

# The future of life science is data-driven

Strategy for the SciLifeLab & Wallenberg  
National Program for Data-Driven Life Science

*Knut and Alice  
Wallenberg  
Foundation*



SciLifeLab

## Table of contents

<b>Executive summary</b>	<b>3</b>
<b>Vision and mission</b>	<b>4</b>
<b>Why is the DDLS program important right now?</b>	<b>4</b>
<b>Strategic objectives</b>	<b>5</b>
<b>Implementations</b>	<b>6</b>
<b>Milestones and deliverables</b>	<b>8</b>
<b>Four strategic research areas</b>	<b>9</b>
Cell and molecular biology	10
Evolution and biodiversity	10
Precision medicine and diagnostics	10
Epidemiology and infection biology	10
<b>DDLS organization and steering structure</b>	<b>12</b>
Working groups set up for DDLS as of April 2021	13
<b>Acknowledgements</b>	<b>13</b>

## **Executive summary**

SciLifeLab & Wallenberg National Program for Data-Driven Life Science (DDLS) is a national 12-year research program funded by the Knut and Alice Wallenberg Foundation (KAW) with SEK 3.1 billion. SciLifeLab (Science for Life Laboratory), as a national infrastructure for life science, coordinates this program in collaboration with ten universities and the Swedish Museum of Natural History. Over the years, the DDLS program will recruit 39 new academic leaders, train over 400 PhD students and postdocs, and work with all stakeholders to profoundly change how life science is practiced today. This document describes the DDLS program's motivation, specific aims, an overall strategy, and the priorities of the four research areas (Cellular and Molecular Biology, Evolution and Biodiversity, Epidemiology and Infection Biology, as well as Precision Medicine and Diagnostics). We believe that the future of life science is data-driven and that an active national collaboration in the DDLS program will promote the government's aim to make Sweden a leading nation in life sciences.

## Vision and mission

**Vision:** The future of life science is data-driven

**Mission:** DDLS is a national research and training program accelerating the data-driven life science paradigm in Sweden, promoting Swedish universities acting at the global frontline with eventual impact on every life scientist and the entire society.

The national DDLS mission is accomplished by joint recruitment of talent, a national program of training and research excellence, launch of a national data platform and set up of an international collaboration hub for engaging academia, health care and industry. Besides top scientific excellence, DDLS also promotes a broad impact for every life scientist as well as for society.

## Why is the DDLS program important right now?

Life science is increasingly data-centric. The European Bioinformatics Institute currently manages about 300 PB of public life science data, and this, along with other data resources are growing rapidly in terms of content, depth and interconnection of data. At the same time, advances in computational capabilities, AI, machine learning and other technologies, provide enormous new opportunities for new objective, unbiased ways to analyze data and to promote biological discovery, insights on life as well as innovation and society benefits. The DDLS program is essential right now because:

- 1. Most life science data are still not FAIR:**

Despite recommendations, most data are still not readily available and FAIR (Findable, Accessible, Interoperable, and Reusable). Often data are not annotated or organized in a standardized, interoperable way to be machine-readable. There is a need for coordinated national efforts facilitating such goals to integrate and make life science data better available.

- 2. Data analysis capabilities need to be further developed and made available:**

Data-driven analysis has developed rapidly as a result of advances in artificial intelligence (AI) and machine learning (ML). As a result, biology can be studied, hypotheses generated, and comprehensive and systematic insights generated in an unbiased manner. However, most scientists are not yet making optimal use of the available data or the latest tools and technologies in their research.

- 3. Life scientists need much more data science expertise and competence:**

Most researchers in life science do not currently have the multi- and cross-disciplinary skills and competences that data-driven science demands. Therefore, training and education is essential to ensure the availability of cutting-edge experts, but also the utilization of data science in the broader life science community.

#### **4. Data science and society needs:**

Industry, health care, decision-makers and the public all need unbiased, data-driven insights of biological processes, human health, and ecosystems. However, there is often a lack of access to data, tools, and technologies as well as expertise. In order for data-driven life science to prosper in the future, a number of policy issues, such as privacy, legislation, and ethical considerations, access to health data, need to be addressed. For example, the COVID-19 pandemic uncovered huge gaps in the flow of data across health care, authorities, academia, the public and decision makers. This contributed to challenges faced by health care and the society at large.








Creation of an entirely new data-driven and hypothesis-generating scientific process would boost discovery opportunities within life science and efforts to examine and understand life processes. A powerful iterative cycle is emerging; from data science to laboratory experiments and back.

We are at the start of a new digitalization era of life sciences which offers exceptional opportunities but also many challenges. We envisage that the DDLS program is essential for Sweden to lead this transformation, not just react to it. Given the importance of data-driven life science, we believe that these actions will substantially promote the government's aim to make Sweden a leading nation in life sciences.

### **Strategic objectives**

The DDLS program will have the following long-term objectives, figure 1:

1. Create a National Framework for Data-Driven Life Science
2. Attract Scientific Excellence
3. Train the next generation of data-driven life scientists
4. Develop national research programs across universities
5. Bridge the gap between the life science and data science communities
6. Create partnerships and impact on society at large
  - industry, health care, and other national and international links
7. Promote policy actions at the national level to provide opportunities for data-driven research

	Create a <b>National Framework</b> for Data-Driven Life Sciences
	Attract <b>Scientific Excellence</b>
	Train the <b>Next Generation</b> of Data-Driven Life Scientists
	Develop <b>National Research Programs</b> Across Universities
	Bridge the Gap Between the <b>Life Science</b> and <b>Data Science</b> Communities
	Create <b>Partnerships</b> and <b>Impact on Society</b> at Large
	Promote <b>Policy Actions</b> at the National Level

**Figure 1:** DDLS program 7 strategic objectives

## Implementations

To realize the strategic objectives of DDLS, we will launch the following specific actions:

**1. Create a National Framework for Data-Driven Life Science:** We will establish a national data platform for data services and life science databases, providing data availability, FAIR data management and advanced data analytics support services. The platform will utilise the top computational resources for the national life science research community in collaboration with all universities. We will coordinate efforts across the country to make FAIR (Findable, Accessible, Interoperable and Reusable) data a norm in academia and to create services and resources to support this. We will facilitate the creation of data resources and services, and organize scientific information, including data, code, methods, and meta-data. The platform will support the needs of the main scientific areas of the DDLS program, and establish domain-specific portals, following the example of the national COVID-19 portal with links to international data resources. We will link up with powerful computational capabilities, advanced data analysis technologies, and AI capabilities and develop new computational methods to facilitate life science research. DDLS will publish a data road map to describe these actions in more detail.

**2. Attract Scientific Excellence.** DDLS will launch the recruitment of 39 junior group leaders (DDLS fellows) in four research areas to the participating universities. After a 5-year fellowship,

the DDLS fellows may have a possibility to be promoted as tenured faculty and continue as key members of the national DDLS program. We expect that the caliber of the DDLS fellows to be recruited will be truly world-leading. In addition, the DDLS fellows should be located in progressive, multi-disciplinary local research environments that form powerful links and synergies with the national DDLS program. Thus, for DDLS to succeed in building a truly globally leading program, it is necessary to attract talented early carrier scientists, but also to link up each university's best research environments and unique capabilities together into a synergistic national DDLS network.

**3. Train the next-generation data-driven life scientists:** Besides the PhD students and postdocs to be recruited to the DDLS fellows' groups, there will be additional positions that will be made available in open calls. Altogether more than 400 PhD students and postdocs are expected to be trained as part of this program. We will launch a national DDLS research school / training programme. The purpose of the training is to educate the future workforce for data-driven life science in Sweden, within academia, industry, health care, and other fields. Courses will be organized together with all universities, SciLifeLab's Data Centre and Bioinformatics Platform, Wallenberg Centre for Molecular Medicine (WCMM), the Wallenberg AI, Autonomous Systems and Software Program (WASP) and many other parties. The training will also aim at making the broad life science community better prepared for the data opportunities and challenges in the next decade.

**4. Develop national research programs across universities:** DDLS will focus on four research areas where all the fellows will be assigned. We will build on this core DDLS community to create a broad national DDLS community, which will participate in research collaborations and training. As DDLS will be anchored at 11 different sites across the country, creation of active national collaborative communities will be key. We expect interactions within each research area, but also cross-disciplinary collaborations across the four research areas.

**5. Bridge the gap between the life science and data science communities:** DDLS program has a unique opportunity to form multi-disciplinary collaborations with the Wallenberg Autonomous Systems and Software (WASP and WASP-HS) Programs and other KAW funded programs. This will enable the life science community to collaborate with the large community of world-leading data science, software, and automation experts. Conversely, this will also provide life science grand challenges to be explored by the WASP community. DDLS will also link up with e.g., the Wallenberg Centre for Quantum Technology and the recently inaugurated Berzelius HPC cluster, Wallenberg Advanced Bioinformatics Infrastructure (WABI) as well as the Wallenberg Centers for Molecular Medicine (WCMM).

**6. Create partnerships and impact on society:** DDLS will boost data-driven life science through partnerships and spread the benefits to the society at large: **A) Industry:** Industry will be a beneficiary of the training programs and junior experts, and it will be important that DDLS is engaged with industry at many levels; including R&D collaborations. Collaboration with the Wallenberg Launch Pad (WALP) program may allow innovative ideas originating from the DDLS research to be developed further into products and services. **B) Health care:** DDLS will promote training of the next-generation of data experts for health care. In the research areas of Precision Medicine and Diagnostics as well as for Epidemiology and Infection biology, we will focus on the links and secure integration of molecular and clinical data in a research setting. DDLS will also work together with the health care regions, biobanks, Genome Medicine Sweden, and the WCMM network on the challenges and policy issues with health care data (see below). **C) Other collaborations nationally and internationally:** DDLS will also engage with other national communities within areas such as biodiversity, environment, agriculture, and forestry. International networking is key to DDLS, and we will build collaborative programs with leading international institutions in data-driven research. DDLS scientists will participate in international (e.g. EU) programs in health care, precision medicine, biodiversity, etc. and will work with its partners to create and promote international standards and practices in data handling.

**7. Promote policy actions at the national level to provide opportunities for data-driven research:** Progress in many areas of life science is highly dependent on regulation and ethical, legal, and social implications (ELSI) and guidelines. These include data security, privacy, ownership, fragmentation and access to health care data that are already being actively debated at the national level. There are also policy questions on biodiversity and sustainability in environmental research. Hence, the DDLS program will work with the community of stakeholders and connect leading experts on ELSI, and related matters, to the program. We will set up a policy action group to address some of the issues that would easily become roadblocks to the transformation to a digital, data-driven future in life science research.

## **Milestones and deliverables**

At the end of the 12 years, we anticipate that the DDLS program has achieved the following outcomes:

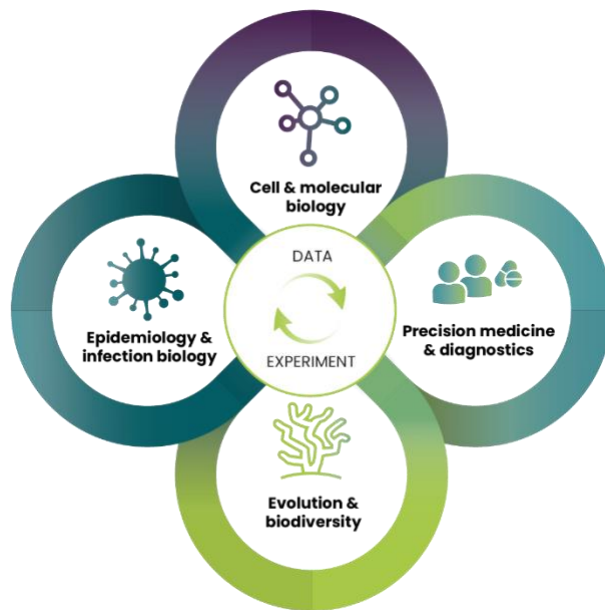
- Accelerated adoption of data-driven life science throughout Sweden and the quality of life science (publication output)
- Sweden and Swedish universities are considered world leading in data-driven life science.
- Outstanding international PI recruitments have taken place.
- A community of 400 PhD students and postdocs has been trained.



- A unique national research program and a networked community established across the 11 partners.
- Major collaborations in place with industry, health care, and other national stakeholders.
- Innovations have taken place and translated via private sector to the society
- Major grants have been acquired by researchers within the DDLS program, such as ERC grants and industry collaboration grants.
- Collaborative interactions with leading international institutions have taken place.
- Established a data platform, broadly enabling and facilitating FAIR data sharing
- Developed high-end computational and ML/AI technologies to transform life science
- Set up advanced computational services for the whole life science sector
- Key policy discussions and actions have helped to take the field forward.
- Taken together, the diverse steps to promote data-driven life science have enabled improved understanding of life and health.

## Four strategic research areas

The program will focus on four broad research areas, where the 39 DDLS fellow positions will be recruited (see figure 2) and where national communities are formed. The aims of these four research areas will be explained below.



**Figure 2:** Main research areas within the DDLS program

### **Cell and molecular biology**

The DDLs program will support research that fundamentally transforms our knowledge about how cells function by peering into their molecular components in time and space, from single molecules to native tissue environments. This research area aims to lead the development or application of novel data-driven methods relying on machine learning, artificial intelligence, or other computational techniques to analyze, integrate and make sense of cellular and molecular data. Our vision for the DDLs program is to support data-driven research that takes advantage of these opportunities, and builds on the state-of-the-art infrastructure and computing capabilities.

### **Evolution and biodiversity**

The DDLs program will support research that takes advantage of the massive data streams offered by techniques such as high-throughput sequencing of genomes and biomes, continuous recording of video and audio in the wild, high-throughput imaging of biological specimens, and large-scale remote monitoring of organisms or habitats. This research area aims to lead the development or application of novel methods relying on machine learning, artificial intelligence, or other computational techniques to analyze these data and to address major scientific questions in evolution and biodiversity. The DDLs and SciLifeLab will also provide state-of-the-art infrastructure, computing facilities and training for data-driven research in evolution and biodiversity.

### **Precision medicine and diagnostics**

The DDLs program will support data-driven research for next generation precision medicine making use of and connecting multiple data layers from genotype to molecular phenotype to clinical data. Molecular precision medicine is about tailoring preventive and therapeutic approaches to the particular characteristics of each person and their disease. Data integration and analysis in DDLs aims to lead to development of molecular patient stratification and discovery of biomarkers for disease risk assessment, prognosis, treatment or prevention. This can include development of data interpretation, visualization and clinical decision support tools. The research is expected to use assets such as high-quality electronic health care data, molecular (e.g. imaging and omics) data, as well as longitudinal patient and population registries, biobanks and digital monitoring data.

### **Epidemiology and infection biology**

Infectious diseases pose significant global threats, including emerging, neglected and chronic infectious diseases, growing antimicrobial resistance, and a lack of antivirals and vaccines. For many host-pathogen systems, multidimensional, genome-scale experimental data can now be processed through computational methods and models to generate testable hypotheses regarding pathogen biology and transmission, as well as to identify antimicrobial or antiviral targets. Population-scale genetic, clinical, or public health data from pathogen surveillance efforts and

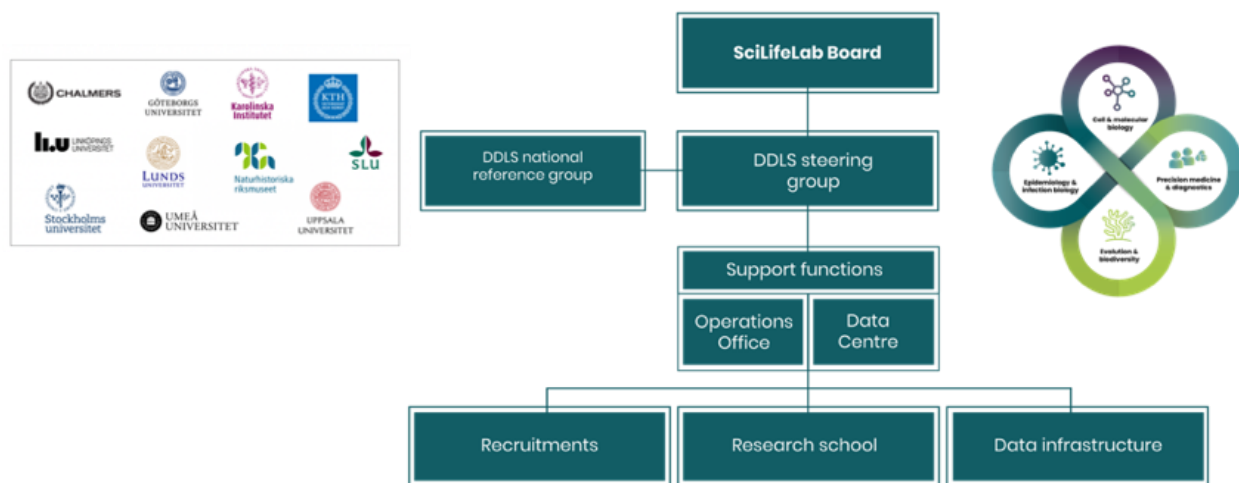
biobanks, on the other hand, offer opportunities for data-driven prediction of the emergence, spread, and evolution of infectious agents, improved diagnostics, and to understand pathogenicity. DDLS work in this research area will use big experimental, clinical, or pathogen surveillance data in innovative ways to transform our understanding of human, animal or plant pathogens, their interactions with hosts and the environment, and how they are transmitted through populations.

## DDLS organization and steering structure

The DDLS program is funded by the Knut and Alice Wallenberg Foundation (KAW) with a total of SEK 3.1 billion over twelve years and the use of the funding is stipulated in the KAW donation letter. Funding will be provided in 3-year allocations, based on a progress report and a detailed strategic plan. The SciLifeLab Board is the decision-making body for the DDLS program, while the Program Director manages the operations together with the DDLS steering group members. A national reference group with representatives from all 11 parties will support and advise on strategic issues and to ensure close links to the operations and leadership at the collaborating organizations, figure 3.

The program's main operations currently include Recruitments, Research school and Data infrastructure. The steering group coordinates these operations to create synergies throughout the program. As the program develops, working groups for other activities will be launched as needed.

As a SciLifeLab coordinated program, DDLS gains substantial synergies in the interactions between infrastructure, research and data at the national level. For example, the SciLifeLab Data Centre is responsible for coordinating data infrastructure and support. In addition, we can make use of the SciLifeLab organization in the coordination and administration of the DDLS program, such as communication, external relations, training, meetings, events, financing and reporting.



**Figure 3:** The DDLS governance, operations, and support functions

## **Working groups set up for DDLS as of April 2021**

### **1. Data strategy working group:**

This group works on establishing a national data-sharing platform, providing access to services including compute and storage e-infrastructure, computational tools, bioinformatic web services, databases, and topic-specific, along with web-based data portals for the four research areas of the DDLS program. The platform will provide a common structure for data-centric services and projects, community-created content, and a single point of contact user support portal. The DDLS program will work together with major Swedish e-infrastructure providers to increase the capability to analyze and share data.

### **2. Recruitments working group:**

This group works on defining the principles of the national recruitment for the DDLS program, and will organize the coordination of recruitments at the national level as well as the adherence to the aims of the program.

### **3. WASP collaborations - working group:**

This group will plan and coordinate the interactions with the WASP community and together with a joint WASP – DDLS working group organize and launch joint calls and networking activities.

### **4. Program coordination, networking and research school – working group:**

This group will plan and coordinate program overarching activities, such as annual conferences and other networking activities, plan for training and research school development and ELSI support, as well as act as support to the other three working groups mentioned above.

## **Acknowledgements**

This first version of the DDLS strategy has been developed by the DDLS steering group with input from the KAW, DDLS national reference group (representatives from the 11 participating organizations Chalmers, GU, KI, KTH, LiU, LU, NRM, SLU, SU, UmU and UU), SciLifeLab Board, Management group, Operations office and the Data Centre.