4 ELIXIR Data Interoperability Strategy

We presume that ELIXIR best practices and directories for programmatic access to data (e.g BioCatalogue; BioMedBridges Directory) is covered by Workstream 1 We see a strong link and collaboration with our stream

Remit. Recognising the importance of effective and sustainable access to reusable research data, there is an increasing requirement from research funders to publish and make the findings from publicly funded research broadly available. Data driven science requires the seamless connection and joint analysis of newly generated data with interoperable legacy information to derive new insights. This explicitly also requires the secure and safe connection of data with managed access consistent with patient data. ELIXIR therefore recognizes, as also demonstrated by several of its early pilots[1]) that it will need to support a mixed open and managed acess data interoperability environment to serve academic, clinical and commercial partners who have access to the datasets.
Optimal data sharing and re-use where possible, is increasingly seen as a basic socio-ethical requirement. An often-implicit assumption in these requirements is that the data will be (a) made available in a machine readable and reusable format – a prerequisite to realise the value of the produced data in new contexts; (b) have sufficient metadata to be intelligible outside its source laboratory; and (c) use standard terms, identification schemes and metadata practices such that data can be systematically discovered, linked and compared by humans and by computers.
Data sharing and re-use is tightly coupled to effective research data management. Thus ELIXIR is concerned with the issue of **Data Stewardship**: that is, the presence of processes and infrastructure to: support data coordination, metadata capture, publication, curation, accesibility and deposition in suitable archives; and support the development and widespread adoption of metadata standards. Data stewardship and data sharing are necessary conditions for data interoperation. The data infrastructure developed and maintained by ELIXIR at the European level must enable data stewardship to make the best use of Europe's collective and expanding capacity. The analysis of industry needs within ELIXIR also revealed data interoperability and stewardship as a core issue (http://connecteddiscovery.com/blog/cd-and-elixir/).

ELIXIR's preparatory phase report (as expected after 5 years) is outdated in its details but still resonates in the fundamentals. Individual isolated high quality database silos are now accompanied by cross database, cross technology and cross domain data integration. The journal (and associated impact) is accompanied by data publication and web (data) services are an established paradigm. Now we have RESTful API's, reusable (data processing) workflows and integrative platforms, publishing using Research Objects, nanopublications, Linked Data approaches and new data types at new scales and new methods to represent them. Big Science Data will accelerate these developments. Some new technologies may be obsolete in another 5 years time. ELIXIR, with a natural role in standards approval (and formal endorsement) and sustainable e-infrastructures for data stewardship at its core can steer this development. ELIXIR should on the one hand 'buffer' against the 'hottest and newest to replace what was hot yesterday' and on the other hand dragging along non-scalable solutions. The ELIXIR strategy should also cover the entire life cycle of data in modern, data intensive science starting ideally already with the engagement of ELIXIR (nodes/hub) data experts in the phase of the experimental design of complex, multi-omics experiments all the way through data integration, modelling and analytics, the latter in close collaboration with other Research Infrastructures.

The chief challenges are four-fold: (a) identifying the pan-European data interoperability needs and priorities; (b) defining strategy, building community agreements and cooperating on data stewardship and data exchange at the national level, the European level and also with numerous global initiatives; (c) taking recommendations, standards, training and practices into data centre practice - where data (and software) is actually forged and where scientists actually exchange (we are primarily concerned with the interoperability and

[1] Link to ELIXIR web site/pilots

exchange of ELIXIR data services, as defined by workstream 1); and (d) keeping up to date with a rapidly changing data and consequently data stewardship landscape.

A further challenge will be to clearly define the remit of ELIXIR data stewardship services. Clearly, ELIXIR can not be the instance dealing with all data coming from life sciences experiments in the member countries. There should be a clear focus on those data and information collections that are or recurrent value for knowledge discovery. It will therefore also important to define quite precisely what ELIXIR will NOT do, which again is different from what individual nodes may be engaged in.

To address data interoperability in practice ELIXIR must address four deeper challenges: (i) map the landscape of dictionaries, vocabularies and reporting standards, their status, use etc; (ii) develop real pathways to adoption of the ELIXIR endorsed standards; (iii) propose pathways and mechanisms to maintain the marked-up core data sets in the face of updates to vocabularies etc; and (iv) devise how to retain the historical trail of annotations. We need improved and semantically enabled 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.

ELIXIR in is a unique position to engage with initiatives aimed at standard development and adoption. It should actively engage with the most important community initiatives but as a preferred endorsemant and adoption authority ensuring the use of these standards in the ELIXIR nodes and countries, rather than as an active paticipant. We should not get involved (at the ELIXIR level) in so many activities that we talk more than we implement. Again this does not preclude very active participation of Node scientists in many of these activities.
From the multi-player initiatives perspective ELIXIR is potentially in danger of being swamped and therefore we should strategize on how we make the initiatives 'come to ELIXIR' rather than the other way round.

The full document will further detail choices on issues such as, do we want to play an active/leading role in:
- The development of practice-based policies, practices and standards for data management and accessibility or do we choose to work closely with initiatives such as Research Data Alliance (RDA), the European Collaborative Data Infrastructure (EUDAT) and the NIH Big Data to Knowledge Initiative (BD2K) to ascertain effective and transparent solutions for the solutions they propose and endorse them.
- DG CONNECT's efforts to implement a pilot action on Open Data in Horizon 2020, sharing best practice from Nodes where appropriate.
- Establish common data stewardship and data exchange/interoperability policies, practices, standards and clearly defined responsibilities with respect to related ESFRI projects and clusters, notably BioMedBridges, BBMRI, EATRIS, EuroBioImaging, INSTRUCT and ISBE.
- the activities of the European level Digital Library preservation efforts, such as - the Alliance for Permanent Access and in institutional libraries - as the blending of library and data stewardship functionalities becomes greater, and as open access repositories increasingly combine article and data publication.
- the Data Stewardship policies and solution in the major Public Private Partnerships (PPP) initiatives aiming at common data practices and policy. These initiatives include key infrastructure and data management projects by the Innovative and Medicine Initiative (IMI), the anticipated BRIDGES programme in the industrial biotech discipline, and other global pre-competitive standards-focused efforts, such as the Pistoia Alliance.
- the activities of pre-existing data stewardship and related initiatives in the fields associated with ELIXIR; for example the community-driven standards such as PSI, MSI, GSC, OBO Foundry, ISA Commons (more listed in the BioSharing catalog) other than fostering harmonization, where relevant.
- activities beyond the European ELIXIR Data Stewardship infrastructure and services; for example, identifiers.org, BioPortal, DataONE, iPlant.
- the European and Global level with data stewardship and data interoperability needs and activities regarding sensitive or non-open data, regarding for instance societal aspects such as patient safety or treatment.

- the rapidly increasing interest and activities of data-focused publishers, such as Nature Publishing Group, Elsevier and BioMedCentral, commercial data providers, such as BaseSpace, and services associated with data processing and data publication, such as Thomson Reuters Web of Knowledge, Sage Bionetwork CleanScience, FigShare and Data Dryad. As a minimum, ELIXIR should ensure that data processing, workflow and publications methods adopted by these private and professional activities are interoperable and rooted in community-driven standards.
- Efforts concerned with other dimensions of experiment description and practices; notably: data identification; data citation; data provenance; and data versioning. Organisations such as the Data Citation Synthesis Group (bringing together CODATA, RDA and Force11), CrossRef, DataCite, CDISC, and the W3C are promoting and defining standards and best practices that cut across discipline boundaries.
- The identification of new and emerging approaches to data stewardship such as reusable workflows, REST, integrative platforms (see PoW1), Research Objects, nanopublications, Linked Data, mobile science etc. ELIXIR should seek to leverage modern semantic approaches.
- Lead general and formal recognition of the skills required for data stewardship at the community, the academic reward system and the government level. For example, by working with the International Society for Biocuration, national and European level funding bodies, and professional societies on skills and recognition policies.

Activities likely to be lead by ELIXIR rather than 'post hoc-endorsed'

1. Identifying "ELIXIR-endorsed" data stewardship services and software, based on the coordinated development, implementation and deployment across Europe;
2. Establishing and recommending best practices for data capture, data publication and data curation and the pathways to adoption, based on 'community compliance', building upon the successful areas of Nucleic acid, Proteomics and Protein Structure schemes with longstanding data capture methods.
3. Establishing and recommending best practices for developing, maintaining, evolving and implementing nomenclatures, vocabularies and ontologies as well as other community standards, such as minimum reporting requirements for guiding deposition and facilitating exchange of information and exchange formats;
4. Campaigning for the awareness and adoption of metadata and new semantic methodologies (such as Provenance-decorated Linked Data) across the whole community, from tool and software developers (Software Carpentry) to data providers and managers (Data Carpentry);
5. Advocating for the development and sustainability of data stewardship services, skills, software, resources and policies at the national, European and international level.
6. Actively promote a 'recognized authority role' for ELIXIR with the established initiatives to achieve all the above.

ELIXIR Pilot candidates for the first phase

We propose to start implementing these strategies in 2014 with a series of interconnected and PoW4 coordinated pilots. Here we list the most obvious candidate pilot actions with the most interested and experienced nodes involved and governed by the Hub:

1. Identify and map the current landscape, liveness and adoption of data stewardship standards and practices, including status, sustainability and usage.
2. Identify the practical bottlenecks of data interoperability and actual exchange.
3.Identify and prioritise critical standards, rules, vocabularies and ontologies, dictionaries etc for ELIXIR services and required maintenance resources (see list at BioSharing registry of standards, highly used ontologies); and the gaps to be bridged in order to fully engage BMS infrastructures?)
4. Identify and prioritise European identifier resolution, data citation, provenance, interoperability services and standards, and full use of earlier initiatives.
5. Establish criteria and processes (including benefits & duties) for data stewardship of formally endorsed ELIXIR data resources (from workstream 1) and ontology/curation services.

6. Establish, implement and maintain a "data registry" and a catalogue of global data initiatives in partnership with other initiatives.

(smaller actions, probably to be merged into a single 'Strategy Pilot'

7. Identify representatives from the nodes for relevant initiatives; reporting/dissemination/infiltration protocols and develop plans for effective collaboration.
8. Develop a data stewardship and data exchange blueprint to fill gaps and meet priorities. (see proposed data stewardship work package for H2020 and beyond)
9. Develop training for Data Carpentry etc (with workstream 5).
10. Develop and implement a policy for (interoperable) data publishing with professional publishers.

Relevant Node Services

As outlined in the Node applications, many of the ELIXIR Nodes offer specific data resources, data services and/or training in the collection and management of specific data resources/services. Some of these resources and services will become "ELIXIR endorsed" as core to ELIXIR (see Workstream 1).
The e-Infrastructure needed to accommodate actual data interoperability and exchange will be defined and mobilized in close coordination with all relevant nodes and the Hub (PoW?)
Training in curation best practices and the dissemination of data curation and interoperability methodologies - "Data Carpentry" - is addressed in Workstream 5, and is explicitly identified as part of the remit of ELIXIR-UK. Few nodes explicitly offer identity management or vocabulary/ontology development services.

Resources

To be defined
Useful organisations
To be defined
http://www.rin.ac.uk/resources/digital-curation-and-preservation/useful-organisations