

### 3 ELIXIR Technical Services

**b. Remit.** *The role of the ELIXIR technical and e-Infrastructure is to provide services for the present and future specialised biomedical data service providers. Technical services must be sustainable and scalable so that data specialists can rely and build on them. The goal is to go forward in close collaboration with the ELIXIR tools and interoperability work streams. Overall ecosystem will enable European researchers to access and use increasing digital output of global biological research.*

Technical services Work stream aims for a transforming solution of wide-spread utility in the life-science domain for accessing and computing with the large data-volumes from modern biology. The community continuously develops scientific software services to meet the biological end-user data analysis challenges. Encapsulation of this specialized know-how in virtual machine templates on top of cloud service offering is a good approach. Building, maintaining and supporting scientific software environments in the domain-specific virtual machines is a highly technical task. The skills for this, however, do not necessarily exist on the technical service providers' side and will be made in a bi-directional exchange with the tools and data interoperability experts. A good forum for these discussions is the ELIXIR node technical coordinator group.

Collaboration with biobanking infrastructures at both the European and national level ensures up to date knowledge on requirements for secure access and exchange of data. ELIXIR promotes federated authentication to enable governance processes and access to personal genome data through institutional logins and automated interaction with the appropriate data access committees. ELIXIR will also collaborate closely with the European e-Infrastructures to address user needs and fully integrate into the European technology landscape. The role of the work stream is catalyse the uptake of the possibilities of the e-Infrastructures in life sciences, preferably in a context of a project addressing a challenge formulated with research funders and scientists.

Sustainable models for resource management and exchange will interface Work stream in the ELIXIR Management Programme.

**c. Overview of services.** *The overall ELIXIR infrastructure should achieve integration despite distribution, with access to the central databases in such a way such that the collective expanding capacity across the continent is optimally used. Some nodes deliver technical services to themselves to succeed in delivering data services to end-users, while other nodes may deliver technical services to the data specialist nodes for them to succeed in the same goal. Bioinformatics tools & data nodes provide bioinformatics tool environment building support. Service change management will be a challenge and mechanisms to identify and take into use cross-cutting technical services are needed.*

- **Physical ICT storage and compute** with virtualisation to bring it near data
  - Emphasis to computable storage (IOPS, bandwidth, capacity)
  - Enabling distribution & federation
  - Collaborate to optimise the resources, carries a major recurring investment
- **Authentication and authorisation AAI Services**
  - User authentication and attribute management services, community membership services ("bona fide biomedical researcher") and access rights management services
  - Collaboration with NRENs (National Research and Education Networks) identity federations and the eduGAIN interederation service of GÉANT
- **Network solutions** with emphasis to enabling distribution and data security
  - Node to Node and Node to Internet connections. Permanent and on-demand. Recommendation is to have readiness for dedicated high-bandwidth (10 Gbps) connections for Node to Node/Hub links
  - Private network and secure protocols to meet data protection requirements. Connections to other relevant biomedical data service providers
  - Collaboration with NRENs and GÉANT, looking especially for advanced services
- **Data services support**
  - Data transfer services, reliability; integration with network services
  - Immutable data identification, annotation, metadata
- Technical service **life-cycle management** that considers impact on the specialised biomedical services.
  - Scientific services could have life of 12-120 months and evolve with the research trends, virtual machine approach is appealing both to enable flexibility and scalability.
- **Single point of contact** for services
- Service Web page, or other gateway(s) to services with **Service level** agreements/descriptions/declarations
- **Technical training**

**d. Relevant, similar or complementary services beyond ELIXIR and Europe.** *ELIXIR must learn from the history of the existing e-Infrastructures: many key concepts of technical services have already been invented. There are, of course, differences e.g., raw sequencing data "atoms" are much bigger than what is a typical data set in other fields. Goal is to recognize technological overlap at the infrastructure level and deliver a plan when it will be possible to build on existing e-Infrastructures for constructing ELIXIR.*



- **Network and AAI.** Network provides key infrastructure components: GÉANT, NRENs (National Research and Education Network providers). ELIXIR nodes need high performance connections to the Internet for end users and between technical services of the nodes that transport data and virtual machines. Specific services by GÉANT, NRENs and global collaborators, like eduGAIN, are a way to bridge between national networks and identity federations. Target is to, for example, enable exchange of services and reliably identify users and user attributes like data access permissions and digital identifiers (ORCID).
- **HPC.** PRACE offers state of the art HPC technology. The current use of PRACE services in biological RI is massive molecular simulations with software that fit the technical platform. Data-intensive computing capabilities would make PRACE more attractive for ELIXIR. It would be desirable to elaborate ELIXIR requirements for IOPS-intensive computing for the forthcoming HPC Center of Excellence call.
- **Distributed ICT.** EGI and OSG in US have services that can help operate and monitor distributed physical RI of ELIXIR. Some legacy technical choices like user certificates have small uptake in the biomedical community. Nevertheless, EGI currently moves towards federated identities and virtualisation of the NGI resources coordinated by EGI is ongoing. Altogether EGI.eu has extensive experience about distributed large scale infrastructure which could be of use for ELIXIR.
- **Data.** EUDAT is reaching maturity and shares similar values with ELIXIR. However EUDAT is a generic data RI and ELIXIR has a biomedical data focus. Research Data Alliance (RDA) aims to be a standardizing forum at a global level regarding data transfers and exchange. We recommend clarifying the area of collaboration by interacting with RDA and forming e.g. an ELIXIR pilot study with generic data e-Infrastructures.
- **Data integration.** BioMedBridges develops infrastructure to allow data integration across the biological, medical, translational and clinical domains. ELIXIR technical service Workstream should closely follow BioMedbridges requirements.

#### e. Priorities, challenges, gaps.

- **Priorities.**
  - Trust-building on technical service components produced by the distributed nodes and hub.
  - Enabling scale-out of services across country borders. Action: lobby for a call to address this.
  - Life sciences need to increase technological overlap with the existing solutions at the infrastructure level.
- **Challenges**
  - Lack of ICT experts that understand biological sciences.
  - Shortage of physical storage capacities that allow computing over large datasets.
  - I/O bottleneck of distributed systems over the network. Mounting storage over the high performance connections will add additional load to storage elements that not necessarily have sufficient surplus capacity to handle that load. Increased latency over the lightpath connection might negate some of the gains of distributed computational capacity. Mitigation: Smart caching of data, replication, and novel algorithms.
  - Privacy of substantial parts of the dataset limits possibility to actually move data. Mitigation: move computing to the data (need compute capacity near to storage), novel algorithms (potential for use of cryptographic algorithms for aggregated computing over distributed immovable dataset).
  - There is a danger of trying to cover too much ground. What are the key things that the technical group is going to do in timeline (process and high-level technical goals) and what input/interaction is needed from elsewhere to do this (NRENs, EGI, PRACE, etc.).
- **Gaps**
  - Cross-domain processes and policies and expertise: technologies exists, but are not effectively deployed for life science data analysis
  - Funding agencies lack instruments to support the entire research ecosystem including infrastructure. For example: Network costs are not budgeted in research grants and long-term data preservation costs are not coupled to a decision to create research data.
  - Legislation for the exchange of sensitive human genetic data is currently developing. Thus security policies that enable shipping of data do not exist. This is more a matter of trust building and compatibility of policies than a technical issue. Mitigation: ELIXIR pilots with non-human data to show technical feasibility and provide social engineering to build the trust is a step towards working with human data.
  - The security level provided by common AAI services is not well defined. It needs to be made sure that the security level of the AAI service provided by the common AAI services like eduGAIN meets the security requirements of ELIXIR. Trusted identification of ELIXIR users is needed for providing authorization services, which in turn is needed to streamline access to sensitive data.
  - Interoperability between Cloud Software Infrastructures is a bottleneck to share virtual machines. Industry is already providing tools to convert virtual machine images, these could be used to build a workflow to convert the images on the fly when necessary. Mitigation: ELIXIR pilot for federation of cloud resources was selected among the winners in the global Enlighten Your Research Competition. Use of external know-how from EGI, HelixNebula etc.

#### f. Objectives and milestones 2014 - 2018 (M1-M60)

M3 Technical service delivery declarations from nodes are collected and "a map of European offerings" is made based on them

M6 New ELIXIR pilots started

M12 Highlighted prime examples of ELIXIR specialised biomedical services between several nodes created from ELIXIR pilots

M12 Report on life sciences technological overlap with existing e-Infrastructure solutions at the infrastructure level

M24 ELIXIR technical infrastructure plan, a document written with the European e-Infrastructures

M 36 ELIXIR technical services life-cycle management process

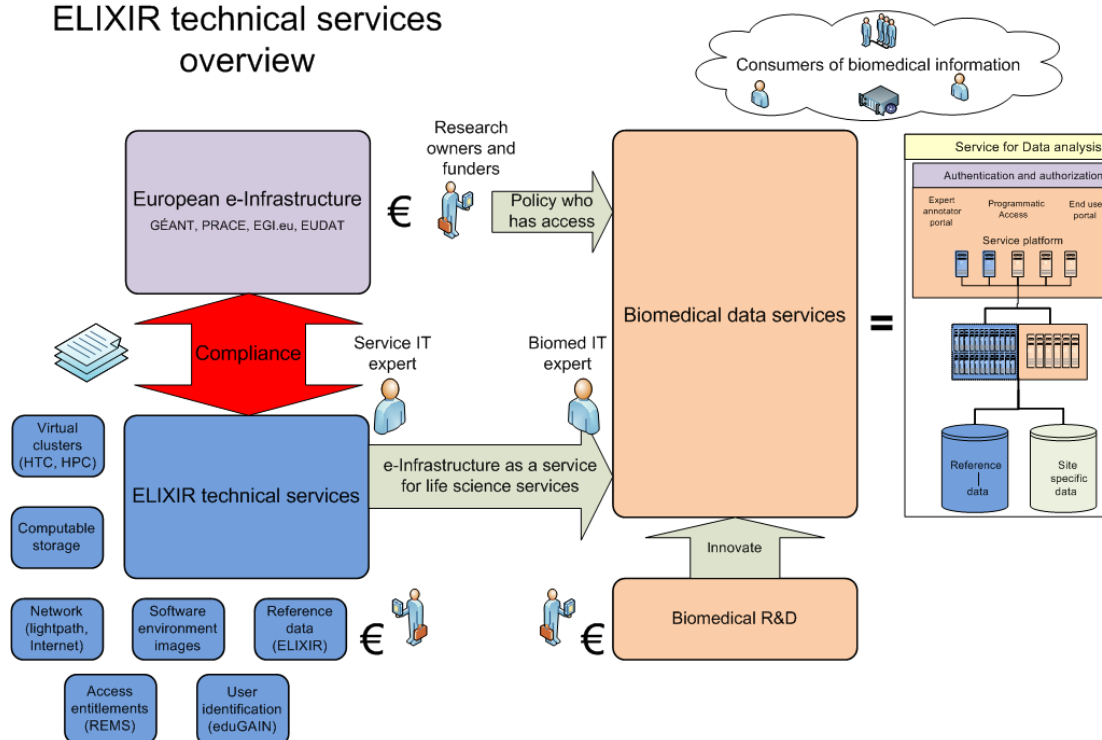
For example, network implementations connecting the ELIXIR nodes and hub are synchronised with the construction of the other parts of the technical distributed infrastructure of ELIXIR. Network upgrade is taken into account when planning upgrades regarding the other parts of the technical infrastructure of ELIXIR like computable storage and cloud compute capacities.

...



M60 ELIXIR service providers are specialized. E.g. ELIXIR life science - aware cloud is provisioned across Europe. Distributed nodes trust each other and build national and regional scientific infrastructure services on each other's services.

## ELIXIR technical services overview



### Background

ELIXIR preparative stage reports <http://www.elixir-europe.org/prep/reports>: Summary [Recommendations from ELIXIR preparative stage](#) from the perspective of Data Resources, User Communities and Technical feasibility of using supercomputing centres (2008-9).

[ELIXIR technical pilots. Summary](#). Key words scalable, life-science data access, high-performance computing, interoperability of public biological and biomedical data resources.

[ELIXIR business case. Summary](#). Technically effective and economically efficient pan-European infrastructure should comprise a data infrastructure, a compute infrastructure, a training infrastructure and a tools and standards infrastructure.

Enlighten your research proposal by ELIXIR was selected to top four in the [SuperComputing 2013 conference](#). Networking solution for data and virtual machines transport will be developed in collaboration with the NRENs based on the document.

[ELIXIR authentication and authorisation pilot report](#). Data access and user identification are a key area in H2020 and are focused to user communities.

[Outline for technical services](#) from the 1st ELIXIR Head of Nodes meeting.

Known technical service offerings by ELIXIR nodes. ([ELIXIR www node overviews](#))

ELIXIR 2014-2018 workplan draft.

#### 1. a. Team of relevant Node staff/experts contributing to the work stream

Requestor: Niklas Blomberg, ELIXIR director & ELIXIR Head of Nodes committee

Working group (Main responsibility to deliver, viewpoints)

FI: Tommi Nyrönen (coordination), Mikael Linden (AAI), Olli Tourunen (cloud), Juha Oinonen (network) (e-Infrastructure, coordination) ELIXIR FI

CZ: Ludek Matyska, (coordination), Michal Prochazka (AAI), Mirek Ruda (cloud and grid), David Antos (data), Petr Holub (network) ELIXIR CZ

ELIXIR Hub: Rafael Jimenez (bioinformatics tools)

Reviewers (Viewpoint on reviewing)

EE: Jaak Vilo, UT ELIXIR Estonia (bioinformatics tools)

EMBL-EBI: Steven Newhouse, EMBL-EBI (e-Infrastructure)

DK: Peter Loengreen, DTU ELIXIR Denmark (coordination, tools PoW)

NO: Lars Bongo, Univ of Tromsø ELIXIR NO (bioinformatics tools)

ES: Josep Lluís Gelpi, ELIXIR ES

ES: Modesto Orozco, BSC ELIXIR ES

IT: Giorgio Maggi (INFN), ELIXIR IT

IT: Gianluigi Zanetti (CRS4), ELIXIR IT (bioinformatics tools, cloud infrastructure)

Reflection Group: ELIXIR node technical coordinators email list