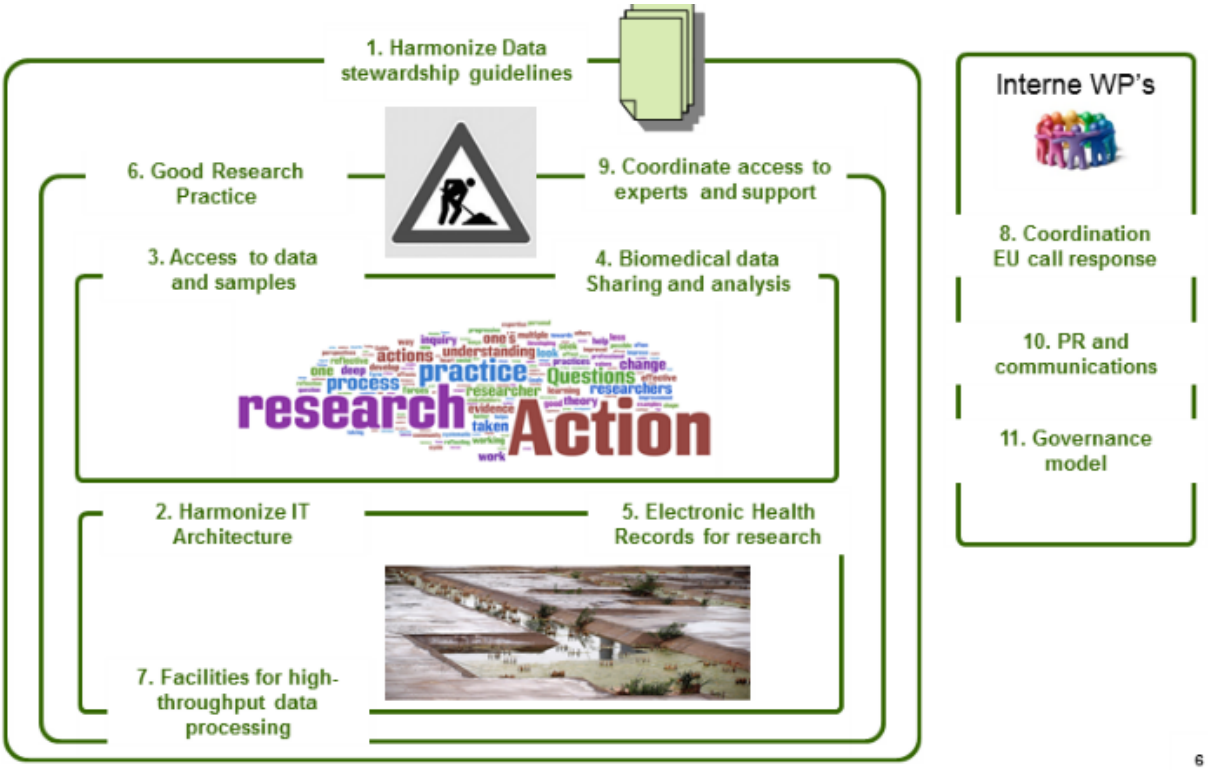
Target audience

This is a high-level project plan and has as target audience the program board of the NFU program Data4lifesciences (D4LS) and the operational board which consists of the D4LS WP leaders.

Contents

1. Introduction

The NFU Data4lifesciences (D4LS) program is initiated by the joined UMCs in The Netherlands aiming to create a shared data infrastructure for biomedical research. Primary objectives are to improve data findability, accessibility, integration, and reusability (FAIR[1]), and promote sharing of expensive IT facilities in such a manner that the entire data infrastructure will appear to the (international) researcher as a coherent set of high-end data services from one virtual 'UMC.nl'. For example, researchers should be able to utilize datasets gathered from any UMC/hospital, use the high-performance compute clusters of all UMCs as well as supporting institutes like SURF and CIT for computer-intensive analyses, and benefit from data handling processes that have been standardized to speed up integration and to improve data quality.
The D4LS program is organized in work packages as outlined in the figure below.



1.1. Original aims stated (September 2014)

In the original work plan for Data4lifesciences (dated September 2014) work package #5 was still called "Using electronic health records for research", and the following goals were specified:

1. Make more and more effective research possible, by using data that are collected in the care process effectively.
2. Reduce patient burden by using the EHRs for the gathering of research data where possible.
3. Reduce the administrative burden of recording patient data for doctors and researchers.
4. Develop 'blue prints' for the implementation of EHRs on behalf of research.

1.2. Barriers to using clinical data for research

A common understanding of the problems introduced by sharing clinical data for research is crucial to a successful resolution to the problems. In this paragraph the, sometimes overlapping, barriers to sharing clinical data are described.

Administrative, political and ethical barriers to sharing data are more problematic than technical barriers

In a 2011 review[1], the main barriers to sharing (and thus re-using) data from clinical care were summarized as "the problem is not really technical […]. Rather, the problems are ethical, political, and administrative." The latter three barriers with examples:

1. Administrative (not enough resources to share)
    - No time to capture data in a structured format in the EHR in a busy clinic

[1] *https://www.force11.org/group/fairgroup/fairprinciples*

- No resources to quality-control the data
- No resources to fill in an external system (*e.g.* electronic Case Report Form (eCRF))

2. Political (not wanting to share)
   - I want to keep control of my own data
   - I want to publish first, have first crack
   - I want to have control in the way my data is analyzed to make sure the conclusions are correct and do not harm me
   - I have promised/sold the data to someone else
3. Ethical (not allowed to share)
   - I can't share data within my legal framework
   - Sharing would breach privacy, bring harm to the patient

Information models control and dynamics are different between care and research
In clinical care an information model[2] is often made only once, at the introduction of an EHR or by a vendor. Changes after that initial model are generally kept small because they have a major impact on the system. In research every project defines its own information model at the start. Often the information model is 'forced' upon the local investigator by an external investigator in multi-center studies, and as such a local investigator has little control over the information model. Information models for care are mostly static, information models for research are dynamic.

Availability and uptake of information model standards in care and research are limited
In both domains there are efforts to define information model standards (*e.g.* CDASH for the research domain and Detailed Clinical Models for the care domain) but these are not interoperable, are not widely taken up, and generally cover a (very) small part of the information needed.

Consent, legal and ethical use of information are different for care and research
The use of information for the benefit of an individual patient in clinical care is permitted without an explicit consent. For research use usually informed consent (incl. what is consented use of the data) is necessary. In the Netherlands, also an opt-out exception is possible for de-identified data if getting informed consent is unfeasible. When using data through this exception, one must note the purpose for using the data and make sure one note which patient was used for what research question. The use of data for research and the difference between using data with informed consent or with an opt-out is usually not well implemented in clinical systems.

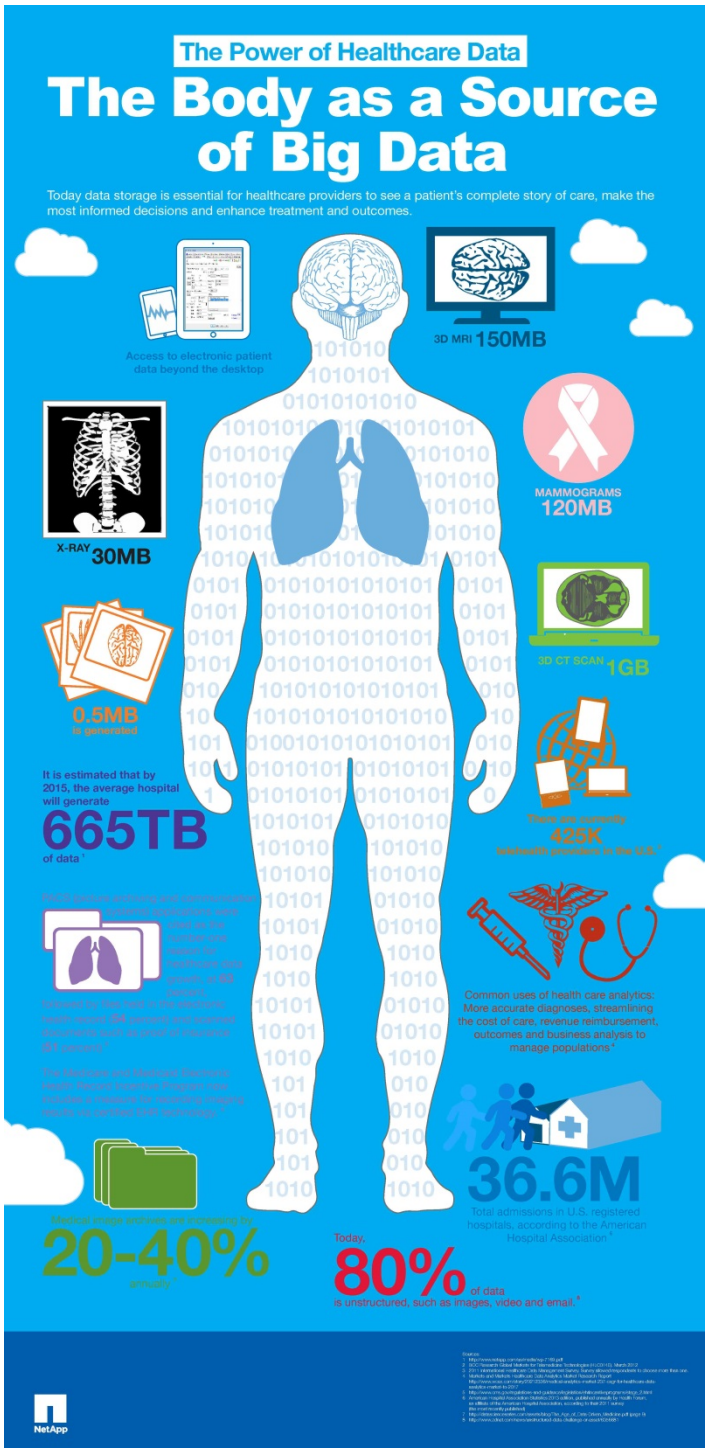


Figure 1: Data overview. From http://www.designinfographics.com/health-infographics/what-is-the-power-of-the-big-data-in-healthcare

Level of interoperability is low both within care and within research and between them it is even lower

[2] *An information model is defined here as a description of the information you want to collect.*

The EU semantic health report defined levels of interoperability which are summarized here as:

- No interoperability: The data is not there or cannot be read
- Syntactic interoperability: The data is there and can be read but is not understood (e.g. foreign language) or using concept codes which are unknown to the receiver
- Semantic interoperability: The data is there, can be read and can be understood (even by a computer)

Given the fact that most care organizations still rely on faxes and emails, and that research organizations rely on paper or electronic CRFs, the current status in NL is syntactic interoperability. Furthermore the efforts towards semantic interoperability in care and research are not well coordinated (*e.g.* Detailed Clinical Models prefer SNOMED CT in NL, while CDASH prefers the NCI Thesaurus terminology).

The majority of data in clinical care is unstructured

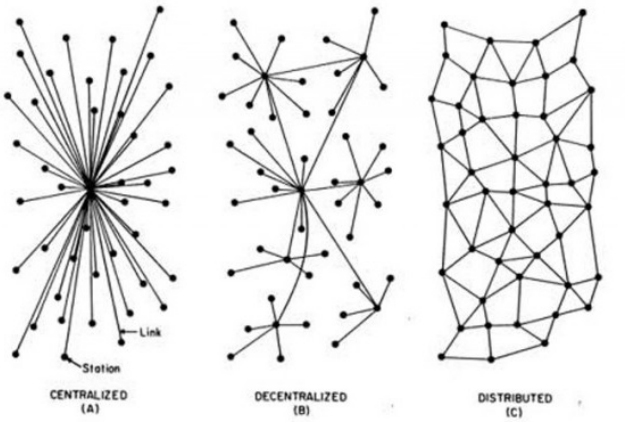
80% of data in hospitals is unstructured and takes the form of images, scanned documents, free text fields with prose and/or shorthand etc. Unstructured data cannot be made interoperable easily and systematically, and cannot be de-identified in a fully reliable manner.

De-identification, pseudonymization and anonymization of data is a battle that is lost

Much of the secondary use of clinical data for research is based on the assumption that data can generally be fully anonymized. This is not true nowadays and will become even more problematic going forward. The amount and content of the information that needs to be released for an individual to answer a research question is often enough to re-identify a patient with modest means. *E.g.* knowing things like gender, age, length, weight, date and location of care encounters, diagnosis, treatments and outcomes very quickly reduces the number of likely candidates. Removing the obvious identifiers (de-identified data) and replacing it (or not) with an unrelated identifier (pseudonymized data) does not change that. Sharing images and genomic data will make this situation worse and will make it easier to re-identify a patient.

The centralization information model of research does not scale to broad clinical data re-use

Often research requires data to be centralized *e.g.* in a clinical trial or audit management system through eCRFs or paper forms. Also images often need to be uploaded to a central FTP site or sent on DVDs to the study coordinator. This centralization ensures that data is available for the researcher to analyze and that the data adhere to a research specific information and data model. However this centralization is unsustainable when sharing clinical data. First, the size of the data to be shared is already large and will continue to grow at a fast rate. Second, centralization requires each hospital to adhere to a study-specific data, data exchange, and information model which are not standardized with the other studies the hospital is involved in; every study defines their own standards. Finally, centralization of data means a loss of control, the hospital has to investigate and trust the external site in terms of technology (*e.g.* security) and (data) governance and has to ensure applicable local or national regulations on the use of data are met by the central location.



2. Deliverables plan

2.1. Overall aim – True North

Given the identified problems described above, the overall aim (True North) of this project is:

What all clinical data of all patients should be made available
Who by every UMC within the Netherlands
Why for every valid research question
When now and forevermore
Where in a scalable, distributed environment
How as semantic interoperable data with a flexible information model, with full protection of privacy of patients, with full control of the UMC and without impacting clinical care

2.2. Specific aims and deliverables – mBSC

The specific aims and associated deliverables are given in a modified Balance Score Card model and the status on 3/2017 is given.

Objective	Norm	Indicator
All clinical data of all patients should be made available	% of data that is FAIR (including extracting FAIR data from narrative text)	80% by 2025
By every hospital of the Netherlands	# hospitals which make clinical data available to Data4lifesciences	All 8 UMCs by 2018 All hospitals by 2022
For every valid research question	Validity of research question defined	Defined by 2017
	Process to guarantee only valid research questions are asked	Implemented by 2020
Now and forevermore	% of data that is FAIR (where the A prescribing long term storage)	80% by 2025
In a scalable, distributed environment	Availability of scalable technology	Technology analysis done in 2017
	Availability of a distributed environment	Technology analysis done in 2017
As semantic interoperable data	% of data which use formal ontologies	80% by 2025
With a flexible information model	Number of extensions to the ZIBs being used in research.	Some research projects are using a combination of ZIBs and custom information models in 2018
With full protection of privacy of patients	Use of privacy-by-design	Technologies offering privacy-by-design evaluated in 2018
	Use of consent	A patient can give consent per data element in 2020
With full control of the UMC	Level of control for UMC	Access to UMC data is governed by UMC administrators in 2020
Without impacting clinical care	Ability to reuse data from clinical routine information systems without routine care being affected	In every UMC data reuse is possible from routine EHR and other hospital systems in 2020

3. Time plan

3.1. Past (2015-2016)

- Getting True North on various roadmaps
  - KNAW Health-RI (January 2016)
  - European Open Science Cloud (April 2016)
- Secure/re-use/leverage funding
  - DTL FAIR-dICT - Topsector Health Holland TKI/LSH Project (approved Q1 2016)
  - Head & Neck Cancer Audit - NFU Registratie aan de Bron Versnellingsproject (approved Q1 2016)
  - PRANA-DATA – COMMITT Zwaluwproject (approved Q1 2016)
  - Field Lab - Smart Industries (submit in 2016)
  - TraIT - CTMM (dates from before D4LS)
  - duCAT- STW (dates from before D4LS)
- Demonstrator project
  - D4LS WP5 pilot: Linked-data based technical implementation for reuse of clinical data for a Head & Neck Cancer trial in synergy with Detailed Clinical Models (Zorginformatiebouwstenen) used in the “Registratie aan de bron” project on a Head&Neck Cancer Audit.
- Semantic-web based technical implementation which re-uses clinical data for the DICA Dutch Lung Radiotherapy Audit (DLRA, 2015-2016).
- Visit to all UMCs (2015-2016) to discuss WP5
- Project plan
- Project environment and –organization operational

3.2. Short term (2017)

- Finishing open projects 2016 (Castor import + export from second EHR)
  - Castor import: setting up a FHIR message
  - Export 2nd EHR 2e EPD: Try MUMC+ solution in another UMC (LUMC)
  - Investigate OpenClinica as an alternative to Castor
- Link up with Medical Intelligence and other WPs
  - Joint meetings and workshops, in particular with WP2 to establish data architecture between clinical care and research
  - Joint project environment
  - Project manager regularly joins other WP meeting to make sure WP5 remains integrated
- Secure/re-use/leverage funding
  - NWO Roadmap (Elixir cluster)
  - KWF application (protons)
  - Health-RI
  - Citrienfonds – Registratie aan de Bron
- Demonstrator projects
  - D4LS WP5 pilots
    - H&N cancer using DICA, Castor, HL7v3, SAP, HIX
    - Second cancer using Linked Data and HL7 FHIR
    - Applying FAIR principles to H&N cancer
  - Use Case
    - Data exchange for proton therapy (MUMC+, UMCG, LUMC, Erasmus MC) using Health-RI & Personal Health Train approach.
  - D4LS-MRDM pilot
    - Exchange radiotherapy data with MRDM using HL7 FHIR
  - Parelsnoer:
    - Investigate option for joint project to build future Parelsnoer data infrastructure based on D4LS principles.

3.3. Medium term (2018-2020)

- Demonstrate scalability of pilot implementation by linking up with other projects
- Getting more resources through grant applications
- Design data access process so that requests may be made

3.4. Long term (2020+)

In the long term, to get to True North, a large-scale facility for medical research on clinical data in the Netherlands is needed. This facility was proposed as the “Health-RI” project and accepted Q1 2016 by KNAW in their “call for dreams”[3].

[3] *http://www.dtls.nl/about/documents/*

4. Budget plan

4.1. Resources/deliverables expected from other WPs

The following resources, deliverables can be drawn from other work packages within the program

Personnel	Contribution
General management and admin support (central)	Reviewing project plans, monitoring progress
Data stewardship guidelines (WP1)	Inform if WP5 solutions meet guidelines
Architecture (WP2)	Inform if WP5 solutions fit the architecture. Help UMCs accept the WP5 solutions
Data sharing and analysis (WP4)	Inform if WP5 solutions can be used in their solution WP5 is “the” data source for the DRE of WP4
Materials & Services	
None foreseen	

5. Organization plan

5.1. Project team

Name	Home Institute/Project	Project Role
Andre Dekker	MAASTRO Clinic Maastricht MUMC+ Maastricht TraIT & Health-RI PRANA-DATA duCAT DLRA	Lead
N.n.	Project manager	Project management
Jan Hazelzet	Erasmus MC Rotterdam	Mentor
Rudi Scholte	AMC Amsterdam	UMC rep
Harry Pijl Robert Veen	UMC Utrecht	UMC rep
Igor Schoonbrood	MUMC+ Maastricht	UMC rep
Bert van Ooijen	Erasmus MC Rotterdam	UMC rep
Louise Veltrop; Karin van der Pal, Rob Cornelisse	LUMC Leiden	UMC rep
Ernst de Bel	Radboudumc Nijmegen	UMC rep
Rene Oostergo Erik van der Velde	UMCG Groningen	UMC rep
Gerard van der Voorn	VUMC Amsterdam	UMC rep

5.2. Stakeholders

Contact	Stake
Fred Smeele	NICTIZ / Registratie aan de bron / Generieke overdracht gegevens / iZiekenhuis / H&N audit
Joyce Simons	Registratie aan de bron
Joep Veraart	H&N audit
Igor Schoonbrood	SIG-PRIMA
Linda Mook	Generieke Overdrachtgegevens / Details Clinical Models
Jan-Willem Boiten	TraIT/ Data4lifesciences / Lygature
Ernst de Bel	Radboudumc / EPIC / H&N audit
Jan-Jaap Hendriks	LUMC / Chipsoft / H&N audit
Mariëlle Ouwens	IQ HealthCare
Pim Koeman	DICA / Dutch Lung Radiotherapy Audit
Wessel Kraaij	PRANA-DATA / TNO
Ruben Kok	FAIR-dICT / DTL

6. Communication plan

The communication plan follows the WP0 lead in its communication. Below are some specific communication planned

6.1. NFU Stakeholders

Organization	Contact person	Min. frequency	Type
D4LS	Mentor	Quarterly	Personal[4]
Organization	**Contact person**	**Min. frequency**	**Type**
NFU-Overall	Project manager	Monthly	Personal
D4LS	WP leaders	Bi-monthly	Operational team
Medical Intelligence	Project manager	Quarterly	Personal

6.2. Other stakeholders

Organization	Contact person	Min. frequency	Type
PRANA-DATA	Wessel Kraaij	Quarterly	Personal
NFU Registratie a/d Bron	Joep Veraart	Quarterly	Personal

[4] *Personal = E.g. a face to face meeting or video- or teleconference with the contact person*

7. Risk plan

Below a risk matrix is given which identifies the major risks as can be seen at this moment in the project. The risk plan is to formally review the risk list every 6 months so that new risks can be added, hazard can be re-estimated and actions be taken. The higher the value, the higher the risk

Risk description	Probability (1-3)	Impact (1-3)	Hazard (P*I)	Action
1. Not enough funding	2	3	6	Contingency – Redraft project plan to match funding Mitigation – Actively look for funding
2. In-kind support UMC insufficient	3	1	3	Contingency – Only work with UMCs which have support Mitigation - Name and shame, work with WP0
3. Patchwork of projects leading to fragmented development	2	2	4	Mitigation – Centralize development work at DTL
4. No involvement of UMC	3	1		Mitigation – Escalate to WP0

8. References

1. Sullivan, R. *et al.* Delivering affordable cancer care in high-income countries. *Lancet Oncol.* **12,** 933–980 (2011).

9. Endnotes / Glossary

- CDASH: Clinical Data Acquisition Standards Harmonization
- CIT: Centrum voor Informatie Technologie, Groningen
- CRF: Case Report Form
- DICA: Dutch Institute for Clinical Auditing
- DLRA: Dutch Lung Radiotherapy Audit
- DRE: Digital Research Environment
- duCAT: Dutch Computer Assisted Theragnostics. An STW project.
- eCRF: electronic CRF: Case Report Form
- FAIR: Findable, Accessible, Interoperable, Reusable
- FAIR-dICT: FAIR data in coordinated transition: a guiding demonstration role for The Netherlands in data driven science, innovation and health. A TKI/LSH project.
- FTP: File Transfer Protocol
- GA4GH: Global Alliance for Genomics and Health
- Health-RI: Health Research Infrastructure for Personalised Medicine & Health Research
- NFU: Netherlands Federation of UMCs
- PRANA-DATA: Privacy Respecting ANAlysis of Distributed patient health. A COMMITT zwaluwproject.
- TraIT: Translational Research IT
- UMC: University Medical Center
- ZIB: Zorginformatiebouwstenen