Sharing:** the threat of rivals gaining a competitive advantage; damage to public reputation through scrutiny or misinterpretation; not getting credit; this leads to “data hugging” and fears of

[59] Borgman, Big Data, Little Data, No Data: Scholarship in the Networked World, MIT Press, 2015.

asymmetric reciprocity – that is failure by consumers to credit the provider, or contribute comment or review. No feedback on results that arose from using the asset fosters a sense of "free riding".
- **Cost of Sharing**: the time and resources needed to prepare; the inconvenience and/or difficulty in preparing to share or sharing; potential long-term sustainability obligations.

There are also risks and reluctance associated with the reuse of others' assets, notably:
- **Reward or obstacles for reuse**: journals with policies of "novel data/model only" acceptance
- **Quality and validation**: concerns that the assets are not of a high enough quality and their descriptions and SOPs too scant to fully reuse.
- **Entrepreneurship, innovation and risk.** Systems biology and synthetic biology represent a move to a new paradigm of research requires investment in time and effort, and this can impact on a researchers willingness to share results and resources. On the other hand a highly innovative and experimental academic culture may foster risk in new forms of asset recognition and scholarly communication.

These risks lead to "data flirting" where scientists strategically (or maybe tacitly) hold back information, communicating just enough to interest their community and publish just enough to preserve their claim for priority on the findings but not enough in practice for competitors to be able to take advantage. Specialized knowledge on experimental details is withheld. Scientific jargon is used to frustrate competitors. This "counterfeit sharing" is prevalent when funders make data sharing directives. The data is deposited, and thus "shared", but it is not FAIR: hard to find and impossible to reuse or reproduce[60].

A survey of 1329 scientists showed willingness to share their data in return for credit[61]. Data and models must be citable and credit given to their authors and stewards, and commoditised so that they can be re-used modularly. Isolated activities or actions won't impact the need for the "republic of science" to abide by shared behavioural norms. Instead:
- Credit needs to underpin the whole system, reaching from strategic planning and overall polices to the mindset and everyday practice of the individual researcher.
- Credit for contributing to data or SOPs (less problematically for models) needs to be de-conflated with authorship. Currently trackable attribution is only possible through authorship of a paper or by a proxy Data Journals.
- Citation of data (and other assets). Several authoritative studies recommend uniform direct citation of data archived in persistent repositories, so that data are to be considered as first-class scholarly objects and be treated similarly in many ways to cited and archived scientific and scholarly literature. Force11 published the Joint Declaration of Data Citation Principles and a framework for operationalizing the JDDCP with and a set of initial recommendations on identifier schemes, identifier resolution behaviour, required metadata elements, and best practices for realizing programmatic machine actionability of cited data[62]. Reverse engineering citations from text is possible through text mining identifiers[63]. The allocation of DOIs to data (and other assets) through DataCite enables bibliographic services to be leveraged.

[60] Nature Editorial "Data's shameful neglect". *Nature* 461, 145 (2009)

[61] Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101

[62] Starr et al (2015) Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

[63] Kafkas et al Database Citation in Full Text Biomedical Articles PLOS (2013)   DOI:10.1371/journal.pone.0063184

- Credit and citation of Research Objects. Systems Biology experiments are compound objects: models, data, SOPs, samples, people structured and linked together. Research Objects enable all the parts of an experiment to be systematically bundled together to form an exchangeable, reproducible and citable unit. The FAIRDOMHub uses the Research Object framework- Figure 4.6 showed a DOI for an entire investigation used to back a publication. The COMBINE Archive format, is a specific kind of Research Object[64]
- Peer review of data, models and SOPs to offer guarantees of quality, putting reproducibility centre stage. Reproducibility of computational research such as systems biology should be a gold standard of quality.
- Getting data and software credits recognised by institution promotion committees and award peer review panels.

Dedicated Stewards and Research Data Engineers, and those Research Software Engineers producing stewardship tools, should be recognised with established and rewarding career paths.

Stewards help to ensure that important digital research data, models and software is adequately safeguarded for future use. Stewards are typically information specialists, archivists, librarians and compliance officers rather than scientists. This is an important role: if data or models have value, someone must manage them, make them discoverable, look after them and make sure they remain usable. However, typically projects and laboratories have at best spare-time, untrained effort and at worst no-one. Large service data centres, such as SIB or the EMBL-EBI, have stewards for public data/models that are in the community public interest. Biocurators have forms their own society (International Society of Biocuration) to advocate on their behalf[65]. Related to stewarding, academic career paths are required for *standards development*.

Research Software (and Data) Engineers A growing number of people in academia combine expertise in programming with an intricate understanding of research. Some Research Software Engineers start off as researchers who spend time developing software to progress their research. Because they enjoy this work and have invested in developing specialist skills, they continue to focus on software and its use in research. Others start off from a more conventional software-development background and are drawn to research by the challenge of using software to further research. Although their combination of skills is extremely valuable, they lack a formal place in the academic system. This means there is no easy way to recognise their contribution, to reward them, or to represent their views. In the UK Research Software Engineers have formed their own society to advocate on their behalf[66].

ISBE's FAIR principles promote the joint making, sharing and reuse of assets, and the recognition for those that undertake these tasks.

This requires ISBE to:

- Advocate for new forms of credit based on contributions (akin to the creative arts industry) rather than authorship;

[64] Bergmann et al (2014) COMBINE archive and OMEX format: one file to share all information to reproduce a modelling project, BMC Bioinformatics 15(1), doi:10.1186/s12859-014-0369-z

[65] http://www.biocurator.org/

[66] http://www.rse.ac.uk

- Allocated persistent identifiers, notably DOIs, to data in public archives and Sys Bio Commons, and implement appropriate bibliometric services. Similarly use ORCID ids for uniquely identifying people. Linking individual researchers to their data and models, and providing persistent links to them, should enable scientists to gain credit for reuse of their datasets and models, encouraging an open, sharing culture
- Support data and model journals until new forms of publishing are established.
- Provide infrastructure and standards for ensuring the reproducibility and validation of Sys Bio outcomes (see 4.1 ad 4.2).
- Advocate policy for rewarding those who share data properly and contribute to standards.
- Reduce the cost of sharing through stewarding services (see 4.5.2).
- Increase the quality of shared content through validation and quality assurance stewarding services and peer review (see 4.5.2)
- Establish partnerships with stakeholders: institutions, funders, publishers, journal editorial boards, learned societies, pressure groups and networks to advocate for the recognition of the skills of asset stewards, research software (and data) engineers and standards makers in promotion and award committees.
- Establish a Software Foundry for Research Software Engineers to share Sys Bio software and practices.

4.4.2 Use Culture

ISBE will expect to support users who are not Systems Biology or Life Science researchers per se and are outside the academic culture. Commercial users, clinical or agri-tech users, citizens, policy makers fall into this category. Here other concerns dominate:

- **Quality and validation**: that the assets are of a high enough quality and have sufficient robustness to be used in the field;
- **Legal access and IP**: that assets are appropriately licensed and IP rights are clear;
- **Entrepreneurship and risk of researchers**: the willingness of researchers to have their assets used by non-researchers. Notable concerns arise about misinterpretation and misuse.

Many of the factors in Academic Culture have an equivalent here. ISBE is required to develop a better understanding of the Use Culture and its impact on asset management.

4.5 Capacity and Capability Building

4.5.1 Skills, training and roles

Stewarding and research asset management provision is not a simple task. It requires knowledge of the research domain; working knowledge of available and appropriate standard formats, ontologies, minimum information models, and community best practice; and access to, and working knowledge of specialised tools, and software, that assist in the stewarding and management processes. It is rare that an individual will possess all skills, and the expertise profiles of individuals will separate them into: curator, steward, manager, producer, and consumer, standards developer.

Table 4.5: The degree of knowledge of specialists roles in a range of areas required for good research asset stewardship and management. 1 star is least knowledge, 3 star is comprehensive knowledge.

Role	Domain knowledge	Format/ontology knowledge	Minimum information models	Community best practice	Knowledge of specialised tools for stewarding.
Curator	***	***	**	**	**
Steward	***	***	***	***	***
Manager	**	*	*	*	*
Producer	***	**	*	*	*
Consumer	***	**	*	*	*
Standards Developer	***	***	***	***	***

As it can be seen from Table 4.5, curators and stewards are knowledgeable within the domain and about standards and minimum information models within the domain. In many cases curators and stewards currently are experienced data and model producers/consumer, and also work within the field of standards development. We expect that dedicated curators and stewards in infrastructures will become less “cutting edge” in their domain specific knowledge, but have a broader knowledge of the field, and standards that can be used for different data and model types. We expect standards developers to still be domain experts, and standards experts. Typical researchers in the field particularly producers and consumers, are expected to have less knowledge about the available standards, minimum information models, community best practice, and tooling.

A key element in getting wider community uptake of ISBE best practice data and model formatting and annotation, will be training. The interdisciplinary nature of systems biology means that, scientifically, the field has many training initiatives in order to enable cross-understanding of the broad fields that comprise it. As an example, the UK has a number of dedicated systems biology centres (Oxford, and Warwick Systems Biology Doctoral Training Centres), which rigorously train the first year systems biology PhD students in subjects such as proteomics, metabolomics, genomics, molecular biology, mathematics, and computational modelling. SysMIC[67] is also a training course that follows similar principles to the doctoral training centres, but provides a collection of experts to deliver online training for early career to late career researchers. The set up of these courses in short blocks of relevant knowledge offer a great opportunity to introduce more specialised stewardship and management aspects to the course.

[67] http://sysmic.ac.uk/home.html

Researchers can also be trained using more specialised workshops, summer schools, and practical courses:

"Bring Your Own Data" BYOD, and capacity building workshops

These are specifically designed to bring curation and stewardship experts into contact with domain experts who have little working knowledge of good practice for data and model management. They aim to teach researchers about the type of standards which are developed and appropriate for their data/models, what tools can assist them in formatting their data/models appropriately, how to annotate their data/models using minimum information models and appropriate software, and what databases/commons resources are available for them to store their published data/models in for future access.

The success of these workshops lie in their *ad hoc* tailored nature: the researcher bring with them the data/models they are actively working on, and the curation and stewardship experts are responsive in their teaching to satisfy the varieties of data and models brought to the workshop. The advantage of these workshops is in the *ad hoc* specialisation as it allows the researchers to gain knowledge that is directly relevant and applicable to their research – increasing the likelihood that the knowledge will be used for further data/models produced by the researcher. Conversely it allows the curators and stewards to better understand the type of data and models that are "cutting edge", and whether standards must be improved to support this, and feed this information back to the standards community. ELIXIR is using the BYOD model to disseminate research asset management knowledge and skills.

Summer schools and practical courses

These provide short, but intense training in a given area of research. They are typically run by a couple of field experts, include invited speakers, and have students that are exclusively selected through rigorous applications. These tend to be funded through initiatives such as EraSysBio, or FEBS, with the explicit goal of upskilling young researchers. Many of these summer schools and practical courses focus on the acquisition of skills within the data and/or model production domain – so specific research skills. This year the International Practical Course for Yeast Systems Biology (ICYSB) in Gothenburg, Sweden, specifically included teaching of data and model management practices. These included standardised formatting and annotation of data/models, versioning, and upload and publication via commons resources. These are ideal locations to expand traditional courses to include some of the core elements of research asset management.

Software Hackathons

Hackathons are events, typically lasting several days, in which a large number of people meet to engage in collaborative computer programming. Examples include the annual BOSC Codefest[68] operated by the Open Bioinformatics Foundation, and platform-specific hackathons for widely used products like Galaxy. Hackathons are a valuable way of training the developers of Systems Biology platforms in interoperability standards (APIs and metadata standards), as well as pump-priming the interoperable services and tools, as well as registering content for registries, for contributing to the ISBE Data Fabric (Section 4.1.2).

Forums and Networks

[68] http://www.open-bio.org/wiki/Codefest_2015

A target for including asset stewardship training are the established networks, forums and initiatives.

- **Networks:** a wide range of networks are available throughout Europe (for example, the Multi-Scale Biology Network in the UK; the FAIRDOM knowledge network)
- **Resource forums**: user and developer forums for Sys Bio resources such as the FAIRDOM user and developer forums, COPASI meetings and so on can be partnered with for co-developing asset stewardship training materials
- **Software Foundry**: is a community effort started by the FAIRDOM initiative to great a forum for developers of Sys Bio tools.
- **Centres and Initiatives:** groups like the UK's Digital Curation Centre provide a further forum for disseminating training materials.

Global Training Initiatives
A target for partnering with training are established training initiatives.

Software Carpentry[69] (SC) was founded in 1998 and is now a Foundation. Well established and highly respected, it supports volunteers to teach basic software skills to researchers in science, engineering, and medicine by running bootcamps, providing open access material for self-paced instruction, and running a training program for people trainers. SC forms a significant plank of the ELIXIR-UK training programme, with plans to work with the wider SC to develop and run bioinformatics-specific SC, extended to Data Carpentry[70] (for Life Science data curation). ISBE should extend this to Model Carpentry (for model curation and making).

GOBLET [71]**(Global Organisation for Bioinformatics Learning, Education & Training)** is a legally registered foundation that aims to provide a global, sustainable support and networking structure for bioinformatics educators/trainers and students/trainees; develop standards and guidelines for bioinformatics education and training; act as a hub for fund gathering and foster the international community of trainers.

IMI EMTRAIN on-course™ is a less comprehensive resource is which contains metadata on more than 6000 training courses.

H2020 Training Initiatives include:

RITrain (INFRASUPP-3-2014) which focuses on the training required to produce skilled managers and operators of research infrastructures. Its primary aim is to establish a training program to allow all RIs to gain expertise on governance, organisation, financial and staff management, funding, IP, service provision and outreach in an international context. It will in effect generate a Masters in Research Infrastructure Management. Whilst this focuses on the skills required for research infrastructure management, it does not provide the skills required for management and implementation of research asset management and stewardship. It would be pertinent to develop similar courses for research asset management and stewardship.

[69] http://software-carpentry.org
[70] http://datacarpentry.org
[71] http://www.mygoblet.org

CORBEL (INFRASUPP-3-2014) which is a pan-BMS RI follow-up to BioMedBridges, incorporating training as well as data management and interoperability. CORBEL aims to make training programmes is more accessible and usable by researchers and make available a network of available trainers. The training programme aims at those who want to learn skills so they can train (train the trainer) or learn skills so that they can use them (classical training).

EXCELERATE (INFRASUPP-4-2014) is the implementation grant of ELIXIR. It has a "training platform" which includes "training the trainers", training researchers and infrastructure providers, and the dissemination of materials and training expertise and events through the TeSS (Training eSupport System) – an aggregator of training resources and repositories

Recommendations for ISBE stewardship training

One of the key objectives of ISBE is to provide systems biology training resources to the community, and ensure that skills are disseminated through all levels of research. Training for stewarding and research asset management should form a key aspect of this for all courses designed and recommended by ISBE.

ISBE should

- Establish a core of ISBE service staff that capable of:
    - Basic data and model management
    - Implementing current ISBE best practice
    - Using the ISBE stewarding services
    - Curation practices
    - Technical tool installation and integration.

  These staff will form a core competency base within ISBE nSBCs.

- Establish and contribute training materials for training using
    - the ELIXIR TeSS training resource
    - GOBLET
    - Other public national and international training repositories (Software Carpentry, Coursera, SysMIC etc).

- Provide training materials for postgraduate curricula for asset stewardship.

ISBE must also look towards maximising the quality and availability of courses by partnering with other Research Infrastructures such as ELIXIR, CaSyM. ELIXIR-UK is leading the coordination of the ELIXIR national node training programmes, and is funded to produce an online training support service. Where appropriate training sources should be reused, and where specialised content needs to be developed the training support content should be developed and updated. ISBE will have particular responsibility for training related specifically to the data management practices in relation to data integration into models, as these practices will not form part of ELIXIR's training scope.

4.5.2 Stewarding services

Technical services for stewards and content generators

Technical services are needed to assist stewards and curators and to support researchers with their own stewardship. Services fall into several categories, including:

- **Stewardship at the point of model and data creation**, to enable greater self-curation, at least for results destined for public archives of experimental reports. For example: smart spreadsheet template tools for structured reporting and metadata annotation of data (e.g. RightField[72], Ontomaton, ISATools); model annotation tools (e.g. OneStop for SBML models); and data management planning tools (e.g. DMPOnline).
- **Automated processes** for curation through automated workflows and specialist analytics[73].
- **Metadata services** such as Ontology services (e.g. Ontology Lookup Service, BioPortal), data to ontology mapping (e.g. Zooma[74]); data driven ontology views; identifier resolution (identifier to entity, concept to entity), name resolution (name to entity, taxonomy mappings), handling different ids in different resources for the same entity, handling id mirrors for distributed resources (e.g. OpenPHACTS IMS, identifiers.org); preservation monitoring tools; parsers for metadata formats; metadata catalogues (e.g. Biosharing.org); model validators (e.g. SBML Validator[75]) and so on.
- **Specialist curation tools for knowledge bases** more will still need to be developed and utilized. ISBE should identify, provision and support such services.

Support services for ISBE brokered projects

When ISBE acts as a broker to bring researchers who generate data into contact with researchers who require data, standards-based and model-compliant data generation must be ensured along with data management planning. Researchers need personal support by skilled stewards to store and explore the links between data, models, protocols and results from ISBE investigations, showing the Systems level details of the experiments, and to understand how separate datasets (e.g. genomics, transcriptomics and proteomics) can be interpreted together, or how they are used for construction or validation of the model, to enable a systems level understanding. ISBE nSBCs will need to be able to deliver professional stewardships services.

4.6 Governance

4.6.1 Legal, ethical and commercial issues

Quite apart from cultural issues that may obstruct data sharing there may be ethical reasons why certain datasets may not be shared; licensing reasons why codes may not be shared; IPR issues for commercial results and legal barriers to recombine data. Even when barriers do not actually exist, ambiguities and misperceptions of the legal and ethical position will deter risk-averse institutions and researchers.

In the case of personal (patient) data, the discipline is well regulated by government, disciplinary bodies, professional societies and institutions. Medical consent forms signed by patients strictly limit what can be done later. The 1000 Genome Project obtained consent from participants for full release of their genomic data, with impressive results. Similar care must be exercised at the beginning of experimental data collection. The benefit of legal and regulatory frameworks lies in clarity, so it is readily apparent

[72] http://www.rightfield.org.uk
[73] Stadelmann et al Toward Automatic Data Curation for Open Data, ERCIM News 100, Jan 2015
[74] http://www.ebi.ac.uk/fgpt/zooma/)
[75] http://sbml.org/Facilities/Validator/

whether and how data may be management, preserved, reused and shared. However, no framework can work around legal prohibition, and where the law is new, untested or ambiguous compliance will force caution.

With respect to the management of ethical responsibilities and norms, researchers will feel more confident about releasing sensitive data is there are established and trusted procedures and services for anonymization, access control etc. Quality control and reproducibility further improves confidence.

Regulation in ISBE is challenging as national and European regulations are at play. The most notable regulation is European Commission's European Data Protection Regulation, which replaces the previous Data Protection Directive. The aim of the new European Data Protection Regulation is to harmonise the current data protection laws in place across the EU member states. The fact that it is a "regulation" instead of a "directive" means it will be directly applicable to all EU member states without a need for national implementing legislation.

ISBE centres must

- Establish clear policies and protocols for the legal issues so that are well understood, and identify and establish ethical data management frameworks, working with funders, professional societies, governing bodies and regulators.
- Establish clear data sharing policies for sensitive data (see 4.3.1)
- Put in place robust frameworks for executing and monitoring policies
- Advise on data collection ethical issues from the start in their projects.

ISBE cSBC must

- Maintain awareness and vigilance with respect to EU and national regulations and compliance mandates.

4.6.2 Economic and business models

Adoption of asset management resources, platforms and technical services must be underpinned by guarantees (as much as anything can be guaranteed) of sustainability. This is also true for training, networking and stewarding services. Sustainability means the long-term securing of resources, usually founded on dedicated funding streams but always. We need to develop a path to support long-term local project archives and the European scale infrastructure. As wide-spread adoption improves chances of sustainability all of activities of WP3 contribute to sustainability, particularly the DMM Network and work with ISBE and ELIXIR.

For example, open source software communities and open knowledge resources like Wikipedia typically depend on in kind contributions from volunteers. However, even these altruistic causes still require cash funding streams for sustaining an active and coordinating core.

Lyon et al propose six interlinked dimensions characterise the funding model for research data infrastructure within a community[76] (see Table 4.6).

[76] Lyon, Ball, Duke, Day Community Capability Model Framework, (2012) https://communitymodel.sharepoint.com/Pages/default.aspx

Table 4.6: The six interlinked dimensions that characterise the funding model for a research data infrastructure within a community.

Projects	**Infrastructure**
Sustainability of project funding • Short-term, quick return • Single-phase thematic investment on a 3-5 year timescale • Multi-phase thematic investment in 5-10 year blocks	Sustainability of funding for infrastructure • One-off investments with no commitment • Slow transition to self-financing • Sustained multi-decade investments in data centres and services
Geographic scale of project funding • Internally or through grants from regional agencies • National funders • International bodies and bi-lateral initiatives between national funders	Geographic scale of funding for infrastructure • Investments by a single funding body at regional or national level • Collaborative development at the national by multiple funders • Collaborative development between international funders
Size of project funding • Small-scale (develop a tool) • Mid-scale (establish a resource) • Major (establish a national capability)	Size of funding for infrastructure • Small-scale tool or resource development • Co-ordinated investment in large or distributed systems • Large central investments in infrastructure, resources or tools

Lyon argues that the degree of centralisation or devolution is key when considering funding models, and this is especially so with a distributed data and model management infrastructure such as ISBE.

- Funding that does not flow centrally but is distributed across PIs or Centres challenges a business model that is centralised;
- Localised resources can run counter to a centralised model of funding.
- Responsibility or ownership of resources may not line up with business models (ISBE will not "own" many of the resources that make up its data infrastructure.

For some aspects ISBE can work with specialist groups - e.g. in the UK, the Software Sustainability Institute (SSI) particularly on sustainability strategies for the software and software training and the Digital Curation Centre (DCC) for data. For long-term data preservation can work with university or national libraries that have a mandate and funding for long-term preservation of digital goods – e.g. ETH Zurich where the ETH library has setup a group for digital data preservation and a current national research infrastructure program (CRUS P-2) has the strategic goal of creating national services in this area. Business plans need to calculate return on investment.

Sustaining what?

ISBE's data infrastructure is made up of a mixture of stakeholders, resources and activities, each with potentially a different sustainability strategy.

- ISBE selected public archives
- ISBE Sys Bio Commons

- ISBE endorsed catalogues and technical support services
- ISBE endorsed software platforms and affiliated tools
- Metadata specifications and templates and ontologies.
- Network community.
- Training programme and materials

Issues include responsibility, delegation, contribution guarantees and productivity/return on investment by contributors. It also depends on whether sustainability is a criteria of ISBE endorsement for contributed public archives. Here we distinguish between resources coordinated by ISBE and resources natively developed by ISBE.

Sustainability Structures

nSBC “node” contribution guarantees

Example: ELIXIR, SyBIT

ELIXIR operates on a National Node contribution model. Resources are contributed by nodes as part of the ELIXIR Infrastructure. These contributions are backed by the node to be sustained and available to the infrastructure, through institutional backing, national backing by funding agencies or, sometimes, pan-national backing. A certain track record of sustained funding and good prospectus of same is needed before a resource or service can be contributed through a formal process.

In SyBIT, Swiss universities have set up permanent support entities as core facilities.

Return on investment by contributor: standing in the community and the potential of new funds through association with the Infrastructure.

ISBE: this model translates to the responsibility for sustainability of resources, services and activities borne by nSBCs, and coordinated by cSBC. ISBE will need to work with national funding partners to establish national services based on a network of institutional facilities or contribute to the resources and with national universities/centres who make up the nSBCs to establish core facilities.

Institutional contribution guarantees

Example: HITS have promised a 10 year guarantee of long-term preservation for the FAIRDOMHub for the ERANet EraSysAPP programme, assured by renewal of servers and funding of a minimal service, to secure hosting and reliable access to support the policies and management mandates of the funding agencies. HITS also support SABIO-RK.

Return on investment by contributor: standing in the community and the potential of new funds through association with the Infrastructure.

ISBE: this model translates to the responsibility for sustainability of resources, services and activities borne by institutes associated with nSBCs , and coordinated by cSBC.

Partnerships with other RIs

Example: ELIXIR MoU/SLA

Other established RIs guarantee or partially sustainability of resources, services and activities of interest to ISBE.

Return on investment by contributor: shared costs and responsibility borne by both parties.

ISBE: this model translates to a SLA with the RI and guarantee by ISBE to bear shared resourcing.

Public-Private Partnerships
Example: Dutch TechCentre for Life Sciences
DTL as a nationwide platform is organised as a public-private partnership: the DTL Alliance, facilitated by the Stichting DTL. DTL is open to universities, university medical centres, universities of applied science ('HBO'), public or private research institutes and companies.

Return on investment by contributors: shared costs and responsibility borne by all parties; preferential consultancy and training rates and access; customised resources; standards and international influence.
ISBE: this model translates to nSBCs becoming PPPs.

Not for Profit Foundations
Examples: APACHE Foundation; tranSMART Foundation; Open PHACTS Foundation; HubZero; VIVO; iPython Foundation; Software Carpentry Foundation; numFOCUS Foundation[77].

Foundations are common for open source software and resources. The foundation forms a legal entity into which IP and funds can be channelled through awards or donations, and membership governance can be developed. Nevertheless, funding streams still need to be found and establishing and running a foundation has a cost. Most Foundations levy membership fees.

Return on investment by contributors: joint legalised ownership and pooled IP; preferential consultancy and training rates and access; customised resources; standards and international influence; shared costs; access to international funding streams (PIC code for EU projects, similar for USA).
ISBE: alliance with established Foundations (e.g. Software Carpentry, numFOCUS, APACHE) and establishing Foundations for specific resources (e.g. FAIRDOMHub).

Volunteerism and in Kind

Examples: open source software, wikipathways

Volunteerism and in kind contributions are prevalent in the software open source community and freely available. Generating a groundswell of contributing developers rather than just user-developers needs work on outreach and greater open development organisation and governance. Nevertheless, core development/contribution/curation will be shouldered by a core team that needs funding, and volunteers are a hard cohort to manage and guarantee.

Return on investment by contributors: joint sense of ownership; public credit profiles; fame and love.
ISBE: build a community of in kind contributors and volunteers around selected resources and activities with strict contribution protocols combined with clear reward, governed and supported by nSBCs.

Sustainability Resource Streams

National level asset management levies
Example: NWO and DTL; SystemsX in Switzerland

[77] http://numfocus.org/

DTL have negotiated a 5% "levy" from all NWO grants to contribute to centralised data stewardship handled by DTL. DTL are the institute upon which the Dutch ELIXIR node and FAIRDOM facility have landed. The

Grants

Examples: National grants, EU grants, e.g. ELIXIR EXCELERATE and CORBEL

An obvious route to funds are national, European and international funds. By establishing foundations or partnerships members can bid to any available source regardless of nationality. The down slide is that cashflow is not guaranteed; peer review is a lottery and funds are periodic and short-medium term. Deliverables in projects may not align with ISBE. ISBE would be relying on the ability of nSBCs and resource/service providers to win awards. Many grants will be partnerships between multiple groups.

Partnerships, Joint Ventures and Foundations

Examples: Public Private Partnerships (DTL), JV (UK's Alan Turing Institute), Foundations (Apache)
Partnerships have annual membership costs in return for some sort of benefit. Costs are tensioned against sector types (commercial, academic, independent).

Subscription / licensing / fees

Examples: an annual subscription for users of public resources or Sys Bio Commons; data / model licensing; licensing support and software maintenance contracts for software platforms; curation support services fees for researchers or publishers; rent-a-feature access to special facilities; rent-a-modeller brokering (similar to ScienceExchange); charges to commercial tool makers to create bespoke plug-ins to platforms.

Contracts

Examples: Publisher companion sites; National funding council CRIS
ISBE enters contractual arrangement for novel resources such as the Commons.

An ERIC or CA

Example: BBMRI, ELIXIR
ISBE have a work package dedicated to developing the legal and financial framework for a distributed data/model management infrastructure which will release national financial commitments, some of which should be used for data management.

Selected Bibliography

- Pryor G (ed) Managing Research Data facet Pubishing, ISBN 978-1-85604-756-2
- ERCIM News 100 Special theme: Scientific Data sharing and reuse
- Borgman CL Big Data, Little Data, No Data, MIT Press ISBN 978-0-262-02856-1
- Riding the Wave: How Europe can gain from the rising tide of scientific data (Oct 2010)
- Lyon, Ball, Duke, Day Community Capability Model Framework, (2012) https://communitymodel.sharepoint.com/Pages/default.aspx

Selected Web Sites

- Research Information http://www.researchinformation.info/
- Data Preservation Coalition http://www.dpconline.org/
- Alliance for Permanent Access http://www.alliancepermanentaccess.org/
- Archive Team http://archiveteam.org/index.php?title=Main_Page
- Digital Curation Centre http://www.dcc.ac.uk
- Software Sustainability Institute http://www.software.ac.uk