ELIXIR CZ Strategy 2020-2025

The Czech Infrastructure for biological data ELIXIR CZ is currently at the beginning of a new stage of its operation, and faces new challenges generated by an increasing number of users and the demands of new technologies in life science disciplines. It is no longer merely a problem of data deluge, but a quite complex issue that creates a great opportunity for general solutions which are interoperable, robust and user oriented. ELIXIR CZ, as part of the pan-European infrastructure ELIXIR, will benefit from the excellent level of its globally recognized expertise. This movement needs a well-formulated strategy for the development of ELIXIR CZ, and maintaining its dedication and compatibility with European partners and global challenges. Reflecting the evaluation of the ELIXIR CZ Scientific Advisory Board, we propose a strategy of infrastructure development based on identified strengths and expertise together with a recommendation of fields for further development. The strategy will serve as the basis for the ELIXIR CZ Scientific Program for the 2020-2025 period, plus also key activities that are critical for the fulfilment of its infrastructure ambitions.

Vision of ELIXIR CZ:

"The vision of ELIXIR CZ is to build a sustainable Czech infrastructure for biological information, supporting life science research and its translation to medicine and the environment, the bio-industries and society"

Mission ELIXIR CZ:

"Provide valuable bioinformatics resources to the Czech and international user communities and consolidate the Czech bioinformatics community with the European activities of ELIXIR."

Motto of ELIXIR CZ Strategy 2020-2025:

"Growing on existing strengths, establishing new fields of expertise"

ELIXIR CZ is composed of 14 partner institutions coordinated by the Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences (IOCB). Being a distributed infrastructure, it covers institutions from across the Czech Republic and provides critical mass of bioinformatics knowledge by integrating them into a single entity to serve a wide Czech and international user community.

ELIXIR CZ is highly respected for its strengths in structural bioinformatics, chemical biology, genomics and data management. This can be documented by an increasing number of users and, consequently, by

an increasing number of publications prepared with the tools and services provided by ELIXIR CZ. It is also underlined by the key role of ELIXIR CZ in establishing the European ELIXIR 3D-Bioinfo Community. Tools co-developed by ELIXIR CZ such as Data Stewardship Wizard or ELIXIR AAI (Authentication and Authorisation Infrastructure), have been considered as Commissioned Services by European ELIXIR.

Through the implementation of this Strategy, ELIXIR CZ is going to contribute significantly to the fulfilment of the national innovation strategy (including national scientific policy) entitled "Czech Republic: The Country For The Future" in several ways: Active engagement within the pan-European ELIXIR family will contribute to the integration of Czech science into the European Research Area and to the visibility of Czech Science, including participation in Horizon 2020 and Horizon Europe funded projects. By cooperating with other ELIXIR nodes, international cooperation and the excellence of Czech science will be strengthened. The international environment will also help Czech Industry to be more competitive internationally. Finally, through engagement in strategic initiatives related to data management, 1 million genomes, the European Open Science Cloud or structural bioinformatics, chemical biology and others, it will contribute to building new competences in Czech science and industry, and thus contribute to its long-term competitiveness.

ELIXIR CZ has the ambition to facilitate the interoperability of data generated by other life science infrastructures, to provide know-how on data management based on approaches and tools developed by ELIXIR and help with data sharing by applying the FAIR principles. ELIXIR CZ has signed a Memorandum of Understanding with most of the Life Science research infrastructures on the Czech infrastructure roadmap, which also anticipates its pivotal role in activities connected with the European Open Science Cloud (EOSC). According to the proposed strategy, ELIXIR CZ primarily established vital collaboration with CIISB – infrastructure for integrative structural biology, CZ OPENSCREEN – infrastructure for chemical biology and genetics, BBMRI-CZ – infrastructure of biobanks and biomolecular resources, EATRIS-CZ – infrastructure for translational medicine, and last but not least, the CCP – infrastructure of Phenogenomics. Besides the Life Sciences sector, a vital collaboration was also established with RECETOX RI – research infrastructure for the management of environmental and health risks related to chemical compounds produced by industry.

As the recent pandemic of COVID-19 clearly shows, the major driving force in research in general is data. Reliable, accessible, useable and most importantly – open. We not only need fast access to data worldwide, but also common standards providing data interoperability for enormous volumes of data. The scientific community will more and more rely on an integrative environment where results are connected with experimental details, literature references, the original source of the data and glued together by common standards in all of the integrated parts. The benefits for users are enormous. On the other hand, we can expect more and more requirements for automatic access to the data, their correct annotation and sustainable archives capable of handling an increasing number of user queries. ELIXIR CZ is a leader in all the above-mentioned user needs and is working in collaboration with its international partners on sustainable solutions.

Strategic areas of ELIXIR CZ strategy

There is a constant development in all areas of bioinformatics. ELIXIR CZ wants to stay at the forefront of such developments and be involved in the emerging areas, which are identified as priority to the Czech and international scientific community. Areas of existing strengths of ELIXIR CZ were identified by its representatives and confirmed by ELIXIR CZ Scientific Advisory Board (SAB) meeting in October 2019. We not only want to cultivate our existing strengths but also contribute to other areas with significant developments. Dramatic progress has been made in the area of human data and personalized/precision medicine. ELIXIR CZ will attempt to also develop activities in this field further and use the expertise offered by other ELIXIR Nodes across Europe.

ELIXIR CZ is committed to building on its strengths and developing in the following areas of strong expertise during its 2020-2025 strategy cycle:

1. *Structural bioinformatics*

Current status

Structural bioinformatics is an important contributor to understanding the molecular basis of biology and medicine, since structure(s) define the function of each biological agent. The resolved biomacromolecular structures are stored in worldwide Protein Data Bank (wwPDB), whose European part PDB Europe (PDBe) is an ELIXIR core data resource. Biologically important data related to the structures are accumulated in PDBe-Knowledge Base (PDBe-KB).

Structural bioinformatics is traditionally strong in the Czech Republic due to historical links to the PDB database, top-class X-ray, NMR and Cryo-EM groups, and groups developing structural bioinformatics software tools and databases used worldwide. ELIXIR CZ initiated the formation of the ELIXIR 3D-Bioinfo community and is deeply involved in the activities of this community.

Key domains and services in ELIXIR CZ are:

- **The visualization of biomacromolecular structures**, including their experimental data and annotations. In this field, we develop the LiteMol suite (and its follow-up Mol*) for the visualization of biomacromolecules and CoordinateServer, DensityServer and PatternQuery for fast data delivery.
- **Analysis of biomacromolecular structures**, e.g. the detection, characterization and storing of information about channels and pores in biomacromolecules (MOLE, Caver, CaverDock, ChannelsDB), prediction of ligand binding sites from protein structure (PrankWeb) and the validation of biomacromolecular structures (ValTrendsDB, ValidatorDB, MotiveValidator, and DNATCO.org).

- **Protein engineering tools**, supporting the design of protein variants with improved functionalities. Specifically, we develop tools for the construction of smart libraries for the screening (HotSpot Wizard), improvement and evaluation of protein stability (FireProt), solubility (SoluProt), and their effect on function (PredictSNP), and newly also on personalized medicine in oncology (PredictSNP onco).
- **Nucleic acid structure research**, e.g., the formulation of community-accepted benchmarks that integrate the different levels of nucleic acid structure descriptions and the development of tools for the modelling, refinement, validation, and annotation of nucleic acid structures. We provide tools at the website DNATCO.org for the analysis of nucleic acid structures. Moreover, we focus on the examination of structure-function relationships in RNAs with special attention on sn- and ncRNAs and develop the tools rPredictor and cpPredictor, rboAnalyzer for this field.

Some of the ELIXIR CZ services are used daily by the life science community, and they are a direct part of PDBe or are integrated into PDBe-KB or PDBsum (i.e., LiteMol suite, Mol*, ChannelsDB, PrankWeb, ValidatorDB). The remaining services deal with PDBe and PDBe-KB data. The providers of these services are integrated or collaborate with the 3D-Bioinfo community.

ELIXIR CZ will serve as a platform for the Czech structural bioinformatics community where the national groups can work together to strengthen international visibility, training and established collaborations with PDBe and other ELIXIR core resources.

Challenges and goals of ELIXIR CZ

Currently, the main challenges in structural bioinformatics are dealing with extra-large biomacromolecular structures (e.g., viruses, hybrid models, organelle models) including additional experimental data, the integration of huge sets of information about structure properties via structure annotation, and interconnection with other structural databases (ligands, nucleic acids, glycans, etc.).

ELIXIR CZ strategic goals in Structural Bioinformatics are as follows:

1) Goal 1: Evolve and progress the key ELIXIR CZ domains and services by:
    - Raising the quality of existing tools and databases to enrich their robustness, stability, interoperability, FAIRification and effectivity
    - Extending ELIXIR CZ tools and databases to handle the above-mentioned challenges, i.e.:
        - Ability to process extra-large and hybrid structures
        - Ability to utilize annotations consolidating results from multiple methods via collaboration with ELIXIR core resources such as PDBe-KB
2) Goal 2: Work on activities and the growth of the ELIXIR 3D-Bioinfo community in accord with the 3D-Bioinfo White paper and interact with its members via publications and joint grants
3) Goal 3: Synergy identification and branding:

- The identification of strong synergies between ELIXIR CZ structural bioinformatics research groups and their activities in the context of their national and international connections to research communities (e.g., Instruct-ERIC, iNEXT-Discovery, EU-OPENSCREEN and BioExcel) and ELIXIR platforms and communities
- Branding of ELIXIR CZ structural bioinformatics services (e.g. their identification on the ELIXIR CZ website, at bio.tools and on the webpages of the tools, joint scientific papers about workflows, inputs in social media)

2. Chemical biology

Current status

Chemical entities of biological interest – such as small molecule metabolites – are essential building blocks of biological systems with crucial roles in human health and disease. Structural classification of these chemical entities is the first step to understanding these roles, but the number and complexity of known and possible chemical structures make any manual classification effort unfeasible. The chemoinformatics characterization of small molecules (e.g., prediction of their key properties and calculation of their descriptors) enhances their utilization in other molecular biology fields.

To effectively utilize large databases of chemical entities in biology, users need well-designed interfaces that guarantee the interoperability of services and seamless processing of chemical data. This is especially important for analyses that concern e.g. drug design and ligand docking, which should employ robust, reproducible and well-benchmarked screening methods.

Key domains and services in ELIXIR CZ are:

- **Support and development of federated services at a global scale**. ELIXIR CZ has a history of building interoperability tools for the utilization of major chemical databases (including ChEBI, ChEMBL, PubChem and PDBChem) in a larger context. The ELIXIR CZ development team has recently focused on fast and user-friendly tools for the similarity- and substructure-based retrieval of chemical entities and building SPARQL endpoints for federated use by other services.
- **Classification of chemical entities**. In collaboration with ELIXIR CH, ELIXIR CZ is currently developing open source tools for the automated structural classification of chemical entities using the reference ontology ChEBI). ChEBI is an expert-curated ontology of small molecules of biological interest, mainly small-molecule metabolites and sugar polymers.
- **Integration of chemical biology resources**. ELIXIR CZ is cooperating with EMBL-EBI to develop services for the chemoinformatic characterization of small molecules using methods derived from QSAR/QSPR descriptors. ELIXIR CZ also integrates selected large databases of small molecules into available computational tools and workflows to make the use of these databases more comfortable and convenient.

Challenges and goals of ELIXIR CZ

In a pan-European context and with strong collaboration with Swiss, UK and EBI partners, ELIXIR CZ is ready to answer new challenges in the development of specialized workflows and focus on creating tools that map chemical structures to arbitrary classifications and characterization. This is absolutely necessary for the successful integration of the chemical space into ELIXIR core data resources and other globally available resources.

Most of the efforts will be focused on semantic interoperability, which defines current trends in chemical biology. This interoperability provides an essential bridge between pure biological disciplines and the chemical space. Several labs of ELIXIR CZ are participating in this trend with IOCB playing a leading role, in which it will be responsible for assisting the smooth cross-discipline utilization of various tools developed for structural bioinformatics, proteomics, metagenomics and human genomics.

ELIXIR CZ strategy goals in Chemical Biology are as follows:

1) Goal 1: Integration of chemical datasets with any resources using ChEBI ontology
2) Goal 2: Interoperability with other ELIXIR Core Data Resources via SPARQL endpoints, making federated cross-database queries possible
3) Goal 3: Application of the new results to workflows for automated drug design utilizing chemical information. The interoperability will provide relevant information to be automatically used for the "in silico" screening of a large body of chemical information
4) Goal 4: Seamless extraction of useful information for computational methods at the interface of proteomics and cheminformatics, supported by specialized software tools for proteome profiling

3. Genomics

Current status

Thanks to the recent advances in sequencing technologies and bioinformatics, the field of genomics has entered the era of ambitious projects such as the Earth BioGenome, aiming to decode the genomic information of all living organisms. Along with the vast impact these efforts will have on science and society, there is now an unprecedented demand for the technical resources and protocols needed to handle, share, and explore the resulting data. While ELIXIR-CZ activities in genomics reflect major trends and challenges in the field, the consortium is aware of the risk of diluting its efforts should it aim to cover the full breadth of the discipline. Instead, it will focus on supporting the selected research areas outlined below that have a long tradition of excellent research work and an active scientific community in the Czech Republic.

Key domains and services in ELIXIR CZ are:

- **Annotation of repetitive DNA elements in eukaryotic genomes**. There are ongoing activities to develop and make publicly accessible various computational tools for identifying and annotating repetitive DNA elements in genome assemblies or next-generation sequencing reads (RepeatExplorer, TAREAN, and DANTE pipelines). Reference databases of repetitive DNA include, for example, the database of mobile element protein domains (REXdb) or human endogenous retrovirus database (HERVd).
- **Phylogenomics and barcoding**. The utilization of large-scale genomic data for resolving phylogenetic relationships between organisms and populations is a powerful application of genomics. Together with DNA barcoding, it has many applications in evolutionary biology, taxonomy, conservation biology, population genomics and epidemiology. Current activities include the development of the AmtDB database.
- **Genomics of selected groups of organisms**, including parasitic protists, neglected crop species, and fungal and microbial communities. These organisms are traditionally of strong interest to the genomics community in the Czech Republic, resulting in substantial data resources and expertise being developed over the last few years (e.g., the GlobalFungi database).

The following goals are proposed to promote further development and to deal with the major challenges of genomic research in the key domains listed above. They address two broad categories of researchers: those involved in the development of bioinformatics tools and the generation of data resources, and those who are mainly users of these resources. Each of these groups faces different challenges, the former mainly due to limitations in terms of computational resources, while the latter is primarily concerned with difficulties in using the tools or accessing genomic data with a limited knowledge of (bio)informatics.

Challenges and goals of ELIXIR CZ

1) Goal 1: Promote the establishment of novel computational tools and databases by providing computational and data storage capacities to the Czech research community involved in their development. Facilitate the integration of these tools into international infrastructures.
2) Goal 2: Support genomic projects by providing hardware resources for the exponentially growing amounts of biological data and expertise for making this data publicly available (interoperability, FAIRification).
3) Goal 3: Provide user-friendly access to the supported computational tools and databases for researchers who are not trained in bioinformatics. Support containerization of the tools to promote reproducible execution. Organize practical training in using the tools.

4. *Bioinformatic data management and access control*

Current status

ELIXIR-CZ deals with two areas that form an integral part of open science:

- Authentication and Authorization Infrastructure (AAI) solving the life cycle of users, their verification, and control of access to services and data,
- Data Management dealing with storing, organizing, and maintaining the data created by experiments.

ELIXIR AAI is a sustainable and user-friendly infrastructure for both users and services across the world. ELIXIR AAI is a part of the ELIXIR Computing platform, which oversees ELIXIR's technical infrastructure. A strong and flexible AAI is needed due to the permanently growing amount of processed and archived scientific data in ELIXIR, including increasingly important sensitive human data and due to a growing amount of related services. ELIXIR CZ, together with ELIXIR FI, is a major contributor to ELIXIR AAI and will continue to participate in its development. ELIXIR CZ partners also participate in major international projects dealing with the implementation of AAI in research infrastructures - GN4, AARC, ELIXIR-EXCELERATE, EOSC-Life. Experiences from the aforementioned projects are utilized in the development of ELIXIR AAI, and thanks to that ELIXIR AAI is currently considered one of the most advanced AAI among the other research infrastructure AAI solutions.

Bioinformatics Data Management (DM) is becoming a critical part of bioinformatics research due to the increasing number of tools that generate such data. Even data that were not originally created or collected for research purposes frequently become 'research objects' at a later stage. As such, DM is perceived as a critical precursor to making data FAIR. It needs to focus on building a knowledge base together with supportive tooling and decision making associated with the planning and evaluation of the FAIRness and preservation of research output. Recently, the COVID-19 situation showed the importance of FAIR DM and the urgency of its development, implementation and embedding in practice. ELIXIR CZ has been playing an active role in these efforts since the beginning in collaboration with other ELIXIR nodes such as ELIXIR NL, ELIXIR SE, ELIXIR LU, and ELIXIR SI. Our main contribution is the development of the Data Stewardship Wizard tool that has become a key ELIXIR solution for DM (Towards Data Stewardship in ELIXIR: Training & Portal implementation study, ELIXIR-CONVERGE). It has also been applied in projects and initiatives of GO FAIR.

Challenges and goals of ELIXIR CZ

1) Goal 1: Development of ELIXIR AAI as an experimental forerunner for authentication and access control for life and beyond sciences.
2) Goal 2: Upgrading current AAI frameworks to provide efficient and flexible access control to sensitive data and to big data.

3) Goal 3: FAIRification of diverse historically independently created distributed data collections and bases and utilizing deep semantics for improving semantic interoperability ("I" in FAIR)
4) Goal 4: Gradual development of a distributed DM toolkit for bioinformatics researchers that comprises the whole DM lifecycle for FAIR data: from data capture, annotation, and sharing; to integration with analysis platforms and making the data publicly available according to international standards.
5) Goal 5: Training of researchers in metadata processing and other DM skills.

ELIXIR CZ pursued field of expertise

ELIXIR CZ will also attempt to further develop activities in the field of precision medicine and use the expertise offered by other ELIXIR Nodes across Europe.

5. Precision medicine

Current status

The focus on precision medicine is a natural extension of the above-mentioned ELIXIR CZ expertise/strategic areas to medicine. As it is primarily genomics-driven, there is a very close connection with the genomics area, as well as with structural and chemical biology via drugs used for treatment. And, of course, there is a specific need for data access and management. The implementation of precision medicine approaches is almost always complex and disease specific. Resources are often fragmented and inherently very specific with respect to the kind of disease. Complex solutions are rare.

Human genomics plays an important role in this part, so it is necessary to consider the 1+MG (1+ Million Genomes) Initiative and B1MG (Beyond 1 Million Genomes) project. ELIXIR CZ is currently not the main driver of these activities in the Czech Republic. An evaluation of its deeper involvement can be done after a detailed mapping of the current situation, because there has not yet been a detailed survey of precision medicine stakeholders, their specific involvement and their needs in the Czech Republic. So, ELIXIR is involved in bioinformatics solutions for precision medicine, substantiated by the strategic partnership of ELIXIR and GA4GH.

The currently identified key domain of ELIXIR CZ is the development of tools and databases for a clinically relevant support system with focus on:

- **Haematological/oncological malignancies**. The processing of personal/biometric data, immunophenotyping, transcriptomic and genomic data (RNA-seq, single-cell RNA-seq, DNA-seq) -

for the analysis of mechanisms of tumours relapse (chemoresistance, immune surveillance), protein mutation involved in cancer development and analysis of the binding process of an FDA/EMA-approved drug to said target proteins, the effect of mutations on protein function, for the detection of pathological mutations on mtDNA molecules, donor choice for hematopoietic stem cell transplantation (HSCT) and real-time interactive data exploration and reporting.

- **Autoimmune and neurodegenerative diseases**. Registry development for setting up tailor-made standardized multidisciplinary healthcare with validated clinical data for *rare disease Amyotrophic lateral sclerosis (ALS)*, the detection of pathological mutations on mtDNA molecules with respect to neurodegenerative diseases, and the study of endogenous retroviruses and their association with diseases of the immune system.

ELIXIR CZ plans to implement a similar strategy to the most prominent ELIXIR members. To directly assist medical professionals as an infrastructure with the necessary complex solutions for using a precision medicine approach in their specific medical field, important steps need to be taken. So, the main challenge is a detailed mapping of the current situation, assessment of stakeholders' needs, implementation of standards and the creation of logistical support in defined medical fields with a special focus on the Czech environment. As for the Federated EGA, ELIXIR CZ wants to focus on supporting those Czech stakeholders that generate data by providing them with sufficient knowledge and facilities, including tools for the subsequent handling and processing of this data for practical use. This action is closely related to the 1+MG activities.

Challenges and goals of ELIXIR CZ

1) Goal 1: Mapping the current situation in the Czech Republic (state of the art, resources/tools used, national/international cooperation, infrastructure and project involvement, ...) including possible ELIXIR CZ activities in 1+MG

Subsequently, depending on the results of Goal 1, the following goals will be pursued in specific medical fields:

1) Goal 2: Setting up a strategy for the efficient and legally correct use and sharing of patient/individual health data
2) Goal 3: Development in ELIXIR CZ identified key domains
    - Improving and adapting existing tools for the specific needs of precision medicine
    - Developing targeted computational and clinical decision support tools and services for diagnosis and therapy
    - Participating in the creation of standardized databases with validated clinical data
3) Goal 4: Establishing and strengthening the interaction between ELIXIR nodes and national infrastructures cooperating internationally, with a specific emphasis on defined Rare diseases
4) Goal 5: Targeting the deployment of R&D tools into clinical practice and possibly industry