Contents

1. Introduction

The NFU Data4lifesciences program's objectives are to make research data FAIR[1]: Findable, Accessible, Interoperable, and Reusable, and co-create and share research data and IT infrastructure across (at least) the Dutch UMCs in such a manner that the entire data infrastructure will appear to the (inter)national 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 from all UMCs as well as supporting institutes like SURFsara 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.
Another objective is to create an environment that is compliant with Good Research Practices[2] for which research data needs to be trackable and traceable to its origins.

The Data4lifesciences program is organized in work packages according to the figure below.

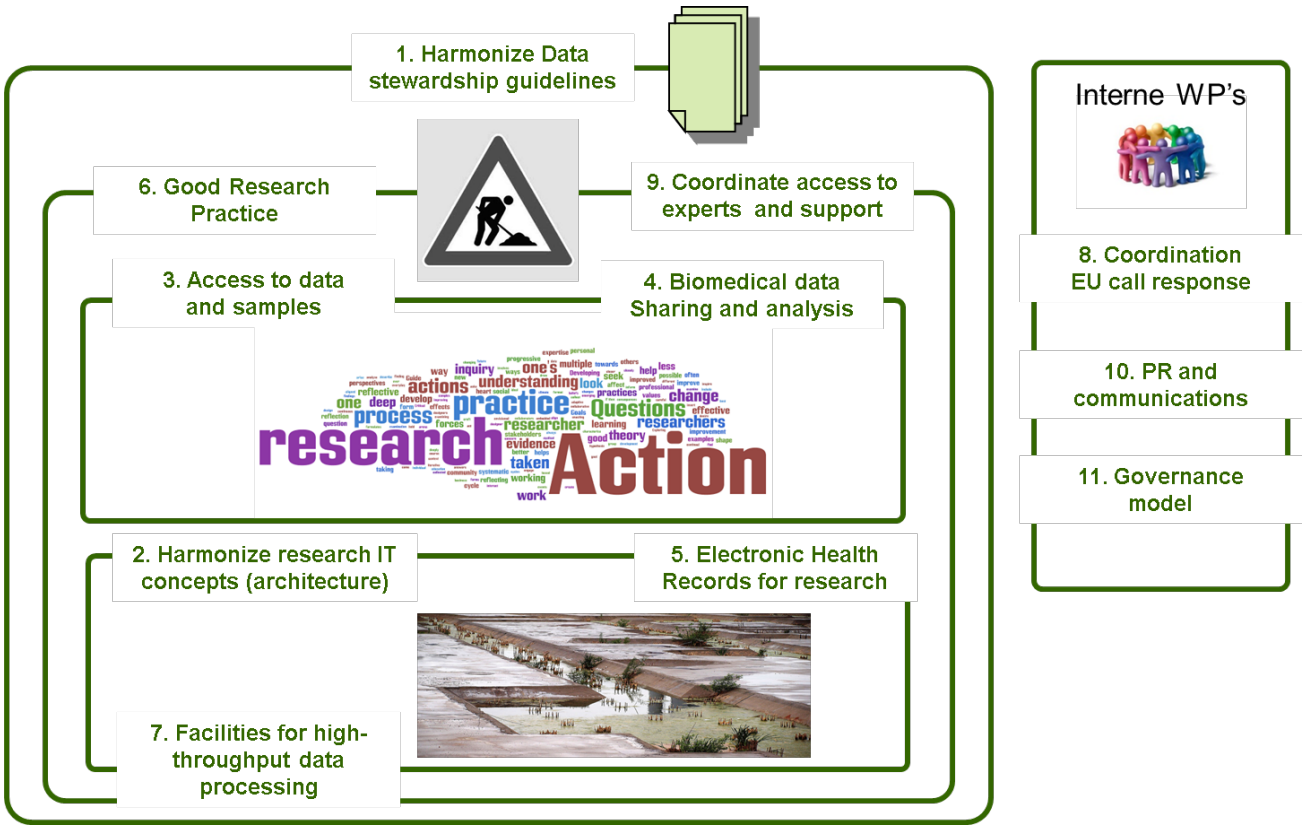

Figure 1 Overview of work packages NFU Data4Lifesciences program

During phase 1 of this program running from 2014 – 2018, the challenge for Data4lifesciences work package 2 was to align IT concepts and (use of existing/established) standards within and across the UMCs in close collaboration with all other work packages (WPs) of this program as well as with relevant partners, such as *SURFsara*, *CIT* (Centrum voor Informatie Technology), *Nictiz*, *STZ* (Samenwerkende Topklinische opleidingsZiekenhuizen) and others, as a basis for all future choices and decisions. To make sure the architecture promotes innovation it is being developed within an open community of scientists, IT specialists, and decision makers. The leading use case from a national perspective is "personalized or precision medicine"[3].

[1] *http://datafairport.org/* and *https://www.force11.org/group/fairgroup/fairprinciples*
[2] *http://www.enrio.eu/codes-guidelines-3/national-codes (http://www.enrio.eu/home)*
[3] Aronson & Rehm, Nature, 15 October 2015, doi:10.1038/nature15816

As the program's target is to develop a unified, coherent biomedical research data infrastructure covering the entire data life cycle (see Figure 2) for at least the eight academic hospitals in the Netherlands, it is desirable to design a consistent architecture for this infrastructure which is supported by the UMCs.

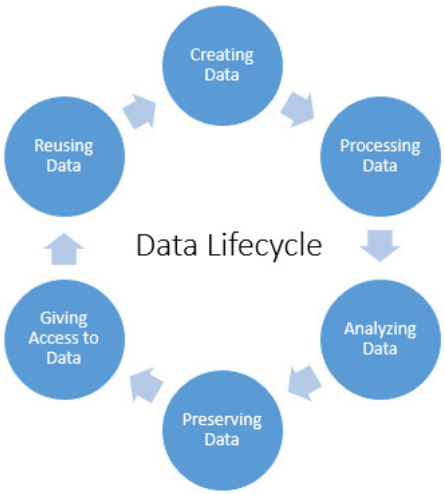

Figure 2 http://www.data-archive.ac.uk/create-manage/life-cycle

A consistent architecture enables smooth data exchange between researchers across UMCs, it aligns emerging IT infrastructures both locally and (inter)nationally, and where feasible it offers the opportunity to make use of each other's IT solutions. Examples include:

- Data capture/entry at the point-of-care (*Registratie aan de Bron*). Traditional research uses separate specific data entry channels. With the current implementations of EHR systems researchers may obtain part of their data (in)directly from those EHRs
- General facilities like storage and computing clusters when in need of peak capacity
- Research-specific facilities such as:
    - analysis pipelines (*i.e.* a set of data processing elements connected in series, where the output of one element is the input for the next one);
    - expensive equipment and software;
    - but also the expertise to run pipelines and operate the equipment and software.

The principles, standards, and frameworks that should concern all in-hospital/UMC as well as inter-hospital/UMC activities related to this research data infrastructure can be distinguished in a number of architectural layers (see figure 3, page 4):

Business layer: this layer describes the links between the various parties active in the field of research data: What is the demand? What is the supply? How do the parties collaborate and how is the corresponding funding arranged? An analysis of the current state of affairs will be the basis for the most appropriate future architecture. This layer encompasses the complex dynamics of sponsors, pharmaceutical companies, insurers, patient organizations, contract research organizations, academic and peripheral hospitals, service providers, consultancy agencies, software suppliers, governments etc. An interesting issue in the business layer is the current increase in the number of registries and cohorts focusing on specific diseases or applications. How can these registries be efficiently embedded in the rest of the playing field? This business layer also extends beyond the Netherlands: many of the players involved operate internationally. However, the architecture focuses initially on the Netherlands.

Process layer: this layer describes the processes concerning research data, such as the data acquisition, obtaining of permission of use, pseudonymisation, processing, analysis, storing and archiving. A large number of questions can be defined in this layer; on a general level, questions regarding the findability of datasets (*e.g.* catalogues), the access to data and under what conditions (open, at request, non-commercial), and the harmonization of informed consents, but also on a more concrete level the connections between the different players as defined in the business layer; how can the processes be organized in such a manner that parties with different goals can still use similar

processes. For example: collective processes for delivery of data to registries, quality assurance, insurers and research cohorts.

Data / information layer: this layer concerns the consistency and presentation of data, which covers questions regarding standardization and adding metadata, but also the topic of pseudonymization.
In this layer, important subjects are: the further roll-out and application of detailed clinical models (or '*zorginformatiebouwstenen*' (ZIBs) in Dutch) in the research domain, the use of ontologies and vocabularies, the standardization of pseudonymization services, use of standard formats, and requirements for quality and reliability of data. An important issue in this layer is bridging differences in standards used by researchers versus standards used in health-care systems in such a way that researchers can still work with the data without needing expensive IT support.

Application layer: this layer deals with the 'application landscape' in (and where applicable between) the UMCs. The diversity of applications will be a starting point, but it is imperative these applications are interoperable, also beyond the borders of the UMCs. Important subjects in this layer are: agreements regarding the standardization of interfaces, concerning the design of data warehouses and storage facilities, harmonization and exchange of analysis pipelines, but also subjects such as authentication and authorization across applications.

Technology layer: this layer describes the technical infrastructure, the hardware necessary to run the research data infrastructure. This includes fast and secure network connections (*e.g.* light paths), and the cohesion of High Performance Computing (HPC) and storage clusters with all its specific security demands.

The following illustration gives an impression of the 5 layers.

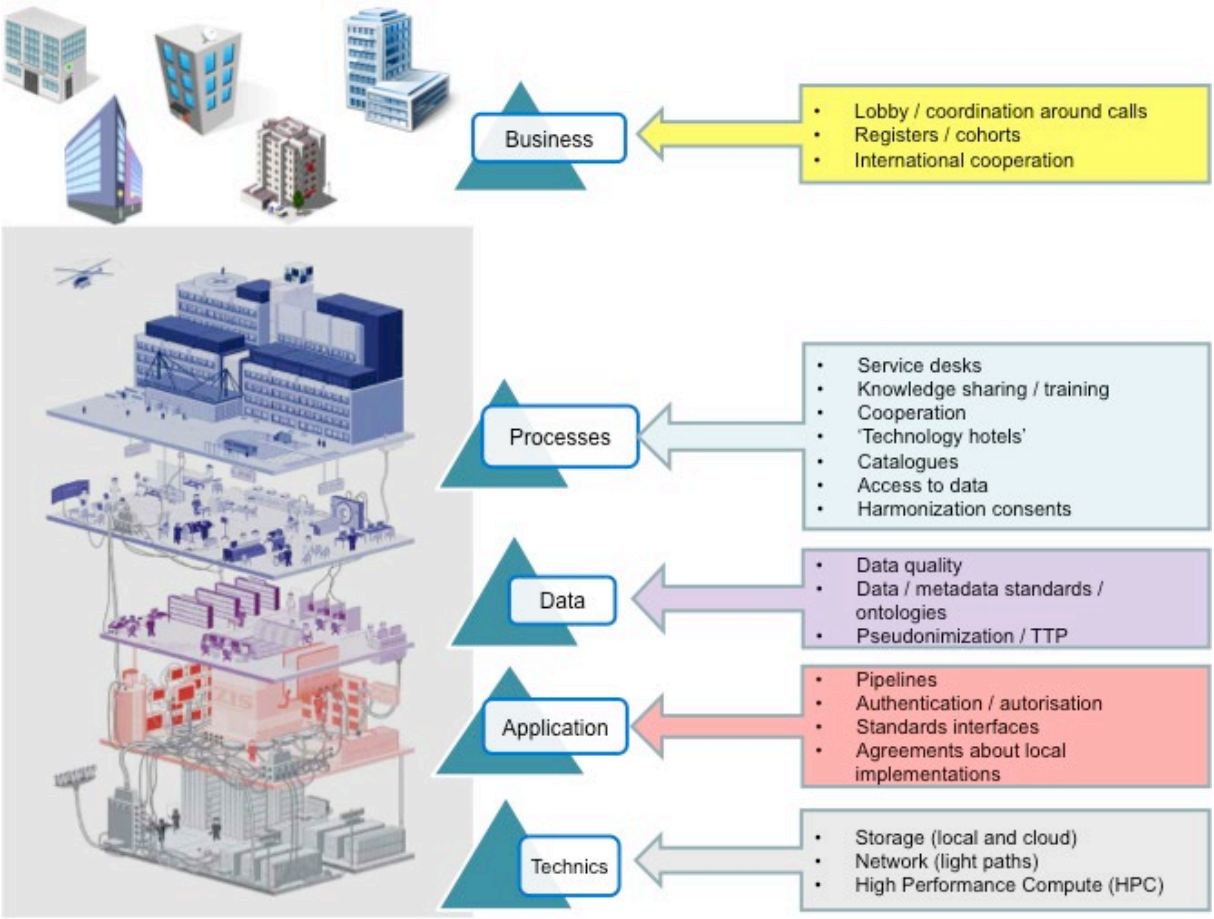

Figure 3: The principles, standards, models and frameworks that should concern all in-hospital/UMC as well as inter-hospital/UMC activities related to this research data infrastructure are defined through various layers in the architecture.

1.1. Current situation

Two national architecture working groups that were already operational, one originally coordinated by CTMM-TraIT, the other, consisting of architects from all UMCs, organized in a Special Interest Group of NFU: SIG-PRIMA, have merged into the SIG-PRIMA. Recently the name of this group changed to "SIG Architecture".
In March 2018 members of HORA, LCRDM and Data4lifesciences have published the report "Geintegreerd ontwerp Architectuur en Onderzoek. Synergie HORA – LCRDM – D4LS" as result of identifying the synergy between the three initiatives working on architecture with emphasis on the domain research. As a result, the working group has published an updated high-level architecture view showing the relationship between the research study life cycle and the research data life cycle (Figure 4).

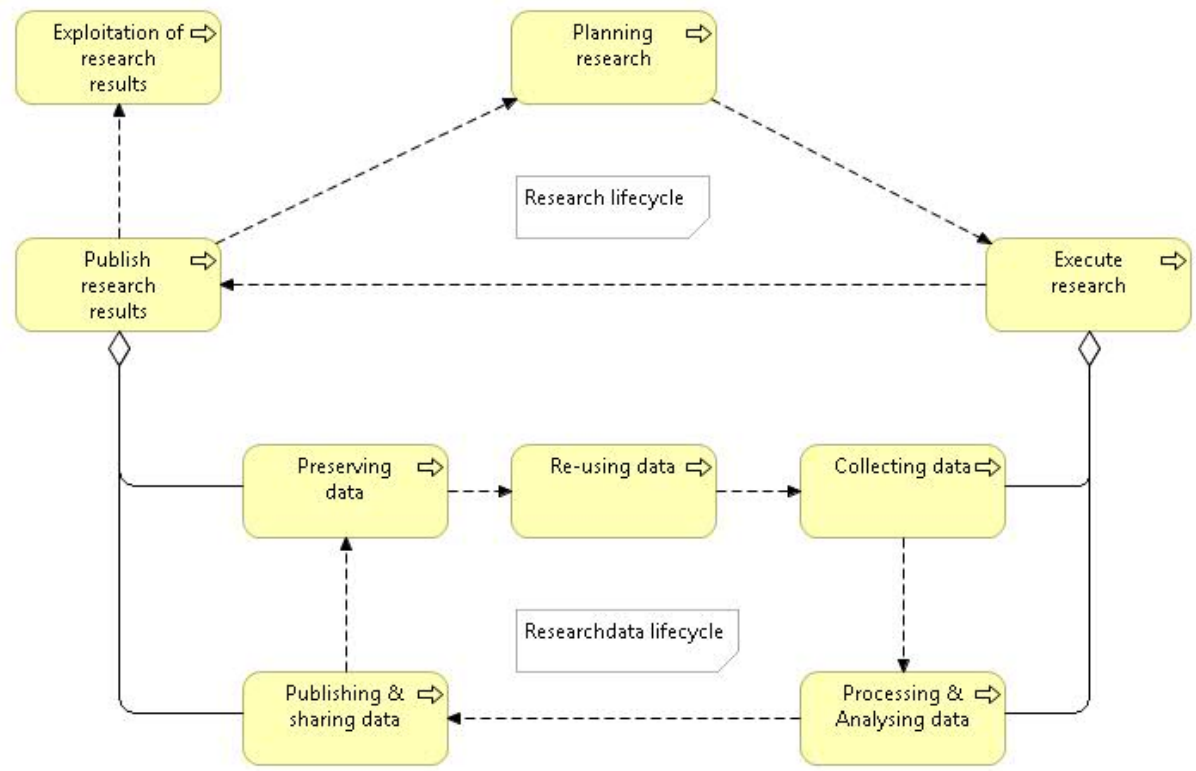

Figure 4 High-level architecture diagram showing the relation between the research (study) and research-data lifecycles (the dashed lines with arrow heads show flow relations, e.g. Planning research flows into Execute research; the diamond shaped lines are aggregation relations).

1.2. Original aims stated

The Data4lifesciences program has specified the following goals for work package 2 "Harmonize research IT concepts (architecture)" in September 2014:

1. To develop standards and architectural principles for the gathering, processing and releasing of data in close cooperation with the program 'Registratie aan de bron' ('Data Capture at the Point of Care').
2. To coordinate the realization of a shared research data and analysis infrastructure and align and integrate results of (inter) national infrastructure projects.
3. To make collective agreements with strategic partners, field parties, among which SURFsara and CIT.
4. To harmonize IT implementation to enable integration and de-duplicate solutions.

In the next chapter the overall aim, also called the True North, of Data4lifescienses and the redefined aims for this work package are described.

2. Deliverables plan

2.1. Overall aim of the Data4lifesciences program – True North

The overall aim or True North of the Data4lifesciences program, as defined by André Dekker (*WP5 lead*), but modified for this WP, is:

What	all clinical, biosample, imaging and experimental data, including metadata (e.g. context, definitions of data) of all patients and/or study-participants should be made available
Who	by every UMC of the Netherlands
Why	for every valid research question
When	now and forevermore
Where	in a scalable, secure, and distributed environment
How	as findable, accessible, interoperable and re-usable (FAIR) data, with full protection of privacy of patients, with full control by the owner (e.g. UMC; taking into consideration the "privacy" of researchers and research consortia to develop products as result of their research) and without disturbing/impacting clinical care

2.2. Specific aims and deliverables for WP2

Proposed approach in original plan
A working group consisting of specialists from all UMCs coordinates the elaboration of the architecture for the research data infrastructure. The original CTMM-TraIT group and the NFU Special Interest Group SIG-PRIMA have been the starting point, although their focus had to be extended in particular for the business and process layers. This working group forms the center of a number of special interest/working groups, organized around specific themes. The views raised by researchers will be leading in these interest/working groups; but also experts working for stakeholders other than the UMCs (such as related sciences, pharmaceutical companies, ICT companies, insurance companies, sponsors, the government) can be member of these groups.
Next, we will contact projects that have handled similar (inter)national or overarching architectural harmonization issues, like U2CONNECT (especially the universities) to obtain exchange of knowledge and experience.

The working group will:

- collect, compare and discuss available architecture pictures/views of the local UMCs;
- perform a quick scan on existing national architecture/initiatives (*e.g.* Parelsnoer (PSI), BBMRI, TraIT, DTL, Nictiz);
- identify other local/national (either top-down or bottom-up) initiatives (*e.g.* STZ, Santeon), and
- in close collaboration/co-creation with the other work packages (were relevant) of this program contributing to a joint reference research IT architecture as depicted in the table at paragraph 4.2;
- to collectively define future harmonization of the national research IT (architecture) concepts against which new/running bottom-up initiatives can be tested.

Further discussion about the required organization and communication strategy of this work package is required. In order to arrive at a functional architecture build, maintained and used by the biomedical research community we will especially stimulate bottom-up initiatives of researchers and other stakeholders dedicated to a specific aspect of the architecture. Otherwise the architecture may become static and risk-adverse, while it should also embrace new developments in science. It is therefore important to link with international developments and standardization trajectories/efforts.

The architecture will be developed step by step. A version of the architecture will be delivered in every plateau of the Data4lifesciences program, which will subsequently serve as a basis for the decision making with regards to the planning of the next plateau(s).

WP 2 specific aims and the associated activities/deliverables contributing to the True North are:

Number	Activity/Deliverable	Approach	Due date	Status 09/2018
1	Governance structure	Join agendas between existing architecture groups: - CTMM TraIT - NFU-SIG PRIMA - SURFsara/SURFnet Determine & establish mandate/authority/governance		Done SIG-architectuur
2	Inventory	Site visits UMCs	Q4 2015	Done
3	Gap analysis	Based on inventory and first workshop determine needs, gaps, and set priorities	Q2-3 2016	Done
4	Workshops	Organize one or two meetings/year to: - Define/update actual use cases - Share best practices - Identify issues, basic architecture principles, and potential solutions - Present/discuss outcome thematic working groups Output will be generated in the form of reports, presentations, etc. contributing to, and outlining, the roadmap to a shared and harmonized (transition) architecture between the UMCs.	Feb 2016 Nov 2016 Mar 2017 2018 2019 2020	Done Being planned
5	Thematic working groups	Start thematic working groups - Determine themes and set priorities - Seek lead per theme - Organize meetings (expected 2-4 per theme). - Establish theme-related deliverables/milestones - Establish resources/budget needed - Write report and present results at next workshop	Fall 2016	Done Ongoing task
6	Draft (transition) architecture in national research data infrastructure handbook	Create and release incremental versions (per D4LS plateau) of handbook on a joined (transition) harmonized architecture	Q4-2016 v1 Q4-2017 v2 Q4-2018 final	delayed
7	Set-up WP2 working environment	Create workspace under D4LS teamsite	Q1 2016	Done
8	Joined architecture view/picture	Joined NFU architecture view/picture including the architecture demands from other D4LS work packages	Q2-3 2016	Approved by NFU AcZie
9	Joined glossary	Common glossary on terminology for research and synchronize these with those set-up by other programs/infrastructures	Q3-2017	Done (hosted by LCRDM)

In relation to the activities and deliverables we will either as part of an activity or deliverable, or across work packages:

- Draw up an inventory of existing architecture (principles/initiatives)
    - Collect, compare and discuss available architecture principles/pictures/views of the UMCs
    - Collect, compare and discuss available data/information models like CDISC-ODM & CDASH, MIABIS (biobanking catalogues)
- Define or choose necessary reference models/architecture (like ISO29585/22221) as well as other relevant standards
    - Including standards and applications to be used for communication within each layer of the architecture, as well as between layers.
- A joint functional architecture handbook, connecting the UMC enterprise architectures and closely related to the HANDS and Health-RI/TraIT websites
    - Define and agree on architecture principles
    - Define and agree on architecture/governance processes
        - comply or explain (when deviating) for new requests
        - create a plan-do-act-check cycle for the architecture handbook/processes
    - paragraphs on components, roadmap, maturity model, reference models/architecture, transition architectures to help harmonize the implementations as well as relevant “maintainable/sustainable” standards
- Interchangeable architecture on all UMC architecture layers
- Share/disseminate/communicate:
    - Communication plan (WP-specific as well as part of overall communication plan)
    - Information, expertise and knowledge
    - Infrastructures
- Network of people/architects to provide input and give feedback
- Link care and research data domains

3. Time schedule

This paragraph gives an overview of the long, mid-term and short-term timelines of which the deliverables as presented in paragraph two (with more detailed information and/or timelines) are part of.

3.1. Long term (2020+)

In the long term, to get to the program's True North, a large scale facility for (bio)medical (translational) research data in the Netherlands is needed (as also pitched at the KNAW-Agenda "Grootschalige Onderzoeksfaciliteiten": *https://www.knaw.nl/en/advisory-work/advisory-reports-and-foresight-studies/lopende-adviezen/agenda-for-a-large-scale-research-infrastructure*) and currently operationalised under the Health-RI initiative.

3.2. Medium term (2018-2020)

- Define a joint generic national architecture vision on research of which the architecture set-up by data4lifesciences is a specific use case
    - Partner with other research infrastructure related projects:
        - national level BBMRI-NL2.0, Parelsnoer (PSI), Health-RI, LCRDM (Landelijk Coordinatie Research Data Management), HORA (Hoger Onderwijs Referentie Architectuur), and other relevant initiatives

  And if feasible link this generic national vision to the international one:
    - Partner with other international research infrastructure related projects:
        - like CORBEL and EXCELERATE
- Implement standards and where feasible include internationally accepted/established standards
    - Team up with (other) standards developing organizations (SDOs) like CDISC
- Obtain structural and innovation budget(s) for sustaining and enhancing the architecture of the Dutch research infrastructure for biomedical research

3.3. Short Term (2018)

Most of the short-term timelines are already presented in paragraph 2. Below, only those are listed for which a more detailed description and/or timeline is available:

- Further detailing of the reference architecture based on priorities and subjects set by D4LS
    - New version of the reference architecture
        - Focus on process layer: see *https://github.com/jambelien/Data4lifesciences-reference-architecture*
    - iRODS, an Open Source Data Management Software solution supporting data management
    - Defining and setting up governance principles and reference architecture to support federated collaboration
- Implementation of Federated Identity and Access Management (FIAM)
    - COmanage: part of the SURF Science Collaboration Zone (SCZ)

4. Resources/deliverables expected from or created in collaboration with other WPs

The following resources, deliverables can be drawn from, or will be (co-)created in close collaboration with other work packages within the program.

Personnel/Products contributing to the D4LS WP2 output
Data stewardship guidelines (WP1) - guidelines that match the reference architecture and can be used within that context
Discovery and access to data (WP3) - reference architecture for national catalogues of research data sets and biosamples
Biomedical data sharing & analysis (WP4) - reference architecture for data acquisition and research collaboration spaces/portals, both for local and multicenter use.
Using clinical data for research (WP5) - reference architecture to link care and research o registration at source, multi-usage, meaningful use of (research) data - framework/reference architecture to obtain care data - medical intelligence reference architecture
Good research practice (WP6) - privacy and security incorporated by design in the reference architecture - NEN norm for pseudonimization services - National TTP service for research - Data modelling (also indicated by increasing amount of registries)
Facilities for high-throughput data processing (WP7) - a reference architecture for HPC peak capacity handling - governance principles and reference architecture to support federated collaboration
Coordinate access to experts and support (WP9) - network of experts that (can/will) contribute to architecture
Public relations and communications (WP10) - Information package in “Jip&Janneke style” to handout to all involved/interested in Data4lifescience program

5. Organization plan

5.1. Project team

Name	Home Institute/Project	Project Role
Jeroen Belien	Amsterdam UMC, Vrije Universiteit Amsterdam, TraIT, BBMRI	Lead
Karel van Lambalgen	LUMC Leiden	Mentor
Frits van Merode	MUMC	Mentor
Hans van den Berg	Amsterdam UMC, AMC	UMC rep and lead architect
Robert Veen (till 31-10-2016) Per 01-11-2016 Martine Ros	UMC Utrecht	UMC architect
Igor Schoonbrood	MUMC+ Maastricht	UMC architect, member SIG-PRIMA
Lisan Scheepe	Erasmus MC Rotterdam	UMC rep/architect
Rob Cornelisse	LUMC Leiden	UMC architect
Ernst de Bel	Radboud UMC Nijmegen	UMC architect
Michael van der Zel Fred Ahsmann	UMCG Groningen	UMC information architect UMC lead architect per 11-2016
Fons Ullings	Amsterdam UMC, Vrije Universiteit Amsterdam Amsterdam	UMC lead Architect
Irene Nooren	SURFsara	WP7-lead & Community Manager Research

5.2. Stakeholders

Contact	Stake
Harry Pijl	Program Manager Research IT UMCU
Erik Flikkenschild	Information Manager Research LUMC
Christiaan Geertsma and/or Arno Sinjewel	ICT Manager Onderzoek & Onderwijs Amsterdam UMC, Vrije Universiteit Amsterdam

6. Communication plan

Information on the progress and results of the Data4lifesciences and when relevant the architecture work package will be actively and openly communicated within the Data4lifesciences project teams, related stakeholders/partners/projects like BBMRI, TraIT, as well as the wider community. This section describes the information plan for people outside the direct scope of this work package.

6.1. WP Stakeholders

Organization	Contact person	Min. frequency	Type
D4LS	Mentors	Monthly	Personal[4]

6.2. D4LS Stakeholders

Organization	Contact person	Min. frequency	Type
D4LS-Overall	Project manager	Monthly	Personal[5]
Operational team		Monthly	F2F meeting or TC

[4] Personal = E.g. a face to face meeting or video- or teleconference with the contact person
[5] 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.

Risk description	Probability	Impact	Hazard (P*I)	Action
1. Not enough funding	6	8	48	Contingency - Redraft and/or reprioritize (parts of) project plan
2. Business case for individual UMC too weak: local IT infrastructure projects prevail above a harmonized national IT infrastructure	6	6	36	Mitigation - Prioritize and continue with those that do support/contribute
3. Not clear how many hours can be spent and are needed per participant.	6	4	24	Mitigation – when possible describe deliverables/tasks as SMART as possible
4. Unclear or non-operational governance on architecture	8	8	64	Make this the first/major deliverable of this work package. Try to set-up governance by starting with a pilot using first (small) result

8. References

http://datafairport.org/ and *https://www.force11.org/group/fairgroup/fairprinciples*

http://www.enrio.eu/codes-guidelines-3/national-codes (*http://www.enrio.eu/home*)

http://www.data-archive.ac.uk/create-manage/life-cycle

Reference architecture Data4lifescience: *https://github.com/jambelien/Data4lifesciences-reference-architecture*

Research Data Management (RDM) glossary: *https://www.edugroepen.nl/sites/RDM_platform/SitePages/RDM%20Glossary.aspx*

Aronson & Rehm, Nature, 15 October 2015, doi:10.1038/nature15816 *http://www.nature.com/nature/journal/v526/n7573/index.html#insight*

9. Glossary and abbreviations

BBMRI: Biobanking and BioMolecular Research Infrastructure, *http://bbmri.nl/* and *http://www.bbmri-eric.eu/*

CTMM: The Center for Translational Molecular Medicine, *www.ctmm.nl*

CTMM-TraIT: A Sustainable Infrastructure for Translational Biomedical Research, *www.ctmm-trait.nl*

ELIXIR H2020 Excelerate: see for more info *https://www.elixir-europe.org/news/elixir-accelerates-major-horizon-2020-funding*

FAIR: Findable, Accessible, Interoperable and re-useable, *http://datafairport.org/* and *https://www.force11.org/group/fairgroup/fairprinciples*

FIAM: Federated Identity and Access Management

Health-RI: Empowering personalized medicine and health research, *https://www.health-ri.org/*

HORA: Hoger Onderwijs Referentie Architectuur *https://hora.surf.nl/index.php/Hoofdpagina*

LCRDM: Landelijk Coördinatiepunt Research Data Management *https://www.lcrdm.nl/*

NFU: Nederlandse Federatie van Universitair Medische Centra, *www.nfu.nl*

SCZ: Science Collaboration Zone

SMART: *https://en.wikipedia.org/wiki/SMART_criteria*