ISBE Infrastructure for Systems Biology Europe

v 28-01-2014

ISBE WP2 report

Deliverable No: D2.2

Preliminary Recommendations for a Data and Model Management Framework

Data and Model Stewardship for the ISBE Infrastructure.

January 2014

Carole Goble, UNIMAN
Katherine Wolstencroft, UNIMAN/UL
Natalie Stanford, UNIMAN
Jacky Snoep, UNIMAN/Stellenbosch
Renate Kania, HITS
Martin Golebiewski, HITS
Wolfgang Mueller, HITS
Sarah Butcher, IC
Nicholas 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 18
Actual date of delivery	Month 19
Deliverable number	D2.2
Deliverable title	Preliminary 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 Stanford (UNIMAN) Jacky Snoep, (UNIMAN/Stellenbosch), Renate Kania (HITS) Martin Golebiewski (HITS) Wolfgang Mueller (HITS), Sarah Butcher (IC), Nicholas Le Novere (Babraham/EBI).
Project coordinator	Richard Kitney
EC Project Officer	Andreas Holtel

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.

Table of Contents

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[1] inspires confidence among the research councils and funding bodies that invest in the work. For stewardship to be effective the technical, social, and educational aspects of its implementation must be well managed.

Technical aspects include: how data and models 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; and how data and model management can become embedded in Systems Biology practice and publishing.

Educational aspects include: educating existing and new Systems Biologists in data and model management; and educating other stakeholders such as funders, librarians and publishers in the importance of data and model management.

[1] https://peerj.com/articles/175/

Objectives

- ➔ To synthesise the findings of *D2.1: Combined report on state of the art and horizon scanning* into a set of recommendations of what data and model management must look like within the ISBE framework in order to meet the future needs of the community.
- ➔ To identify the impacts these recommendations would have upon potential stakeholders of ISBE model and data management.
- ➔ Evaluate the risks that would be associated with such a framework.

Methodology

The foundations of this deliverable lie within the sister 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.

The document was also assembled through close interactions with other work packages within ISBE, 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 and model 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.

Summary recommendations for ISBE stewardship.

The ISBE framework will not dictate a single platform or a tightly integrated data infrastructure. Rather it will focus on: *conventions*[2] that enable data interoperability and stewardship and *compliance*[3] against data and metadata standards, policies and practices.

ISBE infrastructure is separated into three types of centres:

1. Modelling and Integration Centres (MICs); which allow planning of experimental design with a view to model construction, and model generation/simulation.
2. Data Generation Centres (DGCs); which generate Systems Biology standard data.
3. Stewardship Centres (SCs); which ensure provenance, usability, interoperability and longevity of data and models.

[2] A convention is a set of agreed, stipulated, or generally accepted standards, norms, social norms, or criteria, often taking the form of a custom.

[3] Compliance means conforming to a rule, such as a specification, policy, standard, laws or regulations.

The exact functionality of the ISBE centres is still to be agreed, but based upon envisage interactions between ISBE centres (Figure 1) we have firstly proposed centre based recommendations for stewardship. Then cross cutting recommendations for stewardship.

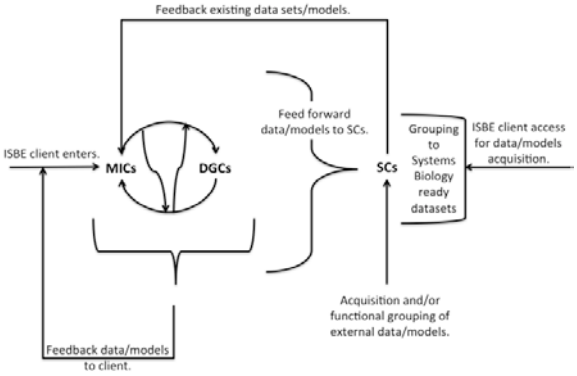


Figure 1. Within the ISBE framework MICs (Model Integration Centres) and DGCs (Data Generation Centres) will be producers of ISBE born models and data. This data will pass into the SCs (Stewardship Centres) where it will be managed over the long term, and made accessible to the public in Systems Biology sets (including both models and data). The SCs will feedback information from current Systems Biology sets and models available so that MICs and DGCs do not reproduce assets unnecessarily.

Stewardship recommendations for DGCs

Must:

- [Defining Conventions and Standards] Work with SCs to ensure standards for structuring and annotation data are relevant for how the data will be used, and as experimental protocols change.
- [Defining Conventions and Standards] Train all relevant staff how to correctly structure and annotate their data, so that this occurs as close to production as possible.
- [Defining Conventions and Standards] Take an active role in defining SOPs to be held centrally within ISBE.
- [Stewardship Services] Use the agreed formats from SCs to structure and annotate ISBE born data (specific to the data types).

Should:

- [Defining Compliance and Process] Develop automation of data structuring and annotation at the collection source.
- [Defining Compliance and Process] Make the prepared data immediately accessible to all ISBE centres.
- [SLAs with RIs] Make the structured and annotated data available to other ESFRIs as single sets.

Could:

- [Stewardship Services] Select important published data and structure and annotate to ISBE standard for use within ISBE.
- [Stewardship Services] Structure and format datasets non ISBE born as a service.

Will not:

- [Defining Conventions and Standards] Dictate what formats data submitted to ISBE must be in, only recommendations will be provided.
- [Defining Conventions and Standards] Be responsible for the quality of data submitted to ISBE, only that of ISBE born data.

- [Stewardship Services] Format data submitted by the public, unless the impact of the data will be high.
- [Defining Conventions and Standards] [SLAs with RIs] Release data to other centres, customers, or ESFRIs unless it has adequate annotation.

Stewardship recommendations for MICs

Must:

- [Defining Conventions and Standards] Work with SCs to ensure standards for structuring and annotation of models are relevant for how the models will be used, and as modelling complexity evolves.
- [Defining Conventions and Standards] Make descriptions of structure and annotation best practice available to the public.
- [Defining Compliance and Process] Cross-link all ISBE born models to the derived datasets that were used to produce them.
- [Stewardship Services] Maintain a database cataloguing the structure of models submitted by the public. This should be used to ensure that associated software (e.g. virtual machines) is available for running the model.

Should:

- [Stewardship Services] Survey available models in order to identify suitable models for formatting and inclusion of ISBE data. (what is meant by this – explain better)
- [Defining Conventions and Standards] Provide a list of preferred softwares for producing model.

Could:

- [Stewardship Services] Offer restructuring and curation of user models as a service so that models can be stored in a preferred ISBE format (this would not guarantee quality, just longevity).

Will not:

- [Defining Conventions and Standards] Take responsibility for the quality or usability of models submitted by the public.
-

Stewardship recommendations for SCs

Must:

- [Defining Conventions and Standards] will develop and agree on standards and formats for exchanging data and models etc. between centres and for ISBE clients based upon the current standards in use within the Systems Biology community.
- [Defining Conventions and Standards] SCs will contribute to the standardisation of formats, ontologies and minimum information checklists.
- [Defining Compliance and Process] will review and evaluate data management processes and standards within ISBE to ensure that ISBE processes continue to run smoothly and new types of data or modelling approaches can be integrated with little disruption.
- [Defining Compliance and Process] must anticipate a mixture of open and commercially sensitive data/models and open and commercial services. There is an expectation that 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. It must also anticipate potential financing as a public private

partnership and the implications this may have on data visibility – its accessibility and accessibility. The extent and under which operating conditions that ISMB should support private and proprietary data needs to be clarified.

- [Stewardship Services] will be responsible for and provide digital curation services for linking data, models, maps, SOPs and experimental descriptions, linking to individuals and/or organisations to ensure credit is awarded to creating scientists. This will ensure that associations between these components are preserved and different versions are recorded properly.
- [Stewardship Services] will provide services for collecting and curating data, models, maps, SOPS, experimental descriptions etc in standards compliant forms. Such services will seek to avoid as much disruption to the scientists working practices as possible, and ensure proper implementation of standards.
- [Stewardship Services] take responsibility for and provide services for transforming relevant existing data sets to formats that can be used in systems biology projects.
- [Stewardship Services] take responsibility for and provide services for digital preservation or migration of ISBE-related data, models and maps to established archives and repositories and develop policies for accessing and using resources.
- [Stewardship Services] take responsibility for and provide services for digital preservation or migration to established archives and repositories. Publishing or deposition of supplementary materials for publications.
- [Stewardship Services] Use discovery services for finding: (people) modellers, software, data, SOPS and models, and more. Monitoring services are needed to identify which resources are used.

Should:

- [Defining Compliance and Process] devise and review compliance with: management and preservation processes; annotation and curation standards; access and responsibility policies; and quality control. It is the Stewardship Centres' responsibility to facilitate: access, archiving, annotation and discovery through portals, cataloguing and indexing, programmatic interfaces etc.
- SCs should host data and model management facilities at the International, National and Centre level supporting selected Sys Bio data and model services and resources such as catalogues, libraries, model repositories, and key data repositories. Such services/resources should be selected and technically reviewed by ISBE as adhering to: the ISBE data interoperability conventions; a prescribed level of quality; compliance with quality of service and metadata standards. They will be scientifically reviewed as to their importance and sustainability prospects. The "certification" criteria of data and model resources and services must be defined. The "certification" process must be defined. Certification should complement that of ELIXIR.
- [Stewardship Services] seek to support hosted data and model management facilities for System Biology private, project and laboratory clients, as a "cloud" service, with suitable access

permissions and backed by scalable computational infrastructure4.

- [Stewardship Services] SCs should recommend and catalogue, and possibly certify and support, data and model management software platforms that can be deployed privately by clients.

Stewardship recommendations for ISBE, as a whole.

Must:

- [Defining Compliance and Process] Make descriptions of the structure and standards available to the public for data, models, and SOPs.
- [Defining Conventions and Standards] The conventions for data and model metadata descriptions be founded on community standards for: identifiers, formats, checklists and vocabularies.
- [Defining Conventions and Standards] Offer consultancy and advice for standards and formats for ISBE clients (that is ISBE participants - guidelines for complying with the ISBE data sharing policy), and for ISBE centres (that is standard operating procedures for ISBE centres)
- SCs will make recommendations for tools and resources to assist with standards compliance, lower the barrier to adoption and to make standards-compliance more efficient and less time-consuming.
- [Stewardship Services] Will ensure all ISBE born data and models are standards compliant and annotated with the correct metadata.
- [Stewardship Services] offer advice with producing data management plans for funding proposals, ensuring consistency of data and models that become ISBE resources and further use of ISBE facilities.
- [SLAs with RIs] ISBE make SLAs with key **domain related RI**, notably the ERANets, ELIXIR, Euro-Bioimaging, BBMRI, and the IMI.
- [SLAs with RIs] ISBE must make SLAs 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.
- [Consultancy and Training Service] Train customers in best practice for data and model management according to the most up-to-date agreements.

Should:

- [Defining Conventions and Standards] The ISBE framework focuses on *conventions*[5] that enable data interoperability and stewardship and *compliance*[6] against data and metadata standards, policies and practices The conventions for data and model services interoperability should be based on the internet and web's minimal "hourglass" approach[7], a specification of lightweight interfaces, standard protocols and standard formats.
- [Consultancy and Training] Provide consultancy to co-develop format/annotation/cross-linking/storage requirements for

[4] It is not in the remit of ELIXIR to provide such a data/model hosting facility.

[5] A convention is a set of agreed, stipulated, or generally accepted standards, norms, social norms, or criteria, often taking the form of a custom.

[6] Compliance means conforming to a rule, such as a specification, policy, standard, laws or regulations.

[7] This is usually called the hourglass model but that terminology is likely to cause confusion in ISBE.

research groups, journals, funding councils needs.

- **[Consultancy and Training]** Provide training for co-develop format/annotation/cross-linking/storage requirements for research groups, journals, funding councils needs.

Will not:

- **[Defining Conventions and Standards]** The framework will not will not dictate a single platform or a tightly integrated data infrastructure for users.
- ISBE infrastructure is a *set of services* to support the stewardship of ISBE data/models, access to ISBE data/models and the technical compliance of data/models against metadata standards, policies and practices. ISBE should not govern the science or scientific methodology that is undertaken using its infrastructure. That is the purview of peer review.

Visions of how stewardship recommendations will impact stakeholders.

Vision 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 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.

Recommendation: The raw data is collected, structured, and annotated according to available and agreed SOPs in two of the ISBE DGCs. The raw data is then stored in an embassy cloud, to be accessed and post-processed by the third DGC, 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 with the MICs 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.

Vision 2: 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.
Recommendation: 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.

Vision 3: 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.
Recommendation:
The Journal can work directly with SCs in order to turn the requirements into a functional set of formats and annotations for authors to follow. DGCs, MICs and SCs can train staff from the journal in data and model curation, submission and interlinking. ISBE can provide 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 can 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 and models 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.

Vision 4: 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.

Recommendation: The NRC can consult 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.

ISBE Infrastructure for Systems Biology Europe

Introduction and Context

ISBE is an infrastructure that should be as inclusive and as flexible as possible, and support researchers in all types of systems biology. In particular it 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. 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 Sys Bio researcher is not ISBE's concern.

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[8] and Coveney[9] 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. The UK House of Lords recently went as far as to state that in fact infrastructure was software and that storage/compute facilities were consumables[10], a sentiment echoed in funding council's roadmaps[11]. The sustainability and maintenance of data and model management software is thus crucial to ISBE infrastructure.

Provisioning a common framework for the ISBE centres and ISBE client will enable data and models arising from the ISBE infrastructure to be retained and managed. Adopting a common framework and standards will enable the exchange of data and models between ISBE centres and will allow scientists to (a) support the reproducibility of results and (b) 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 is a mixture of 'distributed' resource and a 'virtual' resource: with distributed *centres* and virtual access to data by *clients*.

---

[8] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32499/12-517-strategic-vision-for-uk-e-infrastructure.pdf

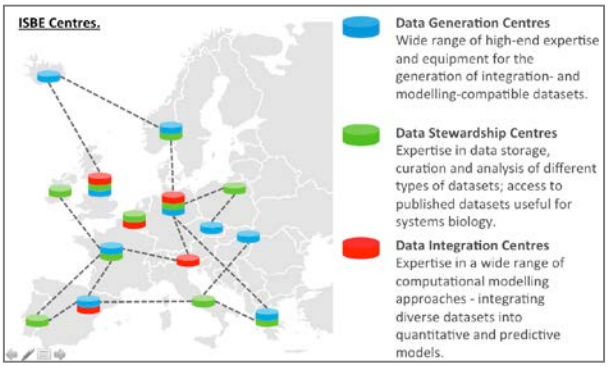
[9] http://wiki.esi.ac.uk/w/files/f/f5/ResearchComputing-glossy.pdf

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

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

ISBE Centres

ISBE proposes an infrastructure based on a matrix of functionally interconnected expertise, distributed throughout Europe, and separated into three separate types of functionality (*WP3 report: Structure and functioning od the ISBE infrastructure*). The centres and associated functions and responsibilities, in brief, are:



Data Generation Centres (DGCs):

- Produce (quantitative) data sets according to the needs of scientists that are fit for model building.
- Are responsible for the SOPs, for acquisition, handling and formats of data, and their harmonisation with non-ISBE data generation centres.
- Advice in the experimental design phase about data storage and analysis.

Stewardship Centres (SCs):

- Provide a unified view over all resources generated and used in ISBE, in the context of the experiments that produced them.
- Allow researchers to explore the links between data, models, protocols and results from ISBE investigations, showing the Systems level details of the experiments.
- Allow scientists 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.
- Organise the gathering of relevant information from literature, databases and from unpublished sources.
- Define, develop and adjust criteria and standards that must be met by data, maps, tools and models, in conjunction with DGCs and MICs.
- Check on the accuracy, reliability and quality of data, models, tools and maps.
- Overall SCs will make the re-use of data sets, models, SOPs etc. possible in future projects.
- The SCs are the prime, but not the sole, place of data and model management.

Modelling and Integration Centres (MICs):

- Will be the primary integrators of multi-type data into models. This is an essential part of Systems Biology research. This integration relies heavily on DGCs and SCs providing standard formats and interfaces for access, storage and exchange.
- Support clients in model based simulation.
- Generate new models based on available datasets.

The centres are to interact using a single point of entry for clients, with different modalities for requested data/model services (Figure 1), although data and models stored within ISBE can be accessed remotely.

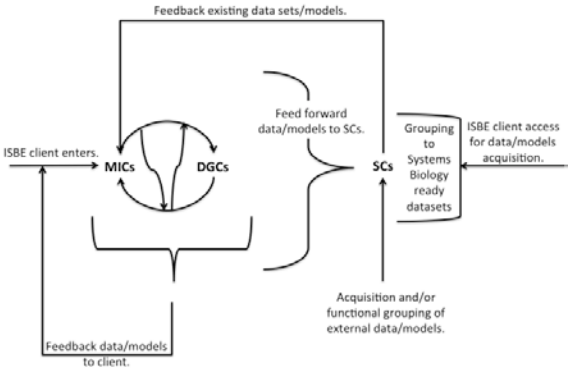


Figure 1. Connection within the ISBE Framework. Within the ISBE framework MICs (Model Integration Centres) and DGCs (Data Generation Centres) will be producers of ISBE born models and data. This data will pass into the SCs (Stewardship Centres) where it will be managed over the long term, and made accessible to the public in Systems Biology sets (including both models and data). The SCs will feedback information from current Systems Biology sets and models available so that MICs and DGCs do not reproduce assets unnecessarily.

ISBE Clients

ISBE clients are foreseen to be from a large breadth of the research community, WP3 have highlighted key clients as: researchers from academia (both novices and experts); industry (SMEs and large); and non-scientists (funders, policy makers, politicians, publishers, digital libraries, patient organisations, press etc.). These clients may approach ISBE as stand-alone clients, or as large, multi-site and multi-partner collaborations.

The ISBE infrastructure is to organise investigation at a pan-systems biology level rather than a pan-project level. This allows small, specialised research groups, individual 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 like virtual liver[12]. Currently this is not very likely due to the range of skills a small group or individual would need access to, but currently does not have or cannot acquire.

ISBE Data and Model Stewardship Ecosystem.

The proposed ISBE structure, and the envisaged breadth of ISBE clientele, form fundamental aspects of the complex data and model stewardship system (sketched in Figure 2). In addition to this, stewardship is impacted by:

- **Standards and formats** in use within the community for describing and exchanging data and models between centres, for ISBE clients and for interoperability and intelligibility.

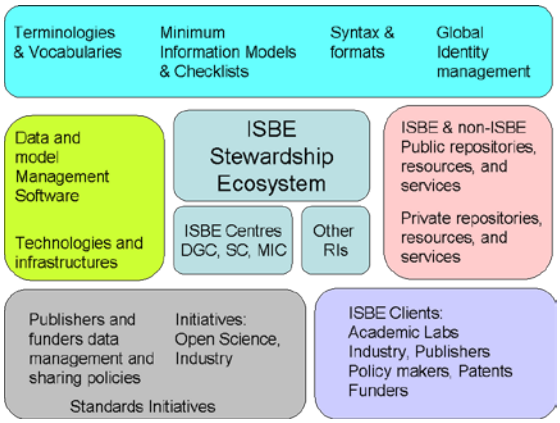


Figure 2: ISBE Data and Model Management Ecosystem.

[12] http://www.virtual-liver.de/wordpress/en/

- **Data and model management software and tools:** including the technologies and infrastructure that they use.
- **The centre based structure of ISBE.** In order to ensure smooth operations the centres must be able to transfer structured and annotated data between themselves without loss of information.
- **Interaction with other Research Infrastructures:** including ensuring that standards used are cross-compliant at the point of data transfer.
- **Repositories, resources and services for data and model stewardship:** hosted by ISBE Centres, by non-ISBE RIs, independently or by clients; most will be in the public domain others may be private
- **The sources and consumers of data and models:** Centres, ISBE Clients and other RIs such as ELIXIR or those outside Europe.
- **Policy, community and standards initiatives of funders, publishers, institutions and other stakeholders governing stewardship**: including access, preservation and the sustainability of software, resources, infrastructure and stewardship services.

Recommendations ISBE functioning within the ecosystem.

- ISBE infrastructure is separated into three classes of centres. However, data and model management is cross-cutting and is *the responsibility of all ISBE centres* (specific responsibilities as described in the Digital Curation Centre Model – Figure 4). The responsibilities of each type of centre must be clarified and coordinated.

- The ISBE framework focuses on *conventions*[13] that enable data interoperability and stewardship and *compliance*[14] against data and metadata standards, policies and practices. We propose that the conventions for data and model services interoperability should be based on the internet and web's minimal "hourglass" approach[15], a specification of lightweight interfaces, standard protocols and standard formats. We propose that the conventions for data and model metadata descriptions be founded on community standards for: identifiers, formats, checklists and vocabularies. The framework will not dictate a single platform or a tightly integrated data infrastructure.

- ISBE infrastructure is a *set of services* to support the stewardship of ISBE data/models, access to ISBE data/models and the technical compliance of data/models against metadata standards, policies and practices. ISBE should not govern the science or scientific methodology that at undertaken using its infrastructure. That is the purview of peer review.

- 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.

[13] A convention is a set of agreed, stipulated, or generally accepted standards, norms, social norms, or criteria, often taking the form of a custom.
[14] Compliance means conforming to a rule, such as a specification, policy, standard, laws or regulations.
[15] This is usually called the hourglass model but that terminology is likely to cause confusion in ISBE.

- Data (and possible models) are *stratified* on several dimensions: open/closed; commercial/non-commercial; secure/public; along scientific lines; ISBE/client; ISBE/other RI. This stratification must be handled by the SCs.

- There is a distinction between (a) ISBE public data/models for which it is primarily responsible for stewarding and (b) ISBE infrastructure for data which it enables and hosts. In the first case there will be ISBE public datasets and models, badged as ISBE, which ISBE takes responsibility for. In the second case lies data and models generated by projects supported by the ISBE infrastructure and training. This distinction needs to be clarified.

- ISBE will support life sciences research, health research and commercial collaborations in these areas. Therefore, clear policies and standard operating procedures are required in ISBE to ensure that private health care information or commercial assets are kept with secure and restricted access.

- 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.

- Data and models must be citable and credit given to their authors and stewards.

- Data and models will be commoditised so that they can be re-used modularly.

Interactions with other Research Infrastructures

Steward Centres 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. The Key EU-RIs are as follows:

ELIXIR

ELIXIR[16] 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.

[16] http://www.elixir-europe.org

ELIXIR's role is primarily the compliance and governance of infrastructure chiefly provided by the nodes. The infrastructure activities are:

(a) 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, such as the UK's training node. All nodes are balancing the need to serve their nation and the requirement to contribute to the overall ELIXIR community.

(b) Pilot projects funded by contributions from nodes and the hub, between nodes and between the node and the hub. First round pilots of interest to ISBE's data and model management ELIXIR pilot projects are indicative of ELIXIR's plans:

- *Safeguarding resources*. Establishing the EGA as a Joint Venture between several ELIXIR nodes.
- *Private, virtual workspaces in the ELIXIR data infrastructure*: ELIXIR-Facing Cloud Support and Virtual Machines. Scientists often wish to compare their research results with large reference datasets, but do not have the capacity to download or manage such massive files locally. The ELIXIR-facing cloud will allow researchers to create a virtual working environment right next to the reference data, with seamless access through their host institute.
- *Seamless, uninterrupted transfer of major datasets across Europe*. Bioinformatics resources can contain several petabytes of data. Europe's Research and Education Networks, including JANET and GEANT, have upgraded the physical infrastructure to allow for dedicated, secure and private transfer of data between European institutes over allocated lines in a timely and predictable manner. The pilot is for the transfer of major European Genome-phenome Archive (EGA) datasets between the UK and Finland.
- *Secure access to genomic data through distributed authentication*. Researchers apply for authorisation to EGA using their host institute credentials, streamlining the process for managing account information and adding an extra layer of accountability. This pilot is to support Data Access Committees with electronic application tools and is endorsed by Geant3Plus.
- *Interoperability* of protein resources for drug discovery. This pilot is for the 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.

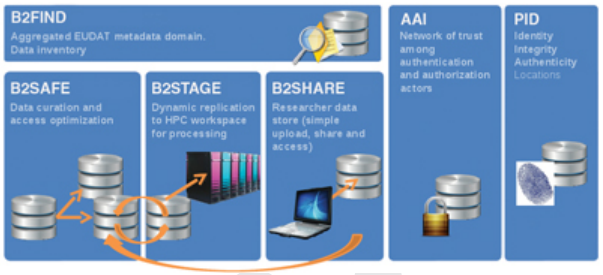
(c) Seven workstreams are defining the Programme of Work of ELIXIR to be delivered by cooperating nodes in partnership with international initiatives and other RIs. The workstreams are:

- *PoW1: Data resources and services*: which (small number) of data resources and services will be "certified" as ELIXIR responsibility. Lead: Ron Appel (CH), Rolf Apweiler (EBI)
- *PoW2: Tools Interoperability and Service Registry*: how will ELIXIR define tool (interface) interoperability, compliance, get tools to interoperate and run a Life Science registry (though BioMedBridges are already doing this). Lead: Bengt Persson (SE), Søren Brunak / Peter Løngreen (DK)
- *PoW3: ELIXIR Technical Services*: cloud storage, compute services, access services, data transfer, locality of data to compute and moving compute to data. Lead: Tommi Nyrönen (FI), Lurek Matyska (CZ)
- *PoW4: Data interoperability, vocabulary and ontology services*: how will ELIXIR define data interoperability, define and adopt metadata annotation standards, manage and motivate compliance, acquire annotation tools, run a Life Science metadata standards registry (with BioSharing). This is directly relevant to ISBE data and model management. Lead: Barend Mons (NL), Carole Goble (UK), Jaak Vilo (EE)

- *PoW5: ELIXIR Training Programme*: a coordinated training programme for the ELIXIR community to use the infrastructures and spread best practice across the full spectrum of players, from developers of tools to bench/bedside scientists. This includes training for data management (data Carpentry) and software development (Software Carpentry): both are relevant to ISBE. Lead: Chris Ponting (UK)
- *PoW6. ELIXIR Domain Specific Services*: Lead: Alfonso Valencia (ES), Inge Jonassen (NO), Jose Leal (PT)
- *PoW7. ELIXIR Management and Operations*. Lead: Niklas Blomberg (ELIXIR)

EUDAT

EUDAT[17] 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 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.

[17] http://www.eudat.eu

- 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.
  EUDAT has begun work on two further services: (1) B2DROP to synchronize file systems with a central store and (2) 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

OpenAIRE[18] 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.

Cross EU-RI Synergies

Work package 5 outlines the synergies anticipated with other EU RIs, we have replicated these in Table 1 for convenience.

[18] http://www.openaire.eu

Table 1. 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; Secure handling and access to data. Data carpentry training.	Tools interoperability; Tools interoperability training Kite-marked tools; Data mining and analytics services.
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.	

Non-European RIs for data and model management

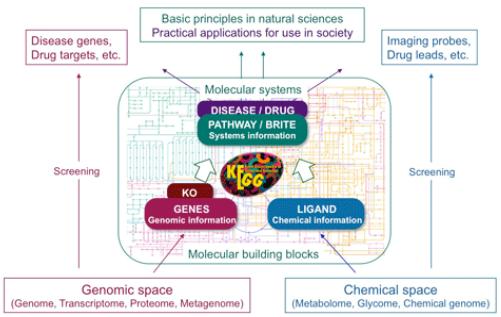
Non-EU RIs are also relevant to for understanding ISBEs place as in infrastructure world-wide. ISBE must compliment the work of other infrastructures on the global scale in order to be relevant to the Systems Biology community. Key non-EU RIs include:

KBase[19]

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[20]



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.

[19] http://kbase.vectorworks.net/
[20] http://www.genome.jp/kegg/

It allows access to information:

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

PMR[21]

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.

NIH BD2K[22]

NIH BD2K is a 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

Recommendations

- 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.

[21] http://www.cellml.org/tools/pmr
[22] http://bd2k.nih.gov/#sthash.FJFonZHS.dpbs

- 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.

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

Data, Models and SOPs in Systems Biology

The ISBE D&M infrastructure needs to be able to provide services which fully support each stage of the systems biology process. Systems biology research operates as a continuous cycle where experiment informs model, and model informs experiment, as shown in Fig. X. 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.

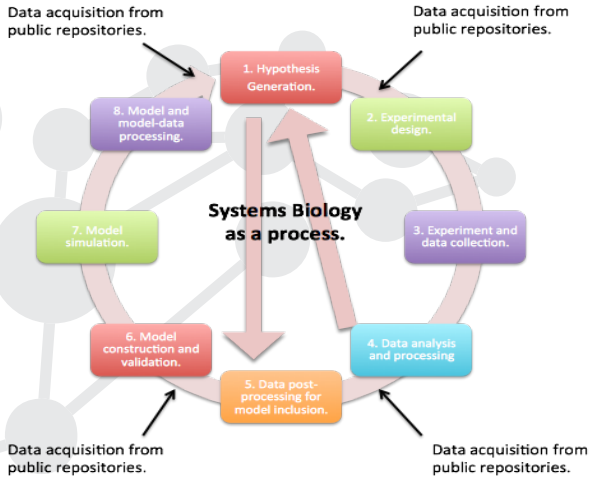


Figure 2 Systems biology as a process.

A D&M management e-infrastructure 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 Sys Bio projects has only been made feasible by introducing bespoke platforms for inter-project data handling (e.g. SEEK[23]).

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 are presented below. We do not propose that all of these steps will be performed by the data stewardship centre,

23 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

rather that the stewardship centre will perform a certain set of these tasks, and advise other centres and customers on how to perform some of the steps themselves.

1. Hypothesis generation.

- Relevant models will need to be accessed so that current understanding of mechanisms and phenomena for a given behaviour can be identified and understood. The model must therefore 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.
- Relevant data needs to be identified and parsed to identify whether there is any support/opposition to the hypothesis.
- Literature that supports the ideas (suitable linking between stored data/models and publication IDs).

2. Experimental design.

- Need to identify 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.
- Relevant SOPs should be easily identifiable 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.
- There should be 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 and data collection.

- Where new experiments need to be designed, new SOPs should be generated.
- Raw data, handled by SC, DGC and other RIs, but it is typically large and difficult to handle (see properties of data). The centres must post-process the data in to sharing format for public release, and general interface management for ISBE.
- Raw data that will come from experiment and how it will be handled/stored.
- We expect data for proteins metabolites, genes, transcripts, kinetics and microscopy data to be produced. This must be available for customers to view in a unified-model centric format.

4. Data analysis and processing.

- SOPs relating to the data analysis need to be identified. 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.
- 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 maybe 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 identify 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, SOPs.

6. Model construction and 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.
- The same model can be parameterised with different data, and released as a separate model and publication, this is especially prevalent in metabolic modelling. Need to unify these models.
- The production of current models is usually for answering a specific question. This procedure forces the model to be valid under a very limited set of conditions. To move towards more robust modelling constructed models need to be subject to parameter sweeps to test robustness.
- Publication of a model is not the end of the process. The model needs to be verified and validated before moving on to steps of replication and reproducibility. [Model life-cycle]
- Do we retain models as an example of the output for the paper/experiment/hypothesis etc. or do we want to organise them so that we can re-use them.
- The model needs to be tested for under-/over-fitting,

7. Model simulation.

- The model simulations need to be compared with the data sets that were used to construct it. 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 and 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 static-link identifiers.

Stewards, Curators and Custodians

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. They are not stewards of private lab data.

Recommendations

- ISBE must put in place skilled digital curators and effective curation lifecycle management for its own data and the infrastructure, and training and services for its clients. The SCs stewardship responsibility for client data must be clarified further.

- ISBE must advocate for stewards to be funded by funders, credited by researchers and have career paths in institutions. ISBE should work with bodies such as International Society for Biocuration, ISCB, The Digital Library Federation, and Software Sustainability Institutes such as SSI in the UK to enable stewards to be first class citizens.

Data, Models and SOPs Stewardship Framework

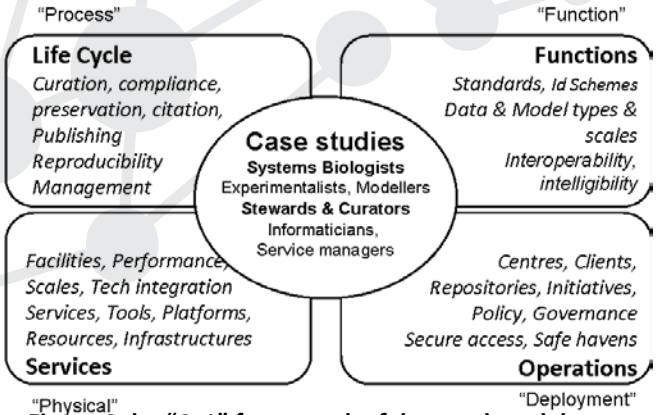


Figure 4 the "4+1" framework of data and model stewardship in ISBE.

Here we describe the overall process of data and model curation in ISBE; and how this process is determined and informed by digital data curation processes, available community standards, and the landscape of data and model repositories and management platforms for Systems Biology, and the stewardship services necessary for the SCs to provision.

We are loosely following the "4+1" model[24] (Figure 4). Each sub-model is driven and influenced by the other three, and all models are examined from the perspective of Systems Biologists using ISBE, and Stewards managing the ISBE data, models and other assets. Case studies include individuals or projects consuming ISBE services (potentially producing new ISBE assets), and ISBE centres performing their routine work of model and data production. The model loosely follows:

Who	Operations	Where	Services, Operations	Why	Case studies

[24] http://en.wikipedia.org/wiki/4%2B1_architectural_view_model

Functions: Properties of Data, Data Types, and Scales.

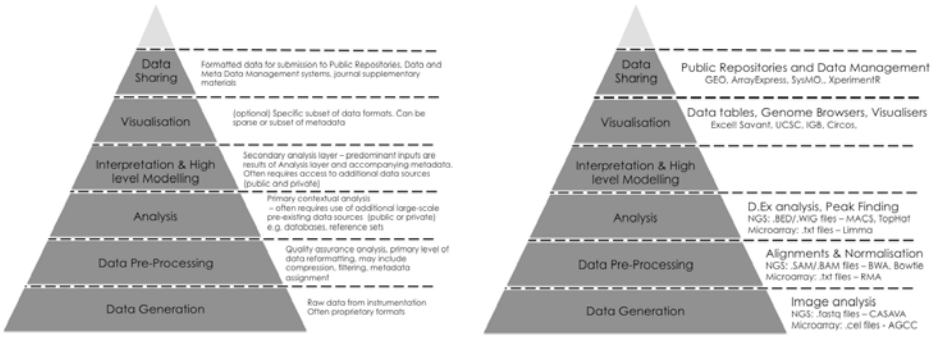


Figure 5 Data property triangle. From data generation through to data sharing.

Raw data and its associated derived data used in systems biology arises from many technological areas. The ISBE-wide survey to date identifies that systems biology researchers are already utilising, or expect to utilise data from multiple technologies (D. 9.1) including microarrays, next generation sequencing, proteomics, metabolomics, single cell-based technologies, and imaging. 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).

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

Illustrations of Changing Data Volumes during Experimentation Analysis Lifecycle

1: The large-scale genome re-sequencing/variant detection project
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:

- 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[25] and Ensembl[26]. Data submitted to the ENA (European Nucleotide Archive Short Read Archive[27] 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.

[25] http://www.1000genomes.org/
[26] http://www.ensembl.org/Homo_sapiens/Info/Index
[27] http://www.ensembl.org/Homo_sapiens/Info/Index

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 paried 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
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 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.

Relationship between Data, Models and SOPs

The systems biology life-cycle 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.

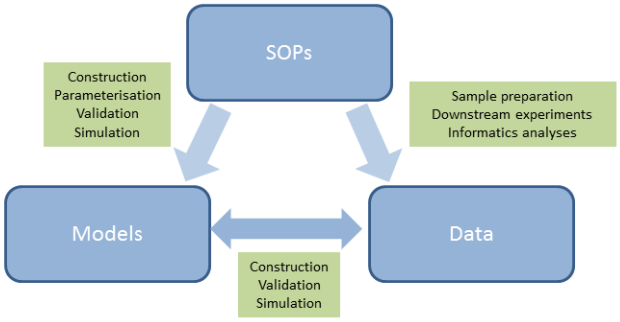


Figure 3 Relationship between data, models and SOPs.

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 x shows the relationships between data models and SOPs in systems biology investigations.

Recommendations for function.

Data recommendations.

- We can characterise data that is ready for the quantitative modelling required by Systems Biology as “born Systems Biology ready”. Data that is generated or already in existence that has not been generated with Systems Biology in mind has to be “made Systems Biology ready”. The SCs must work with non-ISBE Data Generation Centres on data enrichment, standards and interoperability harmonisation; the MICs must provide directives on which datasets need to be made SysBio-Ready and how.

- ISBE should take responsibility for the sustainability of public EU resources that are essential for Systems Biology and fall outside the remit of ELIXIR
- ISBE should provide a Systems Biology “view” over existing public resources managed by ELIXIR or other international organisations.
- ISBE should provide a Systems Biology “view” over existing resources with restricted access policies that are essential to Systems Biology and are managed by international organisations. Service Level Agreements would be required to ensure continued access and use.
- ISBE should provide an aggregated “view” of the results of ISBE investigations and provide a public deposition space for 'homeless' data and model types.
- All data stewarded by ISBE should be accompanied by a Standard Operating Procedure.
- ISBE should signup to the BioMedBridges Data Management and Sharing Charter, and should extend this to a Model Management and Sharing Charter.
- Data security, IP, licensing and sensitive data (e.g. patient).

Model recommendations

- Models generated as part of ISBE projects will, where feasible, be maintained as an ISBE quality model. The model will be made freely available for others to use, and will be published in an established standard (e.g. SBML, CellML, NeuroML).
- ISBE should provide access to a number of high quality models, curated from the literature and stored as ISBE models.
- ISBE will make models submitted to associated databases available for ISBE users to access and run. Their usability will be maintained by ISBE, but their quality, or standardisation will not be.
- All model formats should be supported if the models are relevant, this may involve curation or export to a new format, where necessary.

Metadata recommendations

- Data and model metadata are fundamental for *interoperability* and *intelligibility*. Scientific metadata covers: identifiers, checklists, formats and vocabularies. Cross-cutting metadata covers: author and attribution, provenance, access permissions, availability conditions, dependencies (for example for model execution), versions. Both are necessary and the responsibility of ISBE to identify, recommend and develop pathways for adoption.
- Metadata for ISBE generated data and models (from the academic domain) should be made publicly available through the ISBE infrastructure to encourage reuse by the ISBE community and beyond.
- To address data interoperability ISBE must map the landscape of dictionaries, controlled vocabularies and reporting standards, their status, use etc. ISBE should work with ELIXIR, BioSharing.org, COMBINE, NORMSYS and others to systematically catalogue the landscape of metadata standards and the compliance of current data and model repositories to those standards.
- Biological objects in data, models and SOPs must be annotated with resolvable identifiers from recognised public databases and ontologies. ISBE will recommend which identifiers and ontologies should be adopted. Data and models that are published should be citable through global identifiers such as PURLs and DOIs; their authors should be unambiguously identifiable through global identifiers such as ORCID. The allocation of ids enables: data and model publication and fuels impact tracking services to bootstrap data and model citation statistics.

- Data and models should be annotated using existing minimum information model standards using standard vocabularies. ISBE will recommend which should be adopted and ISBE should provide a catalogue and templates for those standards. Where those standards must be revised and developed, ISBE must work with standards groups such as COMBINE, GSC, W3C and RDA.
- Data and models should be exchanged using existing data and model formats. ISBE will recommend which should be adopted and ISBE should provide a catalogue of templates for those formats. Where those standard formats must be revised and developed, ISBE must work with standards groups such as COMBINE, GSC, W3C and RDA.
- Data, models and SOPs, and other related scientific outcomes, should be organised (e.g. ISA, which forms the basis for Nature Publishing's Scientific Data) and linked (e.g. SED-ML) using metadata standards. ISBE will recommend which should be adopted. Where standards formats must be revised and developed, ISBE must work with standards groups such as COMBINE, GSC, W3C and RDA.
- There is a gap for standards and their adoption; between the state of the art technically on one hand and what system biologists routinely use on the other. To address data interoperability ISBE must develop real pathways to adoption of the ISBE and ELIXIR endorsed standards, propose pathways and mechanisms to maintain the marked-up core data sets in the face of updates to vocabularies etc; and devise how to retain the historical trail of annotations in the face of updated metadata standards.
- ISBE must work with the community to improve and semantically enable curation tools not so much for the ontology / vocabulary development itself but for the data curation (re)using existing vocabularies. This raises technical issues, but more importantly, social issues, notably: in the development of trust in mappings, concepts and in curated data/services.
- Catalogue; meta-directory of directories of models, data and services, leveraging current directories (BioModels, JWS Online, PMR, SEEK, re3data, BioMedBridges)

Lifecycle

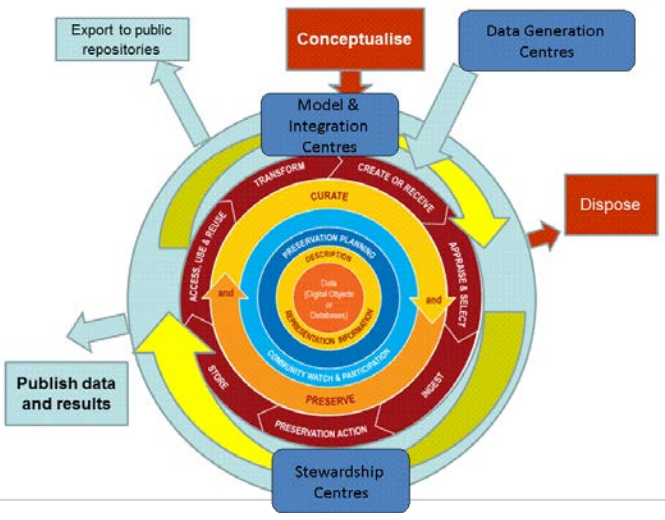


Figure 4 Data and model life-cycle and how different centres relate.

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, developed by the UK's Digital Curation Centre[28] and now widely adopted internationally (Figure 4).
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'.

1. **Create or Receive**

Model Generation

- ISBE data integration centres will produce preliminary models from available resources, in standard formats, in collaboration with ISBE users.
- The community will produce their own models using data from ISBE and in-house data.
- Models should be made available in a standardised format such as SBML or CellML 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 meta-data and should be curated to ensure that it reproduces the behaviours it is supposed to.
- Data protection requirements will be agreed early.

Data Generation

- Once available resources have been identified, missing resources should be supplied by the ISBE Data Generation centres. A registry of which Data Generation Centres offer which types of data would be useful.
- Data generation centres will be responsible for producing data using established SOPs and exporting it to ISBE users, other ISBE centres and/or public repositories.
- The community will produce their own data.
- SOPs should be used for data generation, and where there are no suitable SOPs new ones should be developed. In addition all data available to ISBE should be stored as post-processed, and conforming to ISBE standards for context, syntax, and structure.
- Data protection requirements will be agreed early.

2. **Appraise and Select**

- Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
- All final data and models generated within ISBE, as ISBE data, will be stored and shared within ISBE according to the associated data sharing policy, and for a minimum of 10 years after collection.
- All available high-value data that complete current data sets for modelling, or form parts of new required sets will be stored within ISBE.
- Only post-processed, final data sets, suitable for inclusion within mathematical models will be suitable for storage in ISBE.
- All stored data must meet a minimum requirement for ISBE quality which includes suitable meta-data mark-up
- Data integration centres will incorporate newly generated data into models (again in standard formats) and share these models with ISBE, according to the agreed data sharing policies.

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

- Data and models that do not meet the quality requirements of ISBE can, where appropriate, be linked via ISBE, but not specifically stored as ISBE data.
- The inclusion of data into ISBE should follow data value checklist guidelines (such as NERC, see Sect Functions).

3. Ingest

- For ISBE data (selected for quality based on usability for modelling), SCs are responsible for the final meta-data included on the data.
- For ISBE data, ISBE will be responsible for the final aspects of quality assurance on deposited data.
- For all other data, the listed authors of the data will be responsible for the final aspects of quality assurance on deposited data.
- Ingesting of data will include:
    - Data submitted to ISBE associated/owned sites (e.g. SEEKs, OpenBIS) 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 at point of first construction.
    - Models selected by ISBE, at point of listing, as valuable for long-term storage as an ISBE model.
    - Models submitted to ISBE associated/owned sites (e.g. SEEK) irrespective of privacy settings.
    - 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.

4. Preservation

- All ISBE data will be migrated to a best format, and made available in suitable mediums.
- All modifications made to the data for preservation purposes will be documented with the data.
- All data produced from ISBE DGCs should be annotated according to strict protocols, and checked for compliance by the SCs.
- Data submitted to ISBE externally is graded for usefulness of data with regards to suitable meta-data markup, whilst guidelines for good meta-data should be made available.
- SCs 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.
- Models (just ISBE or models also added to ISBE) can be analysed for quality using an adapted version of the software evaluation: criteria-based assessment.
- Model refactoring - relates to migrating to best formats.

- Models will be built using supported standards where possible (SBML, CellML) what format should ISBE modellers use where this is not possible - does it matter if the model is available on a virtual machine?
- Availability of ISBE data by general users will only be possible through 'packages' of data, containing bundles of complimentary data (by type and also by experimental conditions).
- 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.

5. Store

- ISBE data and models will be stored for specified length of time.
- Data and models added to ISBE will be stored for a specified number of years after the end of the project. The number depends on the data.
- Security of storage.
- 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.

6. Access, Use and Reuse.

- Completed ISBE projects will share data, models, SOPs and any other relevant assets with the rest of ISBE, and the wider scientific community (according to data sharing policies).
- All data, models, SOPs and other relevant 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.
- Accessibility - how people be able to access data and models, programmatically and through user interfaces: catalogues, APIs, portals.
- DOIs that can be linked to from a number of different platforms (publications/ blogs/personal web pages/ LinkedIn/ResearchGate etc).
- 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.

7. Transform

- Create new data from the original, for example: by migration into a different format, or by creating a subset, by selection or query, to create newly derived results, perhaps for publication.
- SCs will be responsible for transforming relevant existing data sets (not born ISBE/Sys Bio) to formats that can be used in Systems Biology projects.

8. Dispose

- Disposing of ISBE data that is not suitable for long-term inclusion in ISBE based on specific policies, guidance or legal requirements:
- migrate to a different storage source.
- if invalid delete permanently.
- secure destruction depending on nature of data.

9. Reappraise

- ISBE data and models which fails standards required is returned to associated ISBE centre for improvement before inclusion in ISBE (or until released to the wider public/customers).
- SCs do not have responsibility for data and models included in ISBE.

10. 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.

Services

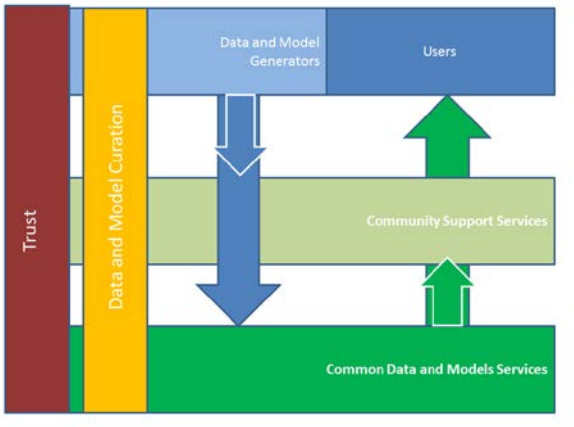


Figure 5 Overall model of e-infrastructure.

Regardless of these specific requirements of integration and interaction between the modelling and experimental elements of systems biology, all scientific e-infrastructures require similar structural elements. Figure 5 is adapted from[29] 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.

- **Common Data and Models Services -** describes the physical infrastructure and the services required to interact with it. It defines where and how data and models will be stored,

[29] 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

how they are identified, the security protocols required to access them, their versioning, backups and federation (in the case of distributed architecture). *For ISBE SCs underpinning services for data storage, access & authorisation, data shipping, data citation, cloud compute, identity resolution, preservation etc will be provided by European RIs already in place, such as EUDAT and ELIXIR (see Sect X) or already widely used in Systems Biology.*

- **Community Support Services -** describes the services required to discover, navigate and annotate contents (i.e. indexes, catalogues, registries and tools for interpretation). For systems biology, this includes tools to explore and run models in addition to data analysis and catalogues. *For ISBE SCs a large ecosystem of data and model management services and platforms, integration platforms and knowledge bases and gateways already exist (see D2.1*). Different types of resources include:
    - Repositories for data, models and SOPs, which should also be the final submission location for much of the data generated within ISBE. These repositories tend to be 'silos', specialising in one type of data collection. For example, PRIDE is a repository for proteomics data and BioModels is a repository for models.
    - Knowledge-bases 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 KEGG pathways database, SABIO-RK and BRENDA.
    - Platforms for Systems Biology. These resources provide infrastructure for Systems Biologists to share their data and models within projects and consortia, and allows them to perform analyses over their content. Many platforms allow public sharing of their content, and therefore become a further source of data and models for Systems Biology activities.
    - Tools for analysis. These resources allow researchers to run simulations over models, or perform informatics analyses over data resource.

ISBE will require a strategy to access and search across distributed resources in order to enable researchers to discover relevant information for their experiments. ISBE will also need a strategy for submitting new data and models back into public repositories, and policies governing when and how this occurs. Interaction with ELIXIR is particularly important for these issues. Many external resources will be ELIXIR resources, or will be co-ordinated by ELIXIR. The adoption of community standards in ISBE, for storing and sharing data, will allow more straight-forward interchange with ELIXIR and other resources.

- **Data and Model Generation -** Data and models generated specifically for the requirements of the infrastructure users. In Systems Biology, quantitative data collected with physiological conditions is most important. For model generation, this would include processes like constructing, parameterization and validating models. *For ISBE, data generated that is not "born SysBio/ISBE" will need to be made so by Community Support Services operated by the DGCs and SCs.*
- **Users -** Services to enable users to interact with the content, allowing the identification and use of resources from within the infrastructure and from external sources.

- **Curation -** Curation of the processes and of the data and models themselves. Compliance to community metadata standards (checklists, minimum information models, identity schemes, format and ontologies), context of the experiments and links between experiments.
- **Trust -** The 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.

Recommendations

- The ISBE framework does not dictate a single platform or a tightly integrated data infrastructure. Rather it will focus on: *conventions*[30] that enable data interoperability and stewardship and *compliance*[31] against data and metadata standards, policies and practices. We propose that the convention for data and model services interoperability should be based on the minimal "hourglass" approach[32] (Figure 7) which is the same as that that underpins the internet, the web and other robust, heterogeneous yet interoperable infrastructures. The hourglass focuses on the specification of lightweight interfaces, standard protocols and standard formats. A similar approach has been proposed during the FAIRPORT[33] meeting attended by representatives of IMI, ELIXIR, ISBE, BBMRI and the USA[34]. ELIXIR's workstreams on Tools and Data Interoperability are also in this direction. We propose that the conventions for data and model metadata descriptions be founded on community standards for: identifiers, formats, checklists and vocabularies.

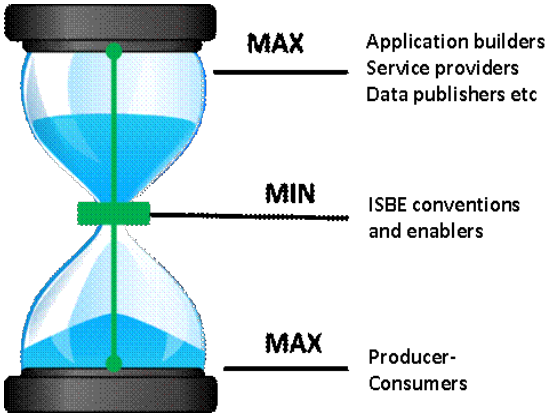


Figure 7 "hourglass" approach.

- Systems Biology inherently draws upon multiple data types. Thus the data management framework proposed by ISBE and managed by the Stewardship Centres must cater for the integration of data. Integration comes in many forms[35], ranging from specialist warehouses to cross-linked indexes, and ranging in the degree to which data remains in its native source and the degree to which is subjected to "Extract Transform Load" pipelines.

- ISBE will operate within an evolving and mixed ecosystem of data resources, of different types of data under the stewardship of different RIs. The federated nature suggests that the data/model framework for ISBE may resemble a federated Aggregated Data Infrastructure (ADIs), interoperating and reusing

[30] A convention is a set of agreed, stipulated, or generally accepted standards, norms, social norms, or criteria, often taking the form of a custom.
[31] Compliance means conforming to a rule, such as a specification, policy, standard, laws or regulations.
[32] This is usually called the hourglass model but that terminology is likely to cause confusion in ISBE.
[33] A FAIRPORT is proposed as a safe and fair data stewardship, trading and routing environment supporting services pertaining to the entire data stewardship cycle.
[34] http://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602&venue=Snellius
[35] Goble and Stevens State of the nation in data integration for bioinformatics, *Biomedical Informatics* 41(5): 687-693, 2008

many data and model resources and platforms. ADIs aim to exploit independent data type and RI specific resources by: (i) cross-resource access, indexing, querying and processing; and (ii) Extract-Transform-Load pipelines to migrate and process data between resources.

Operations.

ISBE Centres

ISBE infrastructure is separated into three classes of centres. However, data and model management is cross-cutting and is *the responsibility of all ISBE centres* (specific responsibilities as described in the Digital Curation Centre Model - Fig#). The responsibilities of each type of centre must be clarified and coordinated.

ISBE infrastructure is separated into three classes of centres. However, data and model management is cross-cutting and is *the responsibility of all ISBE centres*. The responsibilities of each type of centre must be clarified and coordinated.

General properties of ISBE Centres.

Must:

- **[Defining Compliance and Process]** Make descriptions of the structure and standards available to the public for data, models, and SOPs.
- **[Defining Conventions and Standards]** The conventions for data and model metadata descriptions be founded on community standards for: identifiers, formats, checklists and vocabularies.
- **[Defining Conventions and Standards]** Offer consultancy and advice for standards and formats for ISBE clients (that is ISBE participants - guidelines for complying with the ISBE data sharing policy), and for ISBE centres (that is standard operating procedures for ISBE centres)
- SCs will make recommendations for tools and resources to assist with standards compliance, lower the barrier to adoption and to make standards-compliance more efficient and less time-consuming.
- **[Stewardship Services]** Will ensure all ISBE born data and models are standards compliant and annotated with the correct metadata.
- **[Stewardship Services]** offer advice with producing data management plans for funding proposals, ensuring consistency of data and models that become ISBE resources and further use of ISBE facilities.
- **[SLAs with RIs]** ISBE make SLAs with key **domain related RI**, notably the ERANets, ELIXIR, Euro-Bioimaging, BBMRI, and the IMI.
- **[SLAs with RIs]** ISBE must make SLAs 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.
- **[Consultancy and Training Service]** Train customers in best practice for data and model management according to the most up-to-date agreements.

Should:

- **[Defining Conventions and Standards]** The ISBE framework focuses on *conventions*[36] that enable data interoperability and stewardship and *compliance*[37] against data and metadata standards, policies and practices The conventions for data and model services interoperability should be based on the internet and web's minimal "hourglass" approach[38], a specification of lightweight interfaces, standard protocols and standard formats.
- **[Consultancy and Training]** Provide consultancy to co-develop format/annotation/cross-linking/storage requirements for research groups, journals, funding councils needs.
- **[Consultancy and Training]** Provide training for co-develop format/annotation/cross-linking/storage requirements for research groups, journals, funding councils needs.

Will not:

- **[Defining Conventions and Standards]** The framework will not will not dictate a single platform or a tightly integrated data infrastructure for users.
- ISBE infrastructure is a *set of services* to support the stewardship of ISBE data/models, access to ISBE data/models and the technical compliance of data/models against metadata standards, policies and practices. ISBE should not govern the science or scientific methodology that is undertaken using its infrastructure. That is the purview of peer review.
-

ISBE Stewardship Centres

Must:

- **[Defining Conventions and Standards]** will develop and agree on standards and formats for exchanging data and models etc. between centres and for ISBE clients based upon the current standards in use within the Systems Biology community.
- **[Defining Conventions and Standards]** SCs will contribute to the standardisation of formats, ontologies and minimum information checklists.
- **[Defining Compliance and Process]** will review and evaluate data management processes and standards within ISBE to ensure that ISBE processes continue to run smoothly and new types of data or modelling approaches can be integrated with little disruption.
- **[Defining Compliance and Process]** must anticipate a mixture of open and commercially sensitive data/models and open and commercial services. There is an expectation that 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. It must also anticipate potential financing as a public private partnership and the implications this may have on data visibility – its accessibility and accessibility. The extent and under which operating conditions that ISMB should support private and proprietary data needs to be clarified.
- **[Stewardship Services]** will be responsible for and provide digital curation services for linking data, models, maps, SOPs and experimental descriptions, linking to individuals and/or organisations to ensure credit is awarded to creating scientists. This will ensure that

[36] A convention is a set of agreed, stipulated, or generally accepted standards, norms, social norms, or criteria, often taking the form of a custom.

[37] Compliance means conforming to a rule, such as a specification, policy, standard, laws or regulations.

[38] This is usually called the hourglass model but that terminology is likely to cause confusion in ISBE.

associations between these components are preserved and different versions are recorded properly.

- **[Stewardship Services]** will provide services for collecting and curating data, models, maps, SOPS, experimental descriptions etc in standards compliant forms. Such services will seek to avoid as much disruption to the scientists working practices as possible, and ensure proper implementation of standards.
- **[Stewardship Services]** take responsibility for and provide services for transforming relevant existing data sets to formats that can be used in systems biology projects.
- **[Stewardship Services]** take responsibility for and provide services for digital preservation or migration of ISBE-related data, models and maps to established archives and repositories and develop policies for accessing and using resources.
- **[Stewardship Services]** take responsibility for and provide services for digital preservation or migration to established archives and repositories. Publishing or deposition of supplementary materials for publications.
- **[Stewardship Services]** Use discovery services for finding: (people) modellers, software, data, SOPS and models, and more. Monitoring services are needed to identify which resources are used.

Should:

- **[Defining Compliance and Process]** devise and review compliance with: management and preservation processes; annotation and curation standards; access and responsibility policies; and quality control. It is the Stewardship Centres' responsibility to facilitate: access, archiving, annotation and discovery through portals, cataloguing and indexing, programmatic interfaces etc.
- SCs should host data and model management facilities at the International, National and Centre level supporting selected Sys Bio data and model services and resources such as catalogues, libraries, model repositories, and key data repositories. Such services/resources should be selected and technically reviewed by ISBE as adhering to: the ISBE data interoperability conventions; a prescribed level of quality; compliance with quality of service and metadata standards. They will be scientifically reviewed as to their importance and sustainability prospects. The "certification" criteria of data and model resources and services must be defined. The "certification" process must be defined. Certification should complement that of ELIXIR.
- **[Stewardship Services]** seek to support hosted data and model management facilities for System Biology private, project and laboratory clients, as a "cloud" service, with suitable access permissions and backed by scalable computational infrastructure[39].
- **[Stewardship Services]** SCs should recommend and catalogue, and possibly certify and support, data and model management software platforms that can be deployed privately by clients.

[39] It is not in the remit of ELIXIR to provide such a data/model hosting facility.

ISBE Data Generation Centres

Must:

- [Defining Conventions and Standards] Work with SCs to ensure standards for structuring and annotation data are relevant for how the data will be used, and as experimental protocols change.
- [Defining Conventions and Standards] Train all relevant staff how to correctly structure and annotate their data, so that this occurs as close to production as possible.
- [Defining Conventions and Standards] Take an active role in defining SOPs to be held centrally within ISBE.
- [Stewardship Services] Use the agreed formats from SCs to structure and annotate ISBE born data (specific to the data types).

Should:

- [Defining Compliance and Process] Develop automation of data structuring and annotation at the collection source.
- [Defining Compliance and Process] Make the prepared data immediately accessible to all ISBE centres.
- [SLAs with RIs] Make the structured and annotated data available to other ESFRIs as single sets.

Could:

- [Stewardship Services] Select important published data and structure and annotate to ISBE standard for use within ISBE.
- [Stewardship Services] Structure and format datasets non ISBE born as a service.

Will not:

- [Defining Conventions and Standards] Dictate what formats data submitted to ISBE must be in, only recommendations will be provided.
- [Defining Conventions and Standards] Be responsible for the quality of data submitted to ISBE, only that of ISBE born data.
- [Stewardship Services] Format data submitted by the public, unless the impact of the data will be high.
- [Defining Conventions and Standards] [SLAs with RIs] Release data to other centres, customers, or ESFRIs unless it has adequate annotation.

ISBE Model Integration Centres

Must:

- [Defining Conventions and Standards] Work with SCs to ensure standards for structuring and annotation of models are relevant for how the models will be used, and as modelling complexity evolves.
- [Defining Conventions and Standards] Make descriptions of structure and annotation best practice available to the public.
- [Defining Compliance and Process] Cross-link all ISBE born models to the derived datasets that were used to produce them.
- [Stewardship Services] Maintain a database cataloguing the structure of models submitted by the public. This should be used to ensure that associated software (e.g. virtual machines) is available for running the model.

Should:

- [Stewardship Services] Survey available models in order to identify suitable models for formatting and inclusion of ISBE data. (what is meant by this – explain better)
- [Defining Conventions and Standards] Provide a list of preferred softwares for producing model.

Could:

- [Stewardship Services] Offer restructuring and curation of user models as a service so that models can be stored in a preferred ISBE format (this would not guarantee quality, just longevity).

Will not:

- [Defining Conventions and Standards] Take responsibility for the quality or usability of models submitted by the public.
-

The introduction of a European infrastructure for Systems Biology will have an impact on researchers and their institutes (in Systems Biology and in other fields), European funding bodies, publishers and commercial companies engaged in Systems Biology research. The following describes the predicted effects on each of the stakeholders.

Researchers and Institutions

Researchers already working in Systems Biology will have access to more computational and knowledge resources through ISBE, providing an environment to perform Systems Biology systematically and at scale.

- Researchers from outside of Systems Biology will have the opportunity to explore systems approaches to their work
- ISBE will provide a mechanism for ensuring the long-term storage of previous work (in conjunction with other ESFRI initiatives, such as ELIXIR), providing a safe-haven for data, models and other resources and allowing researchers to comply with open data sharing and open access policies.
- ISBE will be a long-term sustainable infrastructure. Time and resources will not need to be consumed on developing local data management solutions and re-implementing similar systems in different institutes.
- The ISBE community portal will highlight European research in Systems Biology, providing greater exposure and impact for individual scientists throughout Europe and internationally. ISBE may also function as a matchmaking service, linking researchers with similar or complementary interests.
- Access to high quality and standardised training that covers the breadth of Systems Biology research.

Funding bodies

- Funding for resources required by the whole European community can be co-ordinated and maximised.
- Less re-implementation and reinvention of similar data and model management systems in individual member states.
- ISBE would provide a framework to structure grand challenge activities.
- Publishers
- ISBE guidelines for metadata and for data and model sharing would ensure minimum standards prescribed by the journals would be more frequently attained.
- ISBE could function as a supplementary data and models store for journals, which would only be possible due to a long-term sustainability model for ISBE.
- Something about validation and testing robustness of models using the ISBE framework
- Best practice training that could be used in conjunction with studentships.

Commercial players - Organisations using or Exploring Systems Approaches

- The ISBE infrastructure and community portal would make it easier for commercial companies to discover and contact researchers with particular expertise.
- Access to well-described, publicly available data and models in a re/usable state would improve the accuracy of their research without raising costs.
- Service level agreements or full collaborations would be possible with ISBE, allowing companies to easily work with consortia of researchers under a single agreement.
- A mixture of public and private storage facilities and policies would be in place for assets generated in ISBE, allowing commercial interests to be protected.

Other ESFRIs -

- ISBE would be a consumer (or “customer”) for data produced in other ESFRIs, such as ELIXIR or EUROBIOMAGING.
- ISBE would also rely on the use of ELIXIR and other ESFRIs physical infrastructure for storage and compute or for data generation.

- The ISBE systems approach could be applied to research in other ESFRIs, providing expertise for new research approaches.

Policy

- Annotation and curation of data and models is an ongoing process and is *the responsibility of all ISBE centres*. The stewardship centres assess annotation and curation completeness and compliance throughout the process.
- All resources generated in ISBE should be annotated with their curators and affiliations to promote Data Citation and to encourage adoption of ISBE for research and sharing.#
- ISBE should issue guidelines for data and model management for projects using ISBE infrastructure and resources.

Social

- Sign up to a sharing charter policy, reward and compliance benefit, policies, control, rewards, DM advocates trainers have to be funded by ISBE (at least partially) (think EraSysAPP)
- Network of DGS - people who use ISBE can commission to generate data that are guaranteed to be compliant to ISBE sys bio standards and governance. Some of those DGSs will be ISBE member and some will be partners or contractors. The responsibility of ISBE is to set the terms of compliance, to monitor the compliance and to adapt when necessary to new compliance requirements. ISBE quality control.
- A network of stewardship centres, that may specialise in particular data types, are responsible for the stewarding of data, models and SOPs at the raw, derived and enriched levels. The linking and aggregation of data, models and SOPs is a specialised ISBE SC activity.

Technical e-infrastructure Requirements

WP9 is responsible for Technology Watch. This is a preliminary draft of e-Infrastructure for data and model management.

e-Infrastructure is a foundation of computing and information services, designed to support multiple geographically dispersed research institutions/groups/researchers, over the internet, in advanced data handling, and computing and information processing services. The data handling covers acquisition, storage, management, integration, mining, visualisation. e-Infrastructure[40] refers to the technology and organisations that support distributed, national, and multi-country collaborations enabled by the internet. Tilsley[41] and Coveney[42] present an infrastructure viewpoint that we will drawn upon here.

[40] http://www.jisc.ac.uk/media/documents/publications/einfrastructure_rtf.rtf
[41] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/32499/12-517-strategic-vision-for-uk-e-infrastructure.pdf
[42] http://wiki.esi.ac.uk/w/files/f/f5/ResearchComputing-glossy.pdf

	Stewardship
Data and data services Metadata services	Primary and secondary repositories, Catalogues and libraries, Search services, citation tracking, Data, model and people matching services, resource sustainability
Software, services and algorithms	Modelling tools, management systems, Curation tools Text mining tools Stewardship and sustainability of specialist data and modelling software Protocols for analysis algorithms in software (e.g. cobra codes up most used stoichiometric algorithms for general use with SBML models).
Storage	Data locality; Cloud data storage service, replication services; archive services; back up services; security; long-term preservation services
Compute	Compute service for multi-scale models, model simulation, movement of compute to data (data locality); elastic compute cloud, HPC, HTC, Grid
Networks	Secure movement of data between centres, between clients Large scale data movement
Security & Access services	Access authorisation, authentication and accounting; Embassy Clouds
People & Skills	Systems Biologists who generate and use D&M, accessing and reusing stewarded D&M, depositing and stewarding D&M. Curators, stewards, systems administrators. Software Carpentry, Data Carpentry Model Carpentry Virtual Research Communities and collaboration
Instruments	LIMS, Remote Access and Remote Instrumentation

Case studies

We have looked at a number of case studies of data use in Systems Biology in the sister deliverable 2.1 (Appendix D). These case studies take a typical aspect of the project, and look at how the data is planned, generated, processed, and finalised. The case studies cover high throughput data design, collection, and analysis; generation of genome-scale metabolic models from gene sequence data; and lastly construction of large-scale kinetic metabolic models from genome-scale reconstructions. These are typical examples of what we might expect to do with Systems Biology data. They show that the scientists involved must use very distributed methods, software, and databases, for each step of their day-to-day data processing. This distribution of resources means that many different resources must be able to understand input/output of other resources, otherwise they must be converted, which is time consuming and could introduce difficult to trace errors (especially if these are converted manually). The most important standards for the case studies are SBML, FASTA, E.C. numbers, MIAME.

SWOT analysis

Strengths • Cross border initiatives increase efficiency • Access to compute and storage • Capacity building in systems biology in Europe • Enables scientists from other disciplines to try systems biology methods. • development of single community standards so asset re-usage is easier.	Opportunities • A framework for large-scale participation in grand challenge initiatives • Single discipline labs have access to multidisciplinary expertise. • Research fellows have broader opportunities for their research without needing multiple skills. • Improving life-time of models and data making research activities more efficient.
Weaknesses • Management and procedural overheads • No established career paths for expert modellers acting as service providers. • Modelling currently involves intellectual input from modellers and therefore cannot be provided as a strict service (intellectual property).	Threats • Stringent standards and formats could be a barrier to uptake • Lack of trust of distributed computing facilities for non-published data • Undefined overlaps between other ESFRIs - • Lack of funding (in some member states), prevention of long-term sustainability. • Community needs evolving faster than infrastructure.

Annex A

The challenges for Data Science highlighted by the Riding the Wave 2010 Report.

Scientific e-infrastructure – some challenges to overcome

Collection	How can we make sure that data are collected together with the information necessary to re-use them?
Trust	How can we make informed judgements about whether certain data are authentic and can be trusted? How can we judge which repositories we can trust? How can appropriate access and use of resources be granted or controlled?
Usability	How can we move to a situation where non-specialists can overcome the high barriers to their being able to start sensible work on unfamiliar data, perhaps using intelligent automated tools for an initial investigation?
Interoperability	How can we implement interoperability within disciplines and move to an overarching multi-disciplinary way of understanding and using data? How can we find unfamiliar but relevant data resources beyond simple keyword searches, but involving a deeper probing into the data? How can automated tools find the information needed to tackle unfamiliar data?
Diversity	How do we overcome the problems of diversity – heterogeneity of data, but also of backgrounds and data-sharing cultures in the scientific community? How do we deal with the diversity of data repositories and access rules – within or between disciplines, and within or across national borders?
Security	How can we guarantee data integrity? How can we avoid data poisoning by individuals or groups intending to bias them in their interest? How can we react in the case of security breaches to limit their impact?

Scientific e-infrastructure – some challenges to overcome *continued*

Education and training	How can the citizen make these benefits available for sensible investigations, and how can they be safeguarded from fakes? How can scientific e-infrastructure foster and increase popular interest and trust in science? How can we foster the training of more data scientists and data librarians, as important professions in their own right?
Data publication and access	How can data producers be rewarded for publishing data? How can we know who has deposited what data and who is re-using them – or who has the right to access data which are restricted in some way? How do we deal with the various 'filters' that different disciplines use when choosing and describing data? What about differences in these attitudes within disciplines, or from one time to another?
Commercial exploitation	How can the infrastructure benefit from commercial developments in data management? How can the revenue-generating expertise of the commercial world be brought into play for the long-term sustainability of these resources?
New social paradigms	How can we learn from the wisdom of crowds about what and whom to trust, while avoiding being misled by concerted campaigns of deceit?
Preservation and Sustainability	How can we be sure that the important information we collect will be usable and understandable in the future; in particular how can we fund our information resources in the long term? How can we share the costs and efforts required for sustainability? How can we decide what to preserve?