NEDERLANDSE FEDERATIE VAN
UNIVERSITAIR MEDISCHE CENTRA

Data4lifesciences (English)
A shared data infrastructure for biomedical research

Handling scientific data responsibly requires a great deal from investigators. The data volume has increased enormously and the requirements placed on handling them are becoming increasingly strict. This applies to requirements regarding privacy, but also to those regarding the use and reuse of data. Research data must be collected, stored, processed, analysed and archived and, what is more, shared with others. Fortunately, some generic solutions for these challenges are available, i.e. research infrastructures, which have been set up to facilitate the handling of scientific data. These research infrastructures comprise not only technical facilities (hard and software), but also focus on quality assurance processes and expertise development of investigators and administrators.

Background
The high quality infrastructure needed to meet data-related requirements exceeds the capabilities of individual investigators and, increasingly, also those of individual university medical centres (UMCs). This is why the UMCs have opted for a joint approach, under the coordination of the Netherlands Federation of University Medical Centres (NFU), in conjunction with national programmes such as TraIT, BBMRI-NL, Parelsnoer, DTL, AcZie, Mondriaan and SURF: Data4lifesciences. This programme will set up an innovative research data infrastructure at, for, by and between the UMCs and their partners.
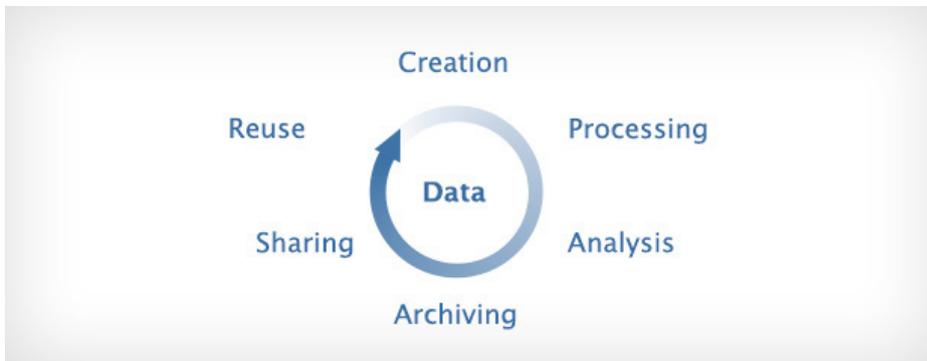
Biomedical research concentrates increasingly on the individual patient. By means of 'personalised medicine', attempts are being made to tailor the methods to the individual patient in order to ensure that he or she receives the best possible treatment. The technological revolution in genetics and imaging (Magnetic Resonance Imaging [MRI], computed tomography [CT], etc.), among other areas, have made this possible but have, at the same time, resulted in an explosive growth in research data, data which are also extremely complex. Moreover, for personalised medicine research, large cohorts are required (cohorts are groups of patients with similar characteristics, such as all men born between 1940 and 1950). Such cohorts can only be compiled through (international) collaboration.

Investigators thus have to deal with much more, and more complex, data, but also with stricter requirements in terms of quality, management and sharing of the data. As already outlined above, this is the reason for setting up specific research data infrastructures, inititially on a small scale at the individual UMCs and within specific disciplines. The purpose of the Data4lifesciences programme is to connect these local initiatives to national and international infrastructures. A joint approach, administrative coordination and coordination at the national level are imperative if this is to be achieved.

Ambition

The Data4lifesciences programme will have achieved its objective when

All doctors and investigators working at the UMCs use the research data infrastructure to retrieve clinical and experimental data on all UMC-related patients and make them available to others. The infrastructure is also used to find and request biological material. This infrastructure does not stop at the walls of the UMCs but is also used and made available to collaboration partners in the Netherlands and abroad.

The infrastructure forms a national virtual collaborative environment in which data is registered, processed, analysed, archived and shared. The data is FAIR (Findable, Accessible, Interoperable and Reusable) and is made available in a scalable, distributed environment, in which the computing capacity needed to process the data comes from national and UMC computing facilities. The infrastructure is accessible to all investigators and doctors, independently of institution or site. Patient privacy is safeguarded in the process.

Investigators and doctors with questions about handling data have an extensive data expertise network at their disposal to answer them. These experts in the UMCs are the first point of contact for these questions, solutions will be sought both inside and outside the UMCs. Training is essential to this end.

So why is this expedient now?
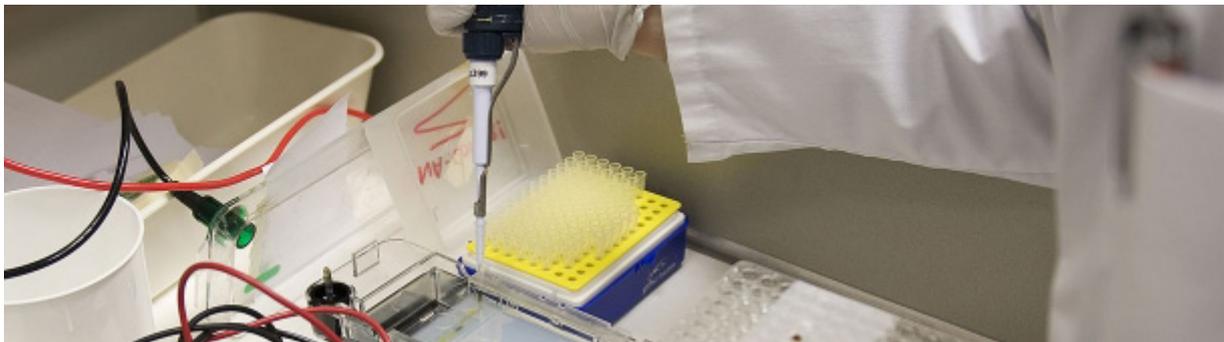
- Sub-projects such as TraIT, BBMRI-NL, Parelsnoer and DTL have achieved a great deal in the field of data infrastructure but coordinated action is needed to sustain and expand these results.

- Urgent reasons for adapting the current infrastructure are the upcoming EU Privacy Regulation, the implementation of new Electronic Health Records (EHR) and more stringent requirements placed on the quality of data management and reusability of data by important research sponsors (for example, the Netherlands Organisation for Scientific Research [NWO], the Netherlands Organisation for Health Research and Development [ZonMW] and the Dutch Cancer Society [KWF]). Data must be FAIR: findable, accessible, interoperable and reusable.

- Joint action by the UMCs will reinforce the competitive position in the Netherlands and in Europe. The current data infrastructure is a good starting point, but this position can only be maintained if the UMCs act jointly to renew the infrastructure regularly. This is a prerequisite for attracting substantial funding in the future, to be able to exploit new European programmes, and to stay ahead of the competition. European infrastructures are, furthermore, playing an increasingly important role in the acquisition of new resources.

- The UMCs are responsible for the quality of research data and the accuracy with which they are collected, stored, processed and archived, as well as for meeting the relevant regulations regarding the protection of privacy and patient safety. An excellent national infrastructure is needed to reduce the likelihood of reputational damage which could arise if individual investigators do not comply with the applicable legislation. This applies to an even greater extent to prevention of abuse, scientific fraud and security leaks.

Shared infrastructure

A data infrastructure is a generic solution for a specific aspect of handling scientific data, intended to lighten the load of investigators and to enable them to make optimal use of the diverse capabilities of IT technology. An infrastructure can take many forms: it may comprise an online catalogue of samples in a biobank, a standard method with which data in an EHR is made available, privacy regulations, the way in which IT is organised at UMCs, a manual which elaborates how investigators should handle data ('data stewardship'), a generic way in which data can be exchanged or an expert who supports an investigator with data issues. The envisaged high quality data infrastructure Data4lifesciences is to deliver will thus consist not only of technical facilities (hard and software), but also systems and processes for assuring quality and the requisite expertise of investigators and administrators. Investigators will, furthermore, be supported by experts in what will be known as shared service centers.

Data4lifesciences will guarantee administrative coordination so that local facilities and expertise networks will be in line with national and international infrastructures and vice versa. It will soon be easy for the investigator to find general information on data-related aspects of research and, for example, the answers to the questions below.

1. I am carrying out an EU study. May I share participants' DNA sequences 'in the cloud'?
2. I am combining care and research data in the EHR. How do I retrieve it?
3. I have to write a data management plan for a grant application. What has to be in it?
4. I need a lot of computing capacity for a short while. Where can I get it?
5. Are the blood samples I need already available somewhere in a biobank so that I do not have to recruit patients again?
6. My research involves a large number of hospitals. How do I collect the data I need and what do I have to arrange to be able to do so?

Data4lifesciences programme lines

The shared research data infrastructure project has been broken down into sub-projects; those listed below will be in full swing by the end of 2015:

1. *Harmonisation of guidelines for data administration ('data stewardship').*
   This programme line entails the development of a guideline for a data stewardship policy in the form of a website with pointers to local and national expertise. A shared policy which is supported by all UMCs is an important condition for the sharing and reuse of research data and for the shared infrastructure ultimately to be achieved. In fact, this forms the basis for the implementation of the infrastructure at the UMCs. The guidelines relate to patient safety, protection of privacy, involvement of patients, ethics, the reliability and provenance of data, monitoring quality, legislation and regulations and so on.
   The Erasmus MC is coordinating this programme line, in cooperation with various specialists from UMCs, universities, knowledge institutes, companies and funders.

2. *Harmonisation of processes and IT architecture*
   This programme line entails the identification and listing of existing IT architectures and guidelines for the collection, processing and provision of data at the UMCs and to collaboration partners. The information will be shared with all those involved, after which a reference architecture will be chosen for the areas which are crucial for high quality research data. The coordination is in the hands of the existing architecture working group of the Center for Translational Molecular Medicine's (CTMM) TraIT and the NFU Special Interest Group PRIMA (in which the IT architects from the UMCs participate).

3. *Access to data and samples (catalogue)*
   This programme line entails making the various collections of biomedical samples (tissue, blood, urine, etc.) and data sets accessible to investigators via a shared catalogue. The registries of, for example, BBMRI-NL (including UMC biobanks), LifeLines (a large-scale genetic study into the mechanisms behind healthy aging), Parelsnoer, TraIT and PALGA (a nationwide network and registry of histopathology and cytopathology in the Netherlands) will be connected so that samples and data can be sought at the national level. The procedures for providing data and samples will be harmonised with one another and there will be proper assurance of privacy protection and informed consent. BBMRI-NL is coordinating the programme line on the basis of a national catalogue working group with representatives from the aforementioned initiatives, supplemented by contributions from European projects such as BioSHaRE and BioMedBridges.

4. *Sharing and analysis of biomedical data*
   This programme line focuses on the challenges faced by large national studies in collecting and integrating data from different hospitals and the subsequent joint analysis of these data. To this end, a shared research data platform made up from reallife use cases will be made available. This will take place on the basis of existing best practices from national programmes such as Parelsnoer, CTMM-TraIT, BBMRI-NL and Mondriaan, as well as from the UMCs.

5. *The use of digitally stored patient data for research*
   As yet, there is little direct reuse of health care data for biomedical research; manually re-entering health care data into research systems is still the standard approach. The purpose of this programme line is to change this situation. To this end, it has close connections with the NFU programme 'Registration at the source' which regulates the clear, one-off registration of patient data. The first results will comprise a series of pilot projects together with Registration at the source, CTMM-TraIT and Parelsnoer.

6. *Good research in practice*
   Privacy is a controversial issue in scientific research. The patient must be confident that medical data will not be made public, but this is becoming increasingly difficult as a result of the technological developments in 'big data'. This programme line implements procedures and technologies for privacy and security aspects in accordance with the guidelines specified in

programme line 1 (data stewardship). This includes for instance the pseudonymisation of personal data via Trusted Third Parties (TTP), the encryption of citizen service numbers, and implementation of security standards.

7. *Facilities for high quality data processing*
This programme line concentrates on the computing capacity needed to process research data, primarily via High Performance Computing Cloud systems. Issues which play a role here include being able to provide extra peak capacity easily, the sharing of best practices and joint action towards suppliers. SURF is responsible for the coordination. The programme is nationally and internationally embedded via BBMRI-NL/EU, EYR, CTMM/TraIT, and ELIXIR.

8. *Access to experts, training and support*
Investigators with questions about data stewardship (see also the data stewardship programme line) can address them to local experts at the UMCs. A national expertise network will be set up for this purpose, in which the current UMC data desks will form a vital link. DTL has an excellent network for promoting knowledge exchange and integration between the UMCs and other stakeholders and will be coordinating this programme line.

Organisation and information

The UMCs and many niche collaboration partners are participating in the Data4lifesciences programme. The programme committee, many working groups and other consortiums are working on the realisation of an integrated research data infrastructure under the direction of chairmen Frank Miedema (UMC Utrecht) and Folkert Kuipers (UMCG) and programme manager Jan Willem Boiten (CTMM).

The operational committee comprises:

Name	Organisation	Partly on behalf of programme
Jan-Willem Boiten (programme manager)	Lygature	TraIT
Jeroen Beliën	VUmc	TraIT
Hans van den Berg	AMC	
Rob Cornelisse	LUMC	
André Dekker	Maastro Clinica	TraIT
Petra Drankier	NFU	
Erik Flikkenschild	LUMC	PSI
Rob Hooft	DTL	
Arnoud van der Maas	Radboudumc	
Irene Nooren	SURFsara	
Bert van Ooijen	Erasmus MC	
Petra van Overveld	DTL	

Name	Organisation	Partly on behalf of programme
Harry Pijl	UMCU	
Ronald van Schijndel	VUmc	
Igor Schoonbrood	MUMC+	
Morris Swertz	UMCG	BBMRI-NL
Jan Jurjen Uitterdijk	UMCG	

The programme committee comprises:

Naam	Organisatie	Partly on behalf of programme
Frank Miedema (chairman)	UMC Utrecht	
Jaap Verweij (chairman)	Erasmus MC	
Ameen Abu Hanna	AMC	
Edvard Beem	ZonMW	
Jan-Willem Boiten (programme manager)	Lygature	TraIT
Arjen Brussaard	VU Medical Centre	
Petra Drankier	NFU	
Alain van Gool	Radboudumc	
Jan Hazelzet	Erasmus MC	Registration at the source
Olaf Klungel	Utrecht University	Mondriaan
Ruben Kok	DTL	
Gabriël Krestin	Erasmus MC	Population Imaging
Karel van Lambalgen	LUMC	Vz AcZIE
Gerrit Meijer	NKI (the Netherlands Cancer Institute)	TraIT, EATRIS
(European Advanced Translational Research InfraStructure)	Maastricht UMC (MUMC+)	
Frits van Merode	MUMC+	
Barend Mons	LUMC	ELIXIR
Anwar Osseyran	SURFsara	
Ronald Stolk	UMCG	Mondriaan, PSI (the Parelsnoer Institute)
Cisca Wijmenga	UMCG	BBMRI-NL

Informatie

For more information about Data4lifesciences, see www.nfu.nl

or contact Carolien Bouma, senior policy officer at the NFU: bouma@nfu.nl or call +31 (0) 30 2739 632.