Summary Report of the ELIXIR Industry Advisory Committee meeting

Hinxton, 26th of February 2015

This report is the public summary of the IAC's recommendations to ELIXIR following its meeting in February 2015.

The value to industry of Open Data resources

From an industry perspective research is driven by data, and with an increasing reliance on open data resources. Companies cannot do research by only looking solely at their own data or commercial data platforms. Public bioinformatics resources have a large user-base in industrial R&D.

The **consumer goods** industry is facing a challenging time in the safety environment with an end to animal testing of ingredients for some product types potentially impacting innovation of new products for some markets. There is an urgent need for improved tools to integrate mechanistic knowledge from chemical interaction to phenotypic output to aid in safety evaluation of novel chemicals, through which high quality external data can be combined with internal data. Likewise, there is the need to develop approaches by which such data can be used to conduct robust safety assessments that satisfy the requirements of both industry users and regulatory authorities. Such tools and approaches will be a vital enabler allowing all industry sectors to innovate in the future.

In both, the **agri-biotech and the pharmaceutical industry**, publically available databases such as ChEMBL/ SureChEMBL, Uniprot, PDBe and the omics databases are valuable in discovery and development research. For example, plant and pest genomes help companies to identify targets for new compounds in particular systems and add value to proprietary data by giving it a point of reference and augmenting the scope. Regulatory and safety aspects mean it is also increasingly difficult to get products to the marketplace, where there has been a change from risk-based to hazard-based assessment.

Open Data can help **publishers** produce more reliable, discoverable and innovative articles as part of their services to researchers. Large life science and bioinformatics databases have helped enable journals to more closely integrate data and peer-reviewed literature, such as through URI and accession number systems. These databases have also enabled mandating the archiving of certain data types, helping to increase the reproducibility of research. The number of Open Data repositories is growing and publishers are interested to work more closely with trusted repositories for Open Data to integrate data and literature peer review and publishing. Newer, more general repositories such as figshare, Dryad and Dataverse provide less standardisation and curation but fill a role in the open deposition of data that cannot be harboured in community repositories. Many publishers have now established clear deposition guidelines that cover both community archives and general repositories.

IT industries use open data resources as input of complex data analytics algorithms aimed to improve business processes, as an integrated part of their proprietary system or service offering like in case of genetic variant interpretation software, or to identify new business opportunities in any of the areas mentioned above. The common aim of most of the IT industries' use of open data resources is to convert data collected from heterogeneous repositories into valuable knowledge.

The need for coordination

The IAC considers that the fragmentation of bioinformatics resources is neither optimal nor sustainable. It is hoped that ELIXIR will help harmonize the informatics landscape and stimulate the R&D community to move away from the do-it-yourself attitude of creating its own unconnected resources, and make better use of open source resources. Ensuring the scientific community is aware that reliable high quality, up-to-date resources are already out there, fostering a culture of better cooperation between industry and academia, and changing the perception that 'Open Source' is not industry-standard practice are all valuable objectives that ELIXIR should strive to meet.

The coordination should be addressed from three different leverages:

ELIXIR plays a key role in the certification of existing data resources and the creation of new ones. With its vast diversity of participants, ELIXIR is in a unique position to take the lead in setting and encouraging the adoption of standards and ontologies to improve interoperability between datasets. The ELIXIR 'quality stamp', through the Core Resources and Named Services, will ensure that industry can trust the quality and sustainability of these resources.

Open data provides a significant opportunity for industrial research and innovation. Industrial researchers need to partner with the research community to help develop and promote standards as well as the exchange of ideas. The Reproducible science continues to be an area of great interest.

The rapid growth of large and heterogenoues datasets produced and used by large and distributed research communities also creates significant opportunities for supply side innovation in the ' data enabling industries'. This includes scalable cloud and data storage solutions but also middleware and tools that ease the service access by providing friendly interfaces. These tools will be based on the interoperability standards collected from data consumers and accepted by ELIXIR and will isolate the user from the computational hurdles

ELIXIR is encouraged to help users embrace this data by ensuring that it is accessible not just by bioinformaticians but by all life scientists that need to use it. Close interaction between data scientists and end users will add value to the broad ELIXIR community and boost ELIXIR activities.

High quality manual curated datasets (e.g. UniProt/Swissprot) are highly valued by industry. ELIXIR's efforts in software carpentry and development of gold standard curation will be important in ensuring quality of data.

Recommendations for providing services to Industry

Being able to trust resources is crucial (this was a clear recommendation from the ELIXIR Preparatory Phase WP3 Industry Report). Resources must be sustainable, and quality assured, with a high level of integration between data sets, industry level standards and 24/7 reliability, and fast access to the data they need.

Ensuring long term sustainability

The IAC encourages ELIXIR to consider different funding models to ensure the long-term sustainability of ELIXIR's Core Resources and would wish to follow the work of the Working Group established to consider the Long Term Sustainability of ELIXIRs Core Resources, including providing input into the group's final recommendations.

Promoting use of the cloud

Many companies do not have capacity to store their data in house or cover the cost of its upkeep; whereas in the past industry downloaded the data it needed, in the future, as the volume of data becomes larger still, this model may no longer scale so there is an increasing need to put industry data alongside public data. e.g. Embassy cloud model.

Companies are also growing more comfortable with using cloud providers. However, the cloud model is generally still under-utilised by the research community. ELIXIR should play an important role in influencing researchers to make better use of the cloud through helping to address security issues and mistrust of cloud services, and explaining the benefits of distributed computing and how it could better suit their needs. ELIXIRs Authentication and Authorization Identification (AAI) model and the ELIXIR Service registry will be components of this.

Role of web services companies

There is a growing 'research informatics' sector that acts as 'scientific enablers', where they help other companies access data resources. For these companies transparency is important, as is understanding what public resources are available.

ELIXIR should make the processes for accessing its services clear and ensure that checklists for data transfer are agreed.

The IAC encourages the development of the emerging 'ELIXIR Innovation and SME forum', and will review the output of the events with interest.

A high level of interoperability, stability of services and good APIs are also important to ensure interfaces can be built for customers on top of these services that do not break or need to be modified at a later date.

The IAC hopes that the ELIXIR Node network will help provide a point of contact in each country that they can work with.

An information event specific to the research informatics service sector - software developers, HPC and cloud service providers should be considered.

Encouraging industry to share its data

Companies are getting a better understanding of what is pre-competitive, and industry should be encouraged to share their own data where possible. ELIXIR could have a role in highlighting the added value that can be gained from allowing other people to use their data and add to it.

The IAC recommends ensuring versioning systems are in place that will enable correction to the permanent record to be made at a later date should they be required, as this is currently difficult to do once the data has been made available externally.

Improving links between research literature and associated digital objects

The IAC would like to see the use of Digital Object Identifiers (DOIs) extended to other resources, in order to help maintain links between research literature and associated digital objects. Further, it would also like to see closer integration between databases and literature repositories to be able to make more sensitive data available in a public way. ELIXIR should support this.

Benefits of Public Private Partnerships

Many companies are already partners in ELIXIR Nodes (ELIXIR NL and ELIXIR DK for example) and the Public Private Partnerships concept is crucial in facilitating a closer integration between industry and academia. In addition, many SMEs active in this field are spin-outs from either public resources or private companies, so naturally lend themselves to close collaboration with academia. PPPs also provide opportunities for web services and HPC suppliers.

ELIXIR can aid growth of SMEs by allowing them to build tools on top of high quality open data resources e.g. in the area of Next Generation Sequencing there is a healthy eco-system of companies across Europe whose whole business model is based around providing commercial services on the public data infrastructure.

A service model could be envisaged by which ELIXIR open source tools developed by academics are handed over via a support group for product development and made available commercially.

Provision of training to industry

Training is important to ensure scientists are able keep their analysis and data management skills up to date, and due the growth in use of biological data, more people need to be trained in approaches to manipulating and aggregating data. ELIXIR has a key role here.

The ELIXIR IAC will continue to monitor the emerging ELIXIR training programme to ensure it is appropriate and relevant.

Training opportunities can also be opened up though industry exchange schemes between ELIXIR partners and industry (such as the Marie Curie programmes for staff exchanges) and should be considered.

Measuring success

Considering how to measure the impact of ELIXIR on industry is important and something the IAC will continue to consider going forward, alongside the ongoing development of ELIXIR metrics.

Two potential measures of success would be to look at efficiency savings companies make from accessing public resources i.e. how much the company would need to spend to curate and maintain the database in house, multiplied by the number of companies that are using a particular resource; and how many derived datasets can be created. Increased citations of datasets, made accessible via ELIXIR, in the peer-reviewed literature may also be a measure of success for ELIXIR as a whole.

Improving communication with industry

The IAC recommends an ELIXIR communications strategy developed for engagement with industry, in order to ensure industry partners are informed of the latest developments in ELIXIR services and emerging standards. Early engagement with the IAC on ELIXIR grant proposals is recommended, in order that it can make recommendations on the outcome and provide input into the business case.

The IAC also foresees a role in helping ELIXIR shape future calls and identifying relevant calls and opportunities for European funding.

Building links with Industry relevant initiatives

The IAC recommends ELIXIR establish links and develop connections with the following initiatives.

In the area of health and pharma:

- Ongoing and future IMI and IMI2 projects, particularly those relevant to knowledge management and the long-term sustainability of data.
- Pistoia Alliance, particularly with regard to the development of ontologies and standards.
- SAGE bionetworks and the tranSMART foundation.

In the area of agriculture and food:

- The BioBased Industries consortium.
- The iPlant Collaborative, which is actively seeking plant science projects to host.
- Terragenome, the International Soil Metagenome Sequencing Consortium.

In the area of consumer goods:

- The Data Infrastructure for Chemical Safety consortium (diXa)
- The NIH Human Microbiome project
- The Personal Care Association, Cosmetics Europe

The Public Health and Safety Organization, NSF is cited as relevant to all the areas above.

In terms of developing links with SMEs:

- Regional bio-clusters and Brussels based bodies such as the European Trade Association for biopharmaceutical companies (EBE).

In terms of developing links with the publishing community:

- Adopt Force11 principles on FAIR data publishing
- Support Force11 Joint Declaration on Data Citation Principles (already endorsed by EMBL-EBI and many publishers and repositories)
- Digital Object Identifier (DOI) issuing bodies such as CrossRef (DOIs for articles) and DataCite/California Digital Library (DOIs for datasets). (CrossRef and DataCite announced a collaboration in 2014 to look at links between data and literature).
- Consider working more closely with publishers to enable better integration between data and literature and integration of data and manuscript peer review. Repositories such as figshare, Dryad and Dataverse are designed to provide services to publishers/journals. Evaluating or working with these resources could inform future developments of bioinformatics databases to support data-literature integration. Particularly important for growing number of data journals/data articles, such as Scientific Data plus offerings from numerous other publishers, where peer review of data is more closely integrated with manuscript review.
- Cross reference with ORCID persistent digital identifiers for researchers
- Look at the work of the NCSA National Data Service Initiative, in particular 'Open Linked Data Repository for Article-Data Associations' (OLDRADA)
- Enabling peer review and publication of articles linked to sensitive or clinical research data that are restricted access. A working group on this issue will report later this year.

In terms of building links with the HPC and web services community:

- The European Technology Platform for High Performance Computing (ETP4HPC)
- ITEA, which supports, for example, ParMA and H4H projects
- FP7- supported projects like Mont-Blanc and Fortissimo. ELIXIR users could benefit from the already existing scientific cloud (Helix Nebula).
- The Big Data Value (BDV) Public Private Partnership
- Partnership for Advanced Computing in Europe (PRACE).