ELIXIR EUROPE: SUSTAINABLE SHARING OF DATA IN THE LIFE SCIENCES

NIKLAS BLOMBERG, ANDREW SMITH, SUSANNA REPO, AND RAFAEL JIMENEZ. ELIXIR HUB, WELLCOME TRUST GENOME CAMPUS, HINXTON, CAMBRIDGE, CB10 1SD, UK. NIKLAS.BLOMBERG@ELIXIR-EUROPE.ORG

Europe is home to some of the world's leading bioinformatics institutes and resources. ELIXIR represents the collective efforts of 17 European countries to coordinate and sustain these vital data resources, which provide the platform for discovery in the life sciences.

LARGE DATA AND COMPLEX DATASETS TRANSFORM LIFE SCIENCES PRACTICE

The steep drop in costs of high throughput biology has enabled European research laboratories to produce an ever-increasing amount of data. Life scientists are rapidly generating the most complex and heterogeneous datasets that science can currently imagine, with unprecedented volumes of biological data to manage.

Biology has a rich tradition of accurate collection and reuse of data, from the systematic cataloguing specimens that have underpinned our fundamental understanding of species, tissues and cells to todays massive studies of human variation [1], functional genomics [2] and integrated cancer atlases [3]. Such data resources serve as critical reference tools for biologists in laboratories. For example, every day more than 10,000 users from across the world access the data resources held at EMBL-EBI.

The use, reuse and integration of these biological data resources have been the basis of many discoveries, both planned and serendipitous, over the decades. For example: novel risk factors for Alzheimer's disease were identified from large-scale meta-analysis of genomic studies [4]; such efforts depend critically on prior estimates on human genetic variation calculated from public data sets such as the 1000 Genomes. Public life science databases have provided the data that allowed development and validation of the molecular design and docking tools in daily use by molecular modellers and medicinal chemists in drug-discovery laboratories across the world [5]. The development and validation of drug-design tools, many of which are successfully commercialised, all relied on protein structure and inhibitor-binding data extracted from carefully curated public archives such as the PDB [6] or ChEMBL [7].

As data has become an essential commodity for biological research, the importance of making both the narrative and data from publicly funded research openly available is broadly recognised. Data needs to be *findable*, *accessible*, *interoperable* and *reusable* to generate value for a research community beyond

ELIXIR, the ESFRI Research Infrastructure for biological data, brings together Europe's major life-science data archives (at EMBL-EBI) and, for the first time, connects these with national bioinformatics infrastructures and resources throughout ELIXIR member states. By coordinating local, national and international resources, ELIXIR will meet the data-related needs of Europe's 500,000 life-scientists.

Formally established as a legal entity in January 2014, ELIXIR is a distributed organisation that uses a Hub and Node structure. The ELIXIR Hub is based in Hinxton, UK, and is charged with coordination activities, whilst ELIXIR Nodes in participating countries provide the services and resources. This coordinated infrastructure provides data, tools, technical services, training, standards and industry support to users in academia and industry. Prioritised by the European Council and ESFRI in 2014 as one of Europe's three priority new infrastructures, ELIXIR is the initiative that coordinates, sustains Europe's life science data resources.

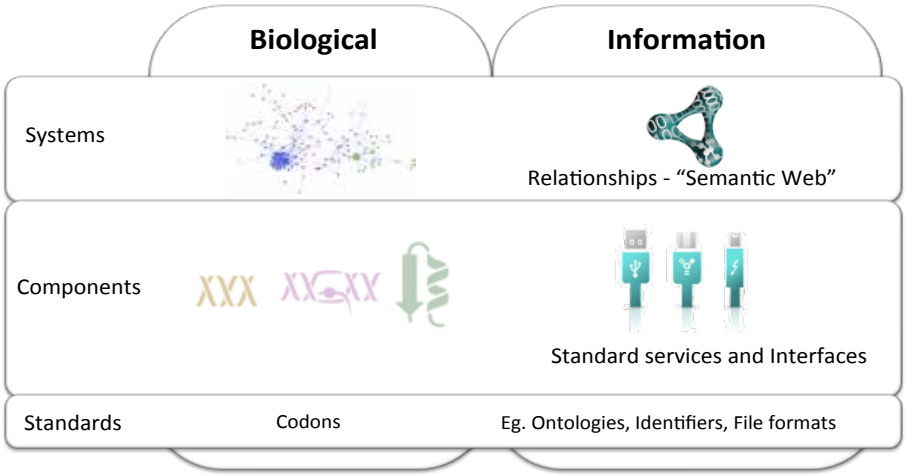

Figure 1: *Just like biology assembles universal building blocks of life into complex systems, information assembles compounds with standard services and interfaces to allow analysis and exploration of relationships in the data.*

the initial researcher's laboratory. It also need robust annotation of experimental conditions and other meta-data to be *assessable* by future researchers [8]. The importance of long-term stewardship is highlighted by the observation that the odds of retrieving the data from a scientific publication decline by 17% per year [9].

As international collaboration is the norm in science, ever-larger volumes of data must flow freely between platforms, laboratories and people. Researchers explore heterogeneous data from multiple sources in order to conduct meaningful analyses and extract knowledge. A life science data infrastructure must be able to cope with this need for aggregation, annotation and functional integration of data from hundreds of laboratories across Europe, as well as the access demands of users worldwide (e.g. the Human Protein Atlas [10] received more than 750,000 visits during 2013). These challenges are too great for any one institute or country to address alone. It also requires alignment and engagement of policy makers, and research funders at both national and EU levels. ELIXIR represents Europe's collective response to this challenge.

A further challenge is the long-term preservation of data across different organisations and countries. Ensuring that our global reference data resources continue to be available for future scientific discoveries requires a coordinated response from researchers in academia, national infrastructures and also industry users. As the size and complexity of the datasets grow so does the cost and challenges for our global archives. By connecting national and international infrastructure, ELIXIR provides a coordinated approach to address this critical, and long recognised, sustainability issue [11] (Figure 1).

Bioinformatics services – biological data resources, tools infrastructure, standards, compute and training – are used across the life sciences [12], not just by bioinformaticians and computational biologists but also geneticists, biochemists, clinical specialists, and plant, environmental and marine scientists. In addition, industry is a major user of ELIXIR resources: users originate from pharmaceutical and biotech industries through to crop science and aquaculture, and encompass both multinational corporations and SMEs alike.

DEFRAGMENTING EUROPE

With growing volumes and complexity of biological data the demand on tools and expertise for managing, integrating and analysing experimental results is rapidly increasing. Many national and European life-science research programmes, as well as public private partnerships (PPPs) such as the Innovative Medicines Initiative (IMI), make significant investments in data and knowledge management infrastructure. ELIXIR's Preparatory Phase report found there are around 1,800 bioinformatics resources in Europe alone [13] (Figure 2). These are rarely coordinated with the core databases and are often difficult for researchers to find. Indeed, the growing number of niche databases makes the bioinformatics landscape needlessly complex for its users and increasingly difficult to sustain.

ELIXIR is founded to coordinate and support sustainable investments in this critical data infrastructure. It is a distributed infrastructure, based on the concept of national ELIXIR Nodes. ELIXIR Nodes represent the national bioinformatics infrastructure, funded by national research organisations, and support life science data management nationally. Some ELIXIR Nodes like SIB, Swiss Institute of Bioinformatics, have been operating for many years, while other Nodes are currently in the process of being established and have received financial support through national funding agencies for this purpose.

The current twelve ELIXIR Members (Czech Republic, Denmark, Estonia, Finland, Israel, Netherlands, Norway, Portugal, Sweden, Switzerland, UK and EMBL-EBI) are represented in the ELIXIR infrastructure. A further six countries (Belgium, France, Greece, Italy, Slovenia, and Spain) are Observers and expected to become full ELIXIR Member States in the near future. Dialogue continues with several other countries to further expand the infrastructure.

An infrastructure of sustained and well-annotated data

Europe's future data infrastructure must ensure that the existing collective data capacity scale to meet rising demand (the "data deluge"). Dealing with these massive data flows requires that we establish common principles for optimising the use of existing data capacity (such as assessing which data should be stored and made available to users, agreements on storage and distribution of the core reference data sets, and so on). This distributed infrastructure must also be coordinated with data standards that enable full data integration so that the collective, expanding capacity across the continent is used optimally.

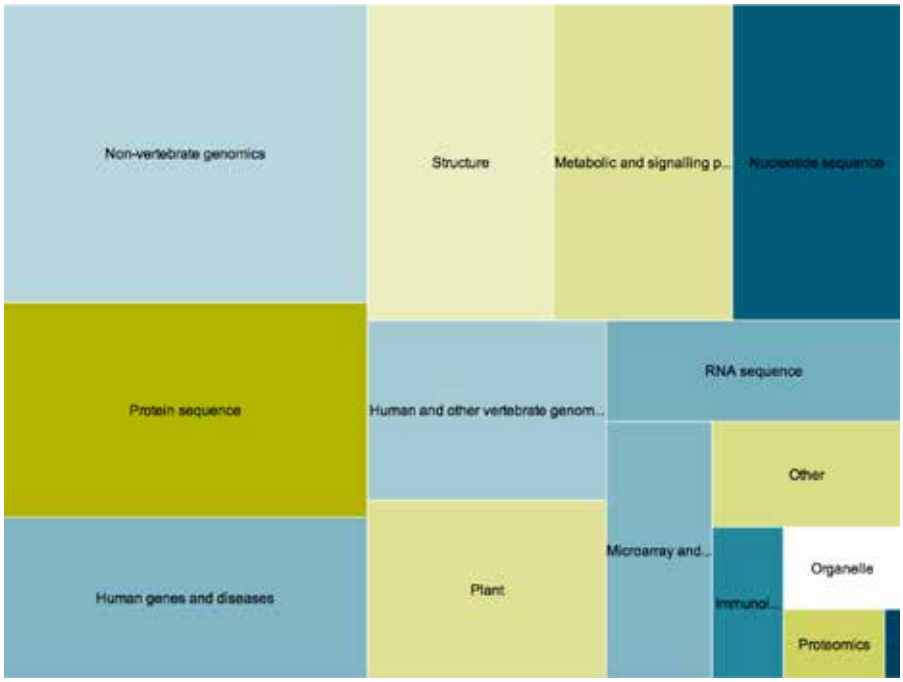

Figure 2: *Europe is home to a very large number of resources covering all aspects of life science*

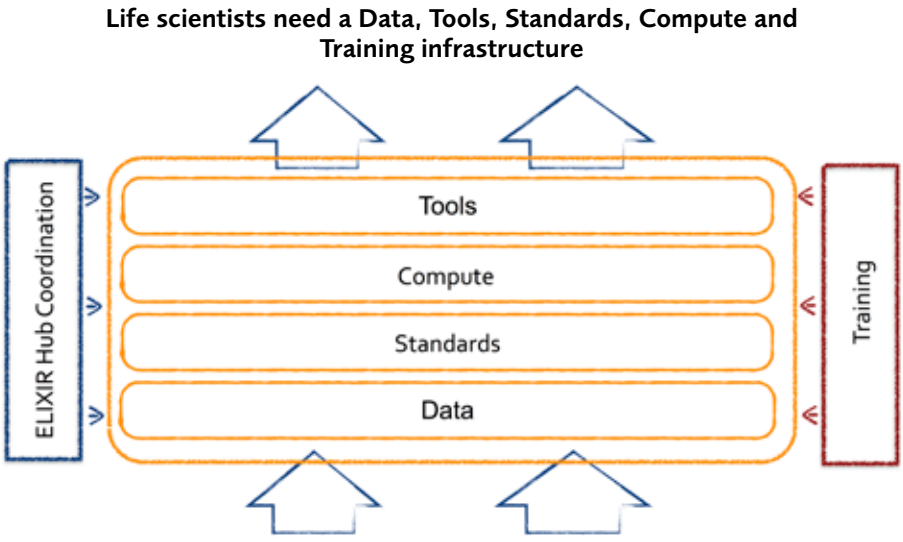

Figure 3: *A bioinformatics infrastructure for European needs. Services for Data, Standards, Compute and Tools powered by skilled people that operate and make use of the services (Training), and supported by an efficient and agile coordinating organisation (ELIXIR Hub).*

And, importantly, users of these data resources should be presented with a transparent interface to a distributed infrastructure (Figure 3).

A universally accessible data infrastructure requires standardisation and guidelines at many levels. Users – the scientists that deposit and access data -need experimental details to be reported for datasets (e.g. by following the well-established minimal information guidelines [14]) to guide deposition and facilitate exchange of the information. Effective interoperability of data also need agreement on nomenclature – harmonisation of names and symbols of biological entities [15] and the use of controlled vocabularies and ontologies to harmonise the terminologies used to describe database content.

Making use of the data: a Tools infrastructure

The data resources provided through ELIXIR will be used in many different ways using a large number of different analytical software tools written by investigators all over the world. Typical analysis use dozens of tools linked together through data processing pipelines. ELIXIR aims to provide an infrastructure that supports the community in managing access to and the discoverability of Tools (through our Tools and Data Services Registry developed by the Danish ELIXIR Node, see next article), encourage robustness and sustainability (through metrics recommendations and services) and support *community-led* efforts to benchmark and enhance scientific quality.

An infrastructure supported by Compute resources, storage

To date, life scientists have rarely needed to use the existing e-infrastructures in Europe, which have been mostly developed to provide services for the physical sciences. As a consequence, the existing compute architecture has been optimised for the requirements of physical science communities, which typically address compute-intensive problems.

In contrast, biologists require data-rich, massively parallel queries. Unless an increase of an order of magnitude in compute infrastructure capacity occurs (including storage and processor power) in a manner suitable to meet demand for life sciences, it is anticipated that the existing European compute infrastructure will not be adequate to respond to challenges driven by the data deluge. This will require enhanced capacity of the existing compute infrastructure to meet demands that evolve from the data deluge but also to maintain integration despite distribution, for instance by establishing agreements and processes that ensure that national and regional compute centres have access to central databases.

A training infrastructure for users

In the past, biological data resources were used by a relatively small community of 'bioinformatics aware' researchers, but as access to data becomes increasingly central to bio-medical research, this user base is growing and diversifying to include, for example, clinicians, experimentalists in the pharmaceutical industry, plant breeders and environmental scientists.

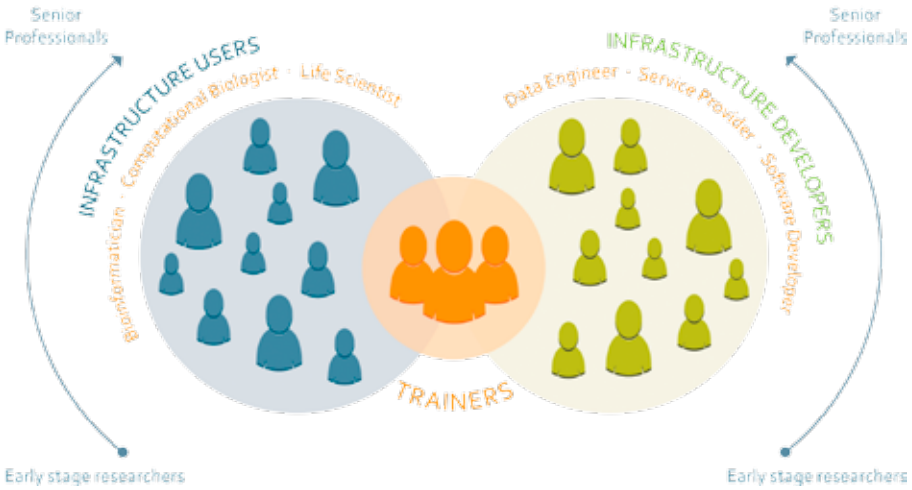

Figure 4: *A training infrastructure that link resource providers with the large and growing user community and provide training support through the researchers career.*

It is anticipated that the increase in demand for online bioinformatics services by new users will be accompanied by an increase in demand for training for the use of such services, and this has been evident by increasing demand for places at training courses organised by EMBL-EBI, SIB and other large national centres. The European training infrastructure is not adequate to accommodate this anticipated increase, this lack of capacity creates a significant bottleneck for users of life science data – as reference datasets become larger and richer they will also increase in complexity and require increased knowledge and skill to access and interpret.

By aligning training with national capacity building ELIXIR allows initiatives to collaborate, exchange training materials and open up their courses and events throughout Europe (Figure 4).

THINK GLOBALLY – ACT LOCALLY

Science is a global effort and the life sciences are no exception, indeed most of the flagship life-science research projects are major global endeavours: the wheat genome consortium, for example, has over 1000 partners in 57 countries. International collaboration is of paramount importance for effective exchange of data within these projects and includes not only technical standards for data exchange but also agreements on compatible legal and ethical frameworks for exchange on human data. ELIXIR brings together Europe's resources to ensure that Europe speaks with one voice on issues such as data standards, data-legislation and exchange mechanisms.

Building on the established links with partners outside of ELIXIR (such as other ESFRI RIs, IMI projects, and the NIH BD2K initiative in the United States) we ensure that ELIXIR solutions are well recognised and integrated into these global discussions.

REFERENCES

1. On not reinventing the wheel. Nature Publishing Group. 2012 Feb 27;44(3):233–3.
2. An Integrated Encyclopedia of DNA Elements in the Human Genome. NIH Public Access; 2012 Sep 6;489(7414):57. Available from: /pmc/articles/PMC3439153/?report=abstract
3. Network TCGAR, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Publishing Group. Nature Publishing Group; 2013 Oct 1;45(10):1113–20.
4. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Publishing Group. Nature Publishing Group; 2000 Jan 1;:–.
5. Gaulton A, Overington JP. Role of open chemical data in aiding drug discovery and design. Future Medicinal Chemistry. 2010 Jun;2(6):903–7.
6. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Research. Oxford University Press; 2014 Jan;42(Database issue):D285–91.
7. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2011 Dec 22;40(D1):D1100–7.
8. Boulton R. Science as an open enterprise. London: Royal Society. 2012;:104pp.
9. Gibney E, Van Noorden R. Scientists losing data at a rapid rate. Nature News. doi:10.1038/nature.2013.14416.
10. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. American Association for the Advancement of Science; 2015 Jan 23;347(6220):1260419–9.
11. Merali Z, Giles J. Databases in peril. Nature. 2005 Jun 23;435(7045):1010–1.
12. ELIXIR. D3.4: User Communities Report. elixir-europe.org.
13. D2.1: Database Provider Survey report for ELIXIR Work Package 2. 2010 Aug 4;:1–58.
14. Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nature Publishing Group. 2012 Feb;44(2):121–6.
15. Juty N, Le Novere N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Research. 2011 Dec 22;40(D1):D580–6.