**Project Description – Project Proposals in the Area of Scientific Library Services and Information Systems**

**LIS Funding Programme or Call “e-Research Technologies, the consolidation and optimization of e-Research technologies”**

**An Analytical Research Platform for Distributed Health Care Networks” (ArCARE)**

**Grant applicants:**

**Prof. Dr. Yahyapour, Ramin, Göttingen**

Director of Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

**Prof. Dr. Reinhoff, Otto, Göttingen**

Director of Department of Medical Informatics (MI), University Medical Center Göttingen

**Prof. Dr. Dihazi, Hassan, Göttingen**

Clinical Proteomics laboratories, Clinic for Nephrology and Rheumatology (NR/BIN), University Medical Center Göttingen

**Prof. Dr. Asif, Abdul Rahman, Göttingen**

Institute of Clinical Chemistry (KC)/ UMG-Laboratories

**Prof. Dr. König, Simone, Münster**

Head of IZKF Core Unit Proteomics (CUP), Medical Faculty, University of Münster

**Project Description**

**“An Analytical Research Platform for Distributed Health Care Networks” (ArCARE)**

# Starting point and preliminary work

“Omic” sciences (“omics”), also referred to as high-dimensional biology, are a comparatively new big-data research approach in the life sciences aiming at the study of structure, functions and dynamics of biomolecules in living organisms. Information technology is an integral part of this research given that the comprehensive analysis of the colossal amount of data spawning by the research instruments is vital. Mass spectrometers, for instance, are some of the most multipurpose analytical instruments and have been extensively used in omics research including proteomics and metabolomics for the identification of analytes such as proteins or small molecules. In a state of the art proteomics lab with several mass spectrometer dozens of data files worth several GBs each are produced and analyzed daily. According to recent statistics, the data store size of one of the world’s largest bioinformatics lab, the European Bioinformatics Institute (EBI) in the UK, reaches 120 petabytes in 2017 [1].

The challenge of vendor lock-in applications and data formats is a known issue in mass spectrometry (MS)-based omics research. There are about five major mass spectrometer manufacturers dominating the market and competing with each other in terms of technology, design and performance. Typically, each vendor supports one or more native data formats of its own and updates them as required by emerging instrumentation. Vendor software typically differs from that of competitors and is therefore non-interchangeable. Consequently, only the vendor proprietary applications can be primarily used to read and analyze the data, impeding the pace of collaborative work of various research labs running different types of mass spectrometers in terms of exchanging and comparing of results of different biological experiments. Nevertheless, efforts have been made internationally to deal with this issue, and a variety of vendor-neutral formats (also called open formats such as mzXML) and open-source analysis tools have been developed to enhance the sharing of results and analysis tools among the researchers in the community. The HUPO Proteomics Standard Initiative (HUPO-PSI), for instance, is one of the most active communities in proteomics working for the development of standardized data representations for data exchange and verification based on vendor APIs (Application Programming Interface).

In addition, enabling reproducibility in omics research is a challenging task not only experimentally but also on the data analysis end. Providing MS data for shared repositories is a complex multistep operation, which involves collecting the raw data from instruments, their extraction to one or more open data-formats and their storage in suitable databases as well as the analysis of the data with respect to biomolecule identification and statistical observations. Depending on the experimental task, the high-throughput workflow is amenable to automation, which saves time and improves consistency and transparency of data processing. Again, vendor-specific applications to that effect such as the Waters Symphony data pipeline software are available, but they do not help research facilities with several instruments from different manufacturers or large omics research networks.

In summary, MS is a key multidisciplinary analytical technique heavily used in omics research such as proteomics, glycomics, lipidomics and metabolomics. A centralized research infrastructure uniting various both local and distant MS labs is vital in particular with regard to the current initiatives by funding agencies to integrate them more in systems medicine. Such infrastructure should provide:

* A secure and highly scalable network environment of the researchers’ instruments.
* Storage of data into a both safe and accessible data repository.
* A user friendly working environment for each lab with respect to data processing.
* Centralized policy management and control to help enforce disciplinary regulations and policies in order to gain proper control and prevent uncontrolled distribution of data.

## Current Practice

The Göttingen Proteomics Forum (GPF) is a local network of scientists and researchers of the Georg-August-University, the University Medical Center Göttingen (UMG), the Max-Planck Institutes (MPI) of Experimental Medicine and Biophysical Chemistry, and the local corporate data processing facility of the University and Max-Planck-Society (GWDG) with the common interest in the MS-based analysis of proteins. The aim of the forum is the pooling of local proteomics expertise and tools to generate synergism and provide mutual support for the scientific community in Göttingen.

At present, the proteomics labs of the GPF network operate a broad range of state-of-the-art MS instrumentations and software (listed in Table 1). Based on its vendor lock-in technology and solutions, each lab follows a set of standardized operational instructions for raw data collection, data conversion to meaningful formats, and the analysis and storage of data for research purposes. GWDG is the data processing center for GPF and supports the proteomics labs. For this, GWDG hosts a MASCOT server, a proteomics analysis software, which has become a standard for protein identification using MS data.

All processes and operations at GPF proteomics labs and GWDG currently are operated by manual routines and hence are prone to reproducibility, productivity and transparency issues. Also, in its current setup, the existing infrastructure does not support common and public data repositories for an open exchange of data across different research settings. Furthermore, the lack of standardized workflow practices among the labs and non-adherence to open data standards also hinders a systematic and spontaneous sharing of results among the institutions.

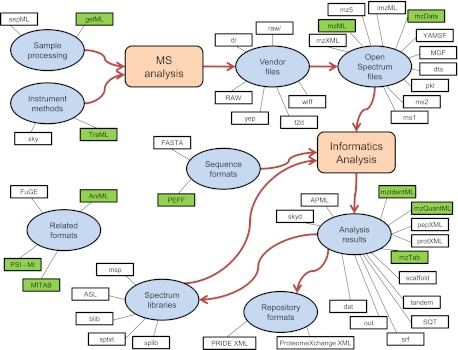
**Table 1:** Mass spectrometers available at NR/BIN, KC

|  |  |  |
| --- | --- | --- |
| **Institute** | **Vendor MS Devices** | **MS Software** |
| NR/BIN, KC | * Q Exactive (Thermo Scientific) * Q-TOF Ultima (Waters Corp.) * TripleTOF 5600+ (AB Sciex) * Rapiflex-Tissue typer (Bruker) | * Mascot (Matrix Science) * Scaffold (Proteome Software) * MaxQuant with Perseus * Biotools * SCILS |

## Available Tools and Technologies for the Development of ArCare

MS is a cross-disciplinary data-driven research technology, which produces enormous amounts of data in omics labs on a daily basis. Currently, researchers follow manual routines of data analysis workflows including but not limited to:

* Reading raw-data files using vendor-specific software.
* Storage of raw-data on their lab data server or at institutional IT facilities.
* Conversion of raw-data into multiple intermediary formats as needed for subsequent processing with advanced software tools (Figure 1).
* Data mining using a broad range of digital tools available from both vendors and open-source communities.
* Generating analysis documentation to be stored alongside the data.



**Figure 1:** Data formats in MS-based proteomics

The automation of the analysis part of omics-derived large data streams would require the distribution of data at two different levels, i.e., locally within a lab, and, second, within shared repositories of various omics labs. Moreover, publishers’ policies often require post-analysis results and raw-data to be submitted along with manuscripts in order to facilitate their exchange within the research community. Therefore, there is a great need of a simple but robust and scalable network backbone that supports the transportation of such complex data efficiently. Current campus networks are not yet ideally designed for, or capable of, coping with these challenges.

Furthermore, data analyses workflows vary from one lab to another due to differences in research task, the available equipment, and the local policies. Therefore, it is important that a software-based automation system should flexibly support different and changing requirements in omics labs.

## 1.2.1 ArCare Proposal

ArCare proposes a secure, robust and automated research platform for omics labs. The heart of the solution is an advanced network infrastructure based on the novel concepts of software-defined networking (SDN) providing a sophisticated control and management plane for increasing the run-time efficiency of data center applications. With a new programming model for packet forwarding devices, SDN supplies a network controller concept (also called SDN controller) which acts as a manager of the network and is able to configure the network switches dynamically based on the instructions of a policy controller. ArCare offers:

* The automation of different MS data analysis workflows (from data collection to data storage and analysis) currently implemented in omics labs.
* The support and automation of standardized workflows (defined, e.g., by HUPO-PSI) based on open data formats and open source tools and software.
* Integration of data into existing transregional repositories, such as PRIDE, for a sustainable and open exchange of data across different research settings.

Figure 2 depicts the ArCare design based on off-the-shelf tools and technological testbeds developed at the Göttingen campus. In a given automation scenario proposed by ArCare, the vendor-specific raw-data will be captured from the MS devices through device-specific SDN agents, and transformed into a suitable open data format (such as OpenMS) with the help of local data adapter applications. This data is subsequently sent to the ArCare SDN controller, where it is stored in a data repository shared among different research labs. Finally, for data analysis and reporting, the ArCare controller further transforms the OpenMS data into other application specific formats as required by subsequent data analysis applications. The available testbed, tools and technologies that will be used in the realization of the ArCare ecosystem are described below:

**OpenNets:**

OpenNets is the implementation of GWDG’s SDN testbed based on the OpenDaylight controller (Lithium version) extended with additional domain-specific SDN applications. OpenNets exposes the REST API of the application affinity service component of the OpenDaylight controller, which allows for automation of provisioning and deletion requests of network connections at the switching/forwarding plane. The capabilities of OpenNets also include the advanced network management of legacy network infrastructure using SDN agents. The SDN agent is a novel concept to bridge the communication gap between the SDN control plane and non-SDN data plane. In order to provide centralized control and automation of MS big data labs, we need to utilize the strength of SDN agents to facilitate the communication between the ArCare controller and the network switches. Also, we need to introduce device agents to collect and transfer the MS data from the vendor devices. (Figure 1)

**MS Proteowizard:**

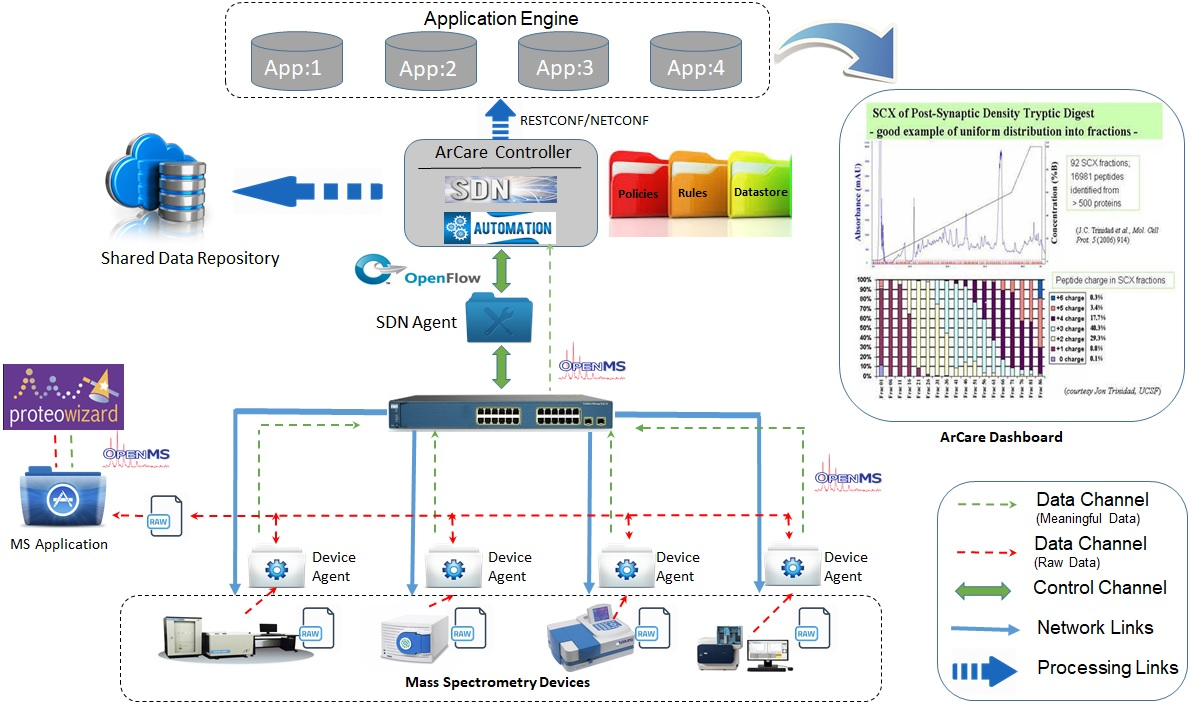
In MS-based research it is often necessary to convert the raw data from vendor-specific binary files to open-format files. In order to unify access to MS data files and enforce standard chemistry analysis and computation, the MS Proteowizard project provides a modular and extensible collection of cross-platform and open-source tools and libraries for proteomics data analyses. It delivers a framework to read different vendor-specific and dedicated formats and transform the data to open formats.

**MS Software Applications:**

During the last two decades, MS has been heavily integrated into life science research, which is largely due to shot-gun profiling of huge biomolecule populations such as proteomes or metabolomes. Both technology and software tools [7] were developed alongside. In ArCare, both proprietary and open-source software, depending on local availability, will be implemented.

**OpenMS:**

OpenMS is a skillful open-source library for the analysis of MS data. It provides a data structure framework from the basic format capture from input/output file and visualization to advanced and mature algorithms. OpenMS allows developers to turn their attention to novel algorithmic approaches instead of focusing on implementation of infrastructure.



**Figure 2:** Overview of ArCare ecosystem design

## Comparison to Alternative Solutions

Scientific research is extremely data driven. Tremendous collected data was produced by plenty scientific research. Currently it is not possible to manage such a massive volume of data in campus networks.

In the area of life-science data, as one of the straight example of big data, is increasing to an impressive rate— over 40% yearly [2]. In the domain of Biological data, the European Bioinformatics Institute (EBI), as a part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year [3].

Today’s data center infrastructures and architecture add to the difficulty, since they are frequently based on branded storage and network technologies which are challenging and expensive to manage and scale. Life-science IT decision makers requires new opportunities to scale quicker and more cost effectively [4]. As a novel technology, recently organizations are using software-defined networking (SDN) solutions as well to form IT systems more easy and flexible. Research and Markets analysts anticipate that the SDN/NFV market is awaited to increase at a CAGR of 48 percent through 2025 because of the requirement of making IT systems smaller and simpler to control [5].

Study shows life-science and Bioinformatics data are one of the best context to deploy SDN/NFV technology [6].

## 1.3.1 SDN Life Science Solutions

## The University of Pittsburgh Medical Center UPMC [10] is a world-renowned healthcare provider with 450 sites including 21 hospitals. Its entire IT infrastructure operation is supported by two data centers. It uses virtualization for solving problems with legacy networks, e.g., stable network topology and long deployment times for their applications. Currently, approximately 90% of all UPMC applications are virtualized. Obviously, traditional networks designed for non-virtualized operation might be problematic in case of using virtualization with huge operational workload joint with legacy networks. In order to overcome networking bottlenecks caused by virtualization, UPMC is using VSP (network Virtualized Service Platform) from Nuage which is a SDN/NFV solution based on hypervisor-agnostic and which is able to work with any type of existing hardware.

## Using SDN/NFV solutions, some of the main networking bottlenecks in a life science environment can be easily solved. SDN brings more functional simplicity, gives improved control over physical workload and provides the ability to specify the necessary bandwidth and assign priority for east-west traffic of any application. Therefore, better QoS (quality of service) for the main applications can be achieved.

**1.3.2 Proteomics Informatics**

For the analysis of MS-data, in particular protein identification, algorithms compare them to sequence information available at public repositories such as GenBank or UniProt. The Technical University of Munich (TUM) together with JPT Peptide Technologies and Thermo Fisher Scientific provides a data analytics platform called ProteomicsDB [12] and a data repository called PRIDE [13]. They have started ProteomeTools [11], a project to translate the human proteome into molecular and digital tools for life science research.

## 1.3.3 Comparison to ArCare

## In contrast to the above mentioned existing solutions, ArCare does not only addresses the limitations of legacy backbone network to process big data health sciences applications, but also provides a streamlined automation framework for omics workflows on top. Such a framework is highly desirable in order to facilitate massive data processing and agility of research operations in life science consortia.

## Project-related publications

### Articles published by outlets with scientific quality assurance, book publications, and works accepted for publication but not yet published

1. Aziz M, Fazely H A, Landi, Gallico D, Christodoulopoulos K, Wieder P (2016) SDN-enabled application-aware networking for data center networks, DOI: 10.1109/ICECS.2016.7841210
2. Landi G, Patronas I, Kontodimas K, Aziz M, Christodoulopoulos K, Kyriakos A, Capitani M, Fazely H A, Reisis D, Varvarigos E, Bakopoulos P, Avramopoulos H (2017) SDN Control Framework with Dynamic Resource Assignment for Slotted Optical Datacenter Networks. DOI: 10.1364/OFC.2017.Tu3L.1
3. LipphardtM\*., DihaziH.\*, Jeon NL., DadafarinS., David W., Rowe DW., Müller GA., Goligorsky MS. Dickkopf-3 in Aberrant Endothelial Secretome Instructs Renal Fibroblast Activation and Endothelial-Mesenchymal Transition. Nephrol. Dial. Transplant. 2018 in press \*: equal contribution
4. Dihazi GH., Olaf J., Tampe B., Zeisberg M., Müller C., Müller GA., Dihazi H. Proteomic analysis of embryonic kidney development: Heterochromatin proteins as epigenetic regulators of nephrogenesis. *Sci. Rep.* 5, 13951; doi: 10.1038/ srep13951 (2015).
5. Khan N, Lenz C, Binder L, Pantakani DV, Asif AR**.** Active and Repressive Chromatin-Associated Proteome after MPA Treatment and the Role of Midkine in Epithelial Monolayer Permeability. Int J Mol Sci. 2016 Apr 20;17(4). pii: E597. doi: 10.3390/ijms17040597.
6. Almamy A, Schwerk C, Schroten H, Ishikawa H, Asif AR, Reuss B. Interactions of antisera to different Chlamydia and Chlamydophila species with the ribosomal protein RPS27a correlate with impaired protein synthesis in a human choroid plexus papilloma cell line. Immunol Res. 2017 Dec;65(6):1110-1123. doi: 10.1007/s12026-017-8952-9.
7. Mohebi A, Aghabozorgi S, Ying-Wah T, Herawan T, Yahyapour R (2016) Iterative big data clustering algorithms: a review. INT J SOFTW ENG KNOW 46 (1): 107-129, doi: 10.1002/spe.2341
8. Zhang F, Liu G, Zhao B, Kasprzak P, Fu X, Yahyapour R (2019) CBase: Fast Virtual Machine storage data migration with a new data center structure. J PARALLEL DISTR COM 124: 14-26
9. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, Knaup-Gregori P, Bavendiek U, Dieterich C, Brors B, Kraus I, Thoms CM, Jäger D, Ellenrieder V, Bergh B, Yahyapour R, Eils R, Consortium H, Marschollek M (2018) HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. METHOD INFORM MED 57(S 01): e66-e81, doi: 10.3414/ME18-02-0002
10. Shen H, Wang H, Wieder P, Yahyapour R (2017) S-fabric: towards scalable and incremental SDN deployment in data centers. In: (Hg.) Proceedings of the 2017 International Conference on Telecommunications and Communication Engineering. ICTCE '17 Proceedings of the 2017 International Conference on Telecommunications and Communication Engineering, Osaka, Japan, October 22 - 24, 2017, 25-29

### Other publications

1. M Sturm, A Bertsch, C Gröpl, A Hildebrandt,R Hussong, E Lange, N Pfeifer, O Schulz ,A Zerck , K Reinert and O Kohlbacher(2008) OpenMS – An open-source software framework for mass spectrometry.BMC Bioinformatics 2008, 9:163, doi:10.1186/1471-2105-9-163
2. J D. Holman,D Tabb, and P Mallick, Employing ProteoWizard to Convert Raw Mass Spectrometry Data (2015),Department of Radiology, Stanford School of Medicine Curr Protoc Bioinformatics. ; 46: 13.24.1–13.24.9. doi:10.1002/0471250953.bi1324s46.
3. Lipphardt M., Song JW., Matsumoto K., Dadafarin S., Dihazi H., Müller GA., Goligorsky MS., The third path of tubulointerstitial fibrosis: aberrant endothelial secretome *Kidney Int.* 2017;92(3):558-568. doi: 10.1016/j.kint.2017.02.033.
4. Zafar S., Behrens C., Dihazi H., Schmitz M., Zerr I., Schulz-Schaeffer WJ., Ramljak S., Asif AR. Cellular prion protein mediates early apoptotic proteome alternation and phospho-modification in human neuroblastoma cells. *Cell Death Dis.* 2017 Jan 19;8(1):e2557. doi: 10.1038/cddis.2016.384
5. Trivedi R., Müller GA, Manohar SR., Mishra DP., Dihazi H. Anti-leukemic activity of shikonin: Role of ERP57 in shikonin induced apoptosis in Acute Myeloid Leukemia *Cell. Physiol. Biochem.* 2016; 39(2):604-16
6. Trivedi R., Dihazi GH., Eltoweissy M., Mishra DP., Müller GA., Dihazi H. The antioxidant protein PARK7 plays an improtant role in cell resistance to Cisplatin-induced apoptosis in case of clear cell renal cell carcinoma *Eur. J. Pharmacol.* 2016 Aug 5;784:99-110*.*
7. Dihazi H., Koziolek M., Datta RR., Jung K., Wallbach M., Heise D., Dihazi GH., Markovic I., Asif AR, Mueller GA FABP1 and FABP3 have high predictive value for renal replacement therapy in patients with acute kidney injury *Blood Purif.* 2016 Jun 30;42(3):202-213.
8. Khan N, Binder L, Pantakani DVK, Asif AR. MPA Modulates Tight Junctions' Permeability via Midkine/PI3K Pathway in Caco-2 Cells: A Possible Mechanism of Leak-Flux Diarrhea in Organ Transplanted Patients. Front Physiol. 2017 Jun 26;8:438. doi: 10.3389/fphys.2017.00438. eCollection 2017.

### Patents

NR/BIN: Patent Title: Biomarkers for renal disease

Pub. No.: US 2010/0035263 A1; Pub. Date Feb. 11, 2010.

# Objectives and Work Programme

## Anticipated total duration of the project

The total duration of the project is anticipated with 36 months (01.01.2020 – 31.12.2022).

## Objectives

The consortium has defined a set of general objectives for the project, which serve as guidelines and lead questions for the overall development:

Objective 1: Data integration into the ArCare domain

Objective 2: Standardized data analysis output and publication support

Objective 3: Transregional research platform

## 2.2.1 Objective 1: Data integration into the ArCare domain

The first main objective is the integration of MS data into the ArCare-domain. Due to wide-scale heterogeneity of vendor specific instrumentations, data types, and data analysis software available in the omics world, it is quite challenging to introduce an all-in-one automation system that fulfills the need to provide for all data analysis workflows. To address this issue, ArCare envisages a two-step approach:

**Compliance to MIAPE guidelines:** In the first step, the objective is to make the ArCare automation system compliant to the MIAPE (The Minimum Information about a Proteomics Experiment) guidelines. MIAPE is a widely accepted set of recommendations for experimentation and analysis that have been developed by the HUPO-PSI community together with the input from a broad spectrum of stakeholders in proteomics. For instance, MIAPE Quant is a module which presents a common set of the most important data types and metadata required for a quantitative experiment to be analyzed and, or a data analysis pipeline to be reproduced [8].

**Introduction of a policy controller:** As a second step, ArCare introduces a policy controller in its solution, which allows the labs to write their own experiment settings and policies into the automation system based on the analysis workflow and the corresponding equipment they intend to use. Nevertheless, the proposal of policy control mechanisms for the operation of the ArCare system implicates to consider ethical, data protection and data consistency issues of different proteomics labs.

**2.2.2 Objective 2: Standardized data analysis output and publication support**

The second main objective of ArCare automation is to define and implement sustainable standards and methodologies to document the outcome of MS data analyses in a single common language, which is a highly demanding feature expected by the proteomics labs in GPF forum. This is required both when a MS data set is stored in a common database for later research and when it is submitted alongside a manuscript to a scientific journal. There is an increasing demand from the researchers’ community for a simple, automated and streamlined approach for the submission of large data sets for publications. To accommodate this need, the project will use MIAPE guidelines to generate a data submission pipeline to publishers.

**2.2.3 Objective 3: Transregional research platform**

The third main objective of the ArCare proposal focuses on the distribution and exchange of life science data among researchers and medical infrastructures at a national level. The effort requires the creation of data models, NoSQL (open-source) database design repositories and configurations, and an automated mechanism of exchange and storage of MS data (raw and processed data) from omics labs in a common data repository. Additionally, the project will consider the integration of the ArCare data system with other life sciences research repositories and data integration projects at the national level, such as PRIDE and HiGHmed [9].

## Work programme and proposed research methods

## 2.3.1 WP1 – MS data collection

Two laboratories will contribute MS omics data to this proposal (Table 2). One is a local Clinical Proteomic laboratories at the UMG (NR/BIN, KC) and the other a distant omics provider (CUP Münster). In this setting, ArCare needs to provide solutions for the different challenges and to adapt to the system inflexibility due to the manifestation of inbred processes. The involved MS laboratories face in their everyday life broad spectrum of challenges. NR/BIN, KC generates data including: protein profiles from biological fluids, peptide sequence information, quantification, information on posttranslational modifications. Moreover, large data from human/animal tissue samples using MALDI-mass spectrometry imaging will also be collected and provided for the project. CUP offers besides bioprofiling also target and quantification experiments. These data collected from CUP and the different proteomics lab in Goettingen Proteomics Forum (GPF) will be provided to ArCare, which have to optimize and provide process for integration of the different data formats generated by the large number of MS instruments.

CUP will provide both benchmarking data, which have been obtained by running standard samples such as *Escherichia coli* or HeLa lysates, and real life samples. CUP will also supply small molecule-profiling data as generated in metabolomics experiments for testing. Different omics use similar but still slightly different experimental conditions for MS measurement and subsequent data analysis, i.e., proteomics workflows involving protein identification cannot be employed for metabolite identification. However, multivariate statistics algorithms work perfectly well on both types of MS data. Therefore, interfaces to this effect need to be programmed into ArCare. In addition, different labs have different access to costly proprietary software, which cannot be shared via ArCare. For instance, CUP uses Progenesis (nonlinear diagnostics), but would be interested in having access to MaxQuant (MPI of Biochemistry), which has been developed for ThermoFisher mass spectrometers. CUP with thus interact with the ArCare bioinformaticians with respect to shared data analysis and incorporation of statistical and comparative software tools across platforms.

Milestones

* MS 1.1 – Identification of the sample data and format, from the available pool
* MS 1.2 – Interacting with ArCare and optimizing the data generation and delivery procedure
* MS 1.3 – Collecting and processing of complex data from MALDI-MS-imaging

Deliverables

In this WP1 there are no deliverables expected as this is part of the GWDG-output.

**Table 2:** MS data provider for ArCare. Available equipment and vendor

|  |  |  |
| --- | --- | --- |
|  | **Facility** | **Type of data and vendor** |
| GPF | Göttingen Proteomics Forum |  |
| CUP | IZKF Core Unit Proteomics, University of Münster | Proteomics experiments using standard lysates  Small molecule profiling experiments biofluids  ->Waters instrumentation |

|  |  |  |
| --- | --- | --- |
| **Work Package 1 Overview – MS data collection** | | |
| WP Lead | **NR/BIN, KC** | |
| Partner | **NR/BIN, KC** | **CUP** |
| PM per partner | 18 | 18 |
| PM total | 36 | |

**2.3.2 WP2 - Formation of SDN infrastructure and integration with omics labs**

A common network infrastructure will be implemented among different types of MS devices available at the premises of the partner labs for MS data collection and transfer to the common repository. Then, an SDN controller which is appropriate for the ArCare project will be selected and implemented. Since SDN controllers have more prominent roles in SDN and functions as strategic control points in order to provide intelligent networks, selecting the appropriate controller is a significant part of this project. There are a variety of Open Source Controllers under development, from NOX to OpenDaylight and ONOS. With our expertise and the fact that OpenDaylight and ONOS can be deployed in diverse production network environments, we selected OpenDaylight and ONOS as potential candidates for this project. As an outcome of this work package, we will set up a SDN testbed for the project, which will provide network service operation and support beyond the project life time. To achieve this, GWDG as the Göttingen campus service provider and the Göttingen eResearch Alliance will provide a prototype data center which will be equipped with essential SDN-Based hardware (Router/Switch) infrastructure.

Milestones

* MS 2.1 – Selection of appropriate ArCare Controller.
* MS 2.2 – Implement physical networking among different omics labs
* MS 2.3 – SDN testbed for ArCare

Deliverables

* D 2.1 - Preliminary report on SDN Infrastructure for ArCare
* D 2.2 - Functional architecture of ArCare automation ecosystem

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package 2 Overview** | | | |
| WP Lead | **GWDG** | | |
| Partner | **GWDG** | **MI** |  |
| PM pro Partner | 20 | 12 |  |
| PM Gesamt | 32 |  |  |

**2.3.3 WP3 - Automation of SOPs in omics labs**

This work package aims at implementation of automated workflow procedures for MS analyses. Currently, omics labs have their individual sets of instructions and procedures (SOPs) in their daily routine. This can and needs to be changed for routine experiments such as expression analyses to automated procedures in an effort to improve transparency and networking. Thus, the ArCare automation setup proposed in Figure 1, will be accommodated according to the needs of each partner lab. To that end, GFP and CUP will agree on experimental procedures, produce data using standard substances and device a data processing routing for ArCare.

Milestones

* M 3.1 – Automation of data analyses processes at different labs

Deliverables

* D 3.1 – Definition of different SOPs from partners to be automated

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package 3 Overview** | | | |
| WP Lead | **GWDG** | | |
| **Partner** | **NR/BIN, KC** | **CUP** | **GWDG** |
| PM pro Partner | 12 | 12 | 12 |
| PM Gesamt | 36 |  |  |

**2.3.4 WP4 - Standardized workflow for reporting and publication**

A clear and standardized reporting of MS results is crucial and of high importance in big-data research. Thus, the MIAPE guidelines will be used as a basis for the implementation of a standardized reporting mechanism of MS experiments. Moreover, reporting routines, the partner labs have in place already, will be accommodated. For instance, CUP uses a locally programmed browser-based searchable internal sample and data management software capable of generating reports. Such local databases will be connected with ArCare to avoid the creation of multiple databases with the similar information. The outcome of this work package will be a standardized workflow of MS-based proteomics and possibly also metabolomics and imaging data for reporting and publication in journals.

Milestones

* M 4.1 – Implementation of an automated submission of publication results to Journals

Deliverables

* D 4.1 – Study and analysis of different modules of MIAPE required for standardized reporting

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Work Package 4 Overview** | | | |  |
| WP Lead | **CUP** | | |  |
| Partner | **NR/BIN, KC** | **CUP** | **MI** | **GWDG** |
| PM pro Partner | 6 | 6 | 6 | 6 |
| PM Gesamt | 24 |  |  |  |

**2.3.5 WP5 - Common data repository and its integration with public repositories**

A data sharing mechanism among different omics labs will be set up. The outcome will be the creation of data models, NoSQL (open-source) database design repositories and configurations, and an automated mechanism of sending and storing MS data (raw and processed) from the labs. Furthermore, this work package also aims at providing an integration mechanism of proteomics data (based on MIAPE guidelines) with available public repositories for a sustainable and open exchange of data across different nation-wide research settings. The candidate systems in this regard are PRIDE, ELIXIR and HiGHMed.

Milestones

* MS 5.1 – Design and implementation of ArCare data repository
* MS 5.2 – Integration of ArCare repository with PRIDE

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package 5 Overview** | | | |
| WP Lead | [**Medizinische Informatik**](https://ecampus.uni-goettingen.de/sb/rds;jsessionid=20DF7A7C3B3DC74119311208F31FB0CB.s46?state=verpublish&status=init&vmfile=no&moduleCall=webInfo&publishConfFile=webInfoEinrichtung&publishSubDir=einrichtung&keep=y&einrichtung.eid=494) **(MI)** | | |
| Partner | **MI** | **GWDG** |  |
| PM pro Partner | 18 | 16 |  |
| PM Gesamt | 34 |  |  |

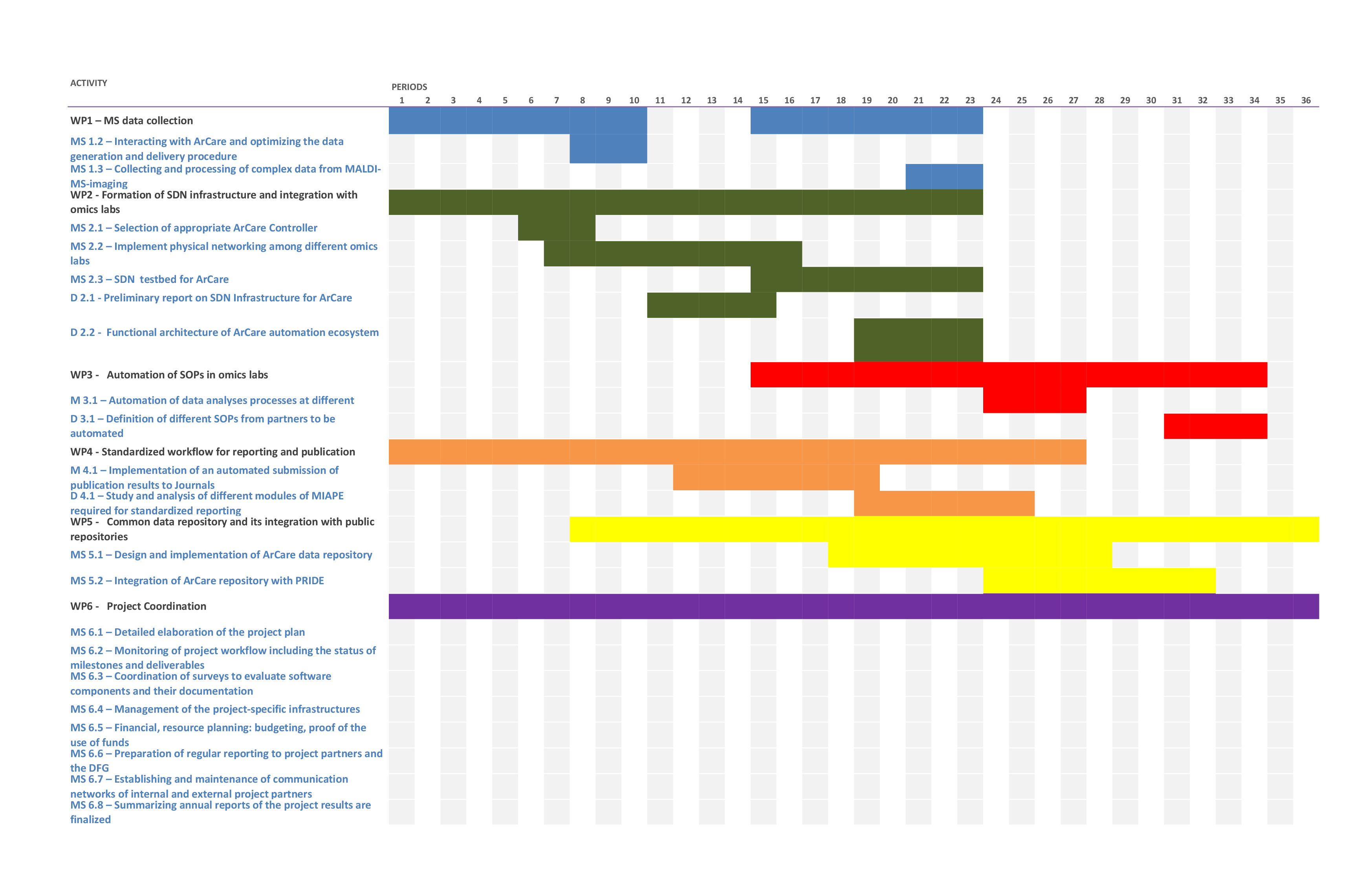
**2.3.6 WP6 - Project Coordination**

Project coordination mainly consists of the following tasks:

Milestones

* MS 6.1 – Detailed elaboration of the project plan
* MS 6.2 – Monitoring of project workflow including the status of milestones and deliverables
* MS 6.3 – Coordination of surveys to evaluate software components and their documentation
* MS 6.4 – Management of the project-specific infrastructures
* MS 6.5 – Financial, resource planning: budgeting, proof of the use of funds
* MS 6.6 – Preparation of regular reporting to project partners and the DFG
* MS 6.7 – Establishing and maintenance of communication networks of internal and external project partners
* MS 6.8 – Summarizing annual reports of the project results are finalized (Months 12, 24 and 30)

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package 6 Overview** | | | |
| WP Lead | **GWDG** | | |
| Partner | **GWDG** |  |  |
| PM pro Partner | 18 |  |  |
| PM Gesamt | 18 |  |  |



## Measures to meet funding requirements and handle project results

Results of the project will, following the FAIR [14] guidelines, be made freely available under a (yet to be decided) open source license. In addition to the distribution of web platform, which is designed for long-term availability, results of this project will be shared by publications in highly reputed scientific journal, and presentations at international scientific meetings and workshops. Journals that allow free electronic access to published manuscripts (open access) will be chosen if possible. Furthermore, the project will use a Git-based archiving tool to archive well-defined states of the project code, like e.g. releases, according to the DFG recommendations for good scientific practice.

## Information on scientific and financial involvement of international cooperation partners

CUP collaborates with a number of scientific institutions including Stanford University (hPOP project), but without financial benefit or involvement.

# Bibliography

1. Cook C.E., Bergman M.T., Cochrane G., Apweiler R., Birney E. The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Res.

2017; doi:10.1093/nar/gkx1154.

1. Laizhong Cui, F. Richard Yu, and Qiao Yan, “When Big Data Meets Software-Defined Networking: SDN for Big Data and Big Data for SDN ”, IEEE Network, January/February 2016
2. EMC Digital Universe, “Driving data growth in HealthCare”,

http://www.emc.com/analystreport/digital-universe-healthcare-vertical-report-ar.pdf

4. Marx, Vivien, “The big challenges of big data”, Nature 2013/06/12/online, Nature Publishing Group, a division of Macmillan Publishers Limited. <http://dx.doi.org/10.1038/498255a10>,<https://www.nature.com/articles/498255a>

5. Intel, “[Transforming Healthcare IT Through Software-Defined Infrastructure”, printed in USA, 334041-001US](file://D:\\Work\\GWDG\\Proteomics\\DFG Proposal\\Transforming Healthcare IT Through Software-Defined Infrastructure\”, printed in USA, 334041-001US, https:\\www.intel.com\\content\\dam\\www\\public\\us\\en\\documents\\casestudies\\transforming-       healthcare-it-through-software-define-infrastructure.pdf)

6. Research and Markets, “[Global $432.2 Billion Software Defined Networking Market Analysis & Trends 2013-2017 & Industry Forecast to 2025 - Emergence of Hyper- Scale Cloud Networking”, Dublin, Feb23, 2017/PRNewswire,](file://D:\\Work\\GWDG\\Proteomics\\DFG Proposal\\Global $432.2 Billion Software Defined Networking Market Analysis & Trends 2013-2017 & Industry Forecast to 2025 - Emergence of Hyper-Scale Cloud Networking\”, Dublin, Feb23, 2017\\PRNewswire,     https:\\www.prnewswire.com\\news-releases\\global-4322-billion-software-defined-networking-Market-analysis--trends-2013-2017--industry-forecast-to-2025---emergence-of-hyper-scale-cloud-networking---research-and-markets-300412458.html)

[https://www.prnewswire.com/news-releases/global-4322-billion-software-defined- networking-Market-analysis--trends-2013-2017--industry-forecast-to-2025--- emergence-of-hyper-scale-cloud-networking---research-and-markets- 300412458.html](file://D:\\Work\\GWDG\\Proteomics\\DFG Proposal\\Global $432.2 Billion Software Defined Networking Market Analysis & Trends 2013-2017 & Industry Forecast to 2025 - Emergence of Hyper-Scale Cloud Networking\”, Dublin, Feb23, 2017\\PRNewswire,     https:\\www.prnewswire.com\\news-releases\\global-4322-billion-software-defined-networking-Market-analysis--trends-2013-2017--industry-forecast-to-2025---emergence-of-hyper-scale-cloud-networking---research-and-markets-300412458.html)

# 7. HIT Infrastructure, “Software-Defined Networking Advances Health IT Connectivity”,

# https://hitinfrastructure.com/news/software-defined-networking-advances-health-it-connectivity

8. <https://en.wikipedia.org/wiki/List_of_mass_spectrometry_software>

9. Salvador Martínez-Bartolomé, Eric W. Deutsch, Pierre-Alain Binz et al.,”Guidelines for reporting quantitative mass spectrometry based experiments in proteomics”, Journal of Proteomics, Volume 95,2013, Pages 84-88,ISSN 1874-3919, <https://doi.org/10.1016/j.jprot.2013.02.026>.

10. <http://www.highmed.org>

11. Nuagenetworks From Nokia, “Nuage Networks Provides Comprehensive Software- Defined Networking (SDN) Solution to Leading Health Provider UPMC”

12. <http://www.proteometools.org/>

13. <https://www.proteomicsdb.org/>

14. http://www.forschungsdaten.org/index.php/FAIR\_data\_principles

# Funding

# 4.1 Requested modules/funds

**4.1.1 Personal Costs**

In total we request funding for 180 PM for five Postdocs/PhDs. The distribution of the positions to the consortium partners is summarized in the following table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Funding for** | 2020 | | 2021 | | 2022 | |
| **Staff** | Quantity | Sum | Quantity | Sum | Quantity | Sum |
| **Prof. Dr. Yahyapour, Ramin (GWDG)** | | | | | | |
| Postdoc  100% | 1 | 72.000 | 1 | 72.000 | 1 | 72.000 |
| PhD  100% | 1 | 66.300 | 1 | 66.300 | 1 | 66.300 |
| **Prof. Dr. Reinhoff, Otto (DMI-UMG)** | | | | | | |
| PhD  100% | 1 | 66.300 | 1 | 66.300 | 1 | 66.300 |
| **Prof. Dr. Hassan Dihazi, Prof. Asif Abdul Rehman** | | | | | | |
| PhD  100% | 1 | 66.300 | 1 | 66.300 | 1 | 66.300 |
| **Prof. Dr. Simone König** | | | | | | |
| PhD  100% | 1 | 66.300 | 1 | 66.300 | 1 | 66.300 |
| Total | 5 | 337200 | 5 | 337200 | 5 | 337200 |

**4.1.2 Consumable for outreach and training**

250 Euro per year are requested for of information and education material which cannot be produced by project partners (e.g., information brochures, posters, material for public relations). In addition, the professional web-design and implementation of a distribution web platform to maximize distribution of ArCare will be done by CUP.

In total, we apply for 750 Euro of funding for consumables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Funding for** | 2020 | 2021 | 2022 |
| **Consumables** | 250 Euro | 250 Euro | 250 Euro |

**4.1.3 Planning of travel expenses**

In order to communicate ArCare and investigate cross-site requirements at different research locations travelling is necessary between Münster and Göttingen. Moreover, five other research sites will be visited. Furthermore, project results will be presented and discussed at national and international conferences of RDM and life science experts. The sum of 6.000 Euro for travel expenses is thus requested.

|  |  |  |  |
| --- | --- | --- | --- |
| **Travel expenses for** | 2020 | 2021 | 2022 |
| **GWDG** | 800 | 800 | 400 |
| **MI** | 600 | 400 | 200 |
| **NR/BIN, KC** | 200 | 800 | 800 |
| **CUP** | 400 | 400 | 200 |
| **Total** | **2.000** | **2.400** | **1600** |

**4.2 Financial contributions**

GWDG will contribute in the following assets:

* + Servers and storage to develop and operate the ArCare infrastructure during and after the project’s life time.
  + Personnel (Proportional) to operate servers, storage, and network.
  + Unfunded personnel (see 5.2) to support dissemination, networking, and outreach of the project.

The Partners contributes the following assets:

* + Training and support for ArCare after the lifetime of the project.

CUP contributes a fully equipped MS-based technology platform to this project including mass spectrometers and peripheric equipment. Technicians and scientists will assist and advise the PhD student hired for this project. CUP can also provide the support of an IT specialist who manages CUP data servers, IT environment and webpages.

# Project requirements

5.1 Employment status information

Permanently funded through Georg-August-Universität Göttingen:

* Prof. Dr. Yahyapour, Ramin, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
* Prof. Dr. Reinhoff, Otto, University Medical Center Göttingen

Permanently funded through University Medical Center

* Prof. Dr. Hassan Dihazi, Clinic for Nephrology and Rheumatology, UMG
* Prof. Dr. Abdul Rahman Asif, Institute of Clinical Chemistry / UMG-Laboratories

Prof. Dr. Simone König, tenured at the University of Münster

5.2 Composition of the project group

* Dr. Philipp Wieder (Deputy Head GWDG and leader eScience group; permanently funded through GWDG)
* Dr. Harald Kusch, (Senior Sceintist; funded through CRC 1002 and University Medical Center (permanently funded through Georg-August-Universität Göttingen)

**CUP:** The core facility was set up (since 1999) by and is headed by Prof. König, who has more than 20 years of practical experience in biomolecular MS and proteomics. Currently, it has two technicians (A. M. Berg, PTA; D. Ackermann, Dipl-Ing. (FH)), one scientist position (M. Bayer, PhD student) and a varying number of master students and applicants for medical doctorate. In addition, the group is supported by Dipl.-Inf. B. Teichert who manages the hard- and software environment as well as data storage and backup.

**5.3 Cooperation with other researchers**

### 5.3.1 Institutions or Researchers with whom you have agreed to cooperate on this project

Letter of intent of the following institutions are attached to this proposal:

* Goettingen Proteomics Forum (GPF)

**5.3.2 Researchers with whom you have collaborated scientifically within the past three years**

**5.3.2.1 GWDG**

* Ben-Gurion University of the Negev, Beer Sheva, Israel
* Cloudius Systems, Herzliyya, Israel
* Datenservicezentrum Betriebs- und Organisationsdaten, Universität Bielefeld (DSZ-BO)
* DLR Institute of Aerodynamics and Flow Technology
* European Grid Infrastructure (EGI)
* German National Library, Frankfurt/Main
* GIGA – German Institute of Global and Area Studies, Hamburg
* Huawei Technologies Düsseldorf GmbH
* IBM Israel –Science and Technology LTD, Petach Tikwa, Israel
* Institut für Sozialwissenschaftliche Forschung München e.V. (ISF München)
* Instiute of Sociology, Friedrich-Schiller-Universität Jena
* Intel Shannon Limited, Leixlip Ireland
* Interoute S.p.A, Rom und Nextworks Srl, Pisa, Italien
* LIPI-MZB Indonesian Institute of Science/Research Center for Biology
* Mellanox Technologies, Israel
* National Technical University of Athens, Greece
* Pipistrel Podjetje Za Alternativno Letalistvo D.O.O., Ajdovscina, Slovenia
* Seagate Technology LLC, Dublin, Ireland
* TMF – Technologie-& Methodenplattform für die vernetzte medizinische Forschung e.V.
* TU Braunschweig, Institute for Theoretical Physics
* TU München, Freising-Weihenstephan
* Universität Stuttgart, High Performance Computing Center Stuttgart
* Universitätsmedizin Greifswald, Institute for Community Medicine
* University of Patras, Greece

XLAB razvoj programske opreme in svetovanje d.o.o., Ljubljana, Slovenien

**5.3.2.2 NR/BIN, KC**

* Ahmed Mansouri: MPI for Biophysics Göttingen.
* Prof. Dr. Oliver Gross: Department of Nephrology UMG.
* Prof. Dr**.** Ali El-Armouche: Technische Universität Dresden Medizinische Fakultät Institut für Pharmakologie und Toxikologie.
* Dr. Gabriela Salinas-Riester director of Transcriptomics Facility at the Medical Center Göttingen
* PD Dr. Olaf Jahn MPI Experimental medicine Göttingen. Director of Proteomics Facility
* Prof. Dr. Visith Thongboonkerd MDMedical Molecular Biology Unit, Bangkok, Thailand.
* Prof. Dr. Michael Zeisberg: Department of Nephrology and Rheumatology, UMG
* Prof. Dr. Michael Goligorsky, Departments of Medicine, Pharmacology and Physiology, Renal Research Institute, New York Medical College at Touro University, Valhalla, NY, USA
* Prof. Silvio Rizzoli Department of Neuro- und Sinnesphysiologie, UMG
* Dr. Samy Hakroush Department of Pathology, UMG
* Dr. Holger Herlyn, University of Mainz, Mainz
* Prof. Bernhard Reuss, University medical center Göttingen

**5.3.2.3 MI**

**5.3.2.4 CUP**

University Clinic Mainz, University of Cape Town, Stanford University, many groups locally

**5.4 Formal assurance**

It is intended to provide the developed software suite as open source software. This includes comprehensive documentation (for further details see sections 2.3 and 2.4)

**5.5 Project-relevant cooperation with commercial enterprises**

None

## 5.6 Project-relevant participation in commercial enterprises

None

# 6 Additional information

**6.1 Abbreviations**

|  |  |
| --- | --- |
| GPF | Göttingen Proteomics Forum |
| GWDG | Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen |
| EBI | European Bioinformatics Institute |
| SDN | Software Defined Networking |
| HUPO-PSI | HUPO Proteomics Standard Initiative |
| UMG | University Medical Center Göttingen |
| VSP | Network Virtualized Service Platform |
| MIAPE | The Minimum Information about a Proteomics Experiment |