Project Description – Project Proposals

[First name last name, city of all applicants]
[Research position (e.g. research assistant, professor)]
[Term of contract (fixed-term, permanent)]

Project Description

1 State of the art and preliminary work

Omic sciences (omics), also referred to as high-dimensional biology, are a relatively new field of research in life sciences aiming at the study of structure, functions and dynamics of living organisms. Information technology is an integral part of this research, given the analysis of colossal amount of data spawning by the research instruments is vital. Mass spectrometer, for instance, being one of the most multipurpose analytical instrument has been extensively used in proteomics (also metabolomics) for the identification of analytes. In a state of the art proteomics lab with several Mass spectrometer instruments, hundreds of data files are produced and analyzed daily, whereas, the amount of data carried out by each single file is of several GBs. According to some recent statistics, the data store size of one of the world's largest Bioinformatics lab, European Bioinformatics Institute (EBI) in UK, reaches to 20 petabytes.

The vendor lock-in applications and data formats is a known issue in mass spectrometry based proteomics research. There is a broad range of mass spectrometer vendors (about 24) available in the market 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 keeps on extending them with new features as required by emerging instrumentation. The native data format of each vendor instrument differs from others and, therefore, non-interchangeable too. Consequently, only the vendor proprietary applications can be primarily used to read and analyze the vendor specific 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 in past to deal with this issue and a variety of vendor-neutral formats (also called open formats) and open-source analysis tools have been developed by various research groups working in the domain to enhance the sharing of results and data analytics among the researchers in the community. HUPO Proteomics Standard Initiative (HUPO-PSI), for instance, is one of the most active community in Proteomics working for the development of standardized data representations for data exchange and verification based on vendor APIs.

Reproducibility is another known issue in proteomics that researchers have to deal with. The data analysis studies in mass spectrometery is a complex multistep operation, which involves collecting the raw data from instruments, extraction of spectral data to one or more open data-formats, submitting the data into shared repositories, and analysis of data with some available software tools. The lack of automation of this high-throughput workflow, therefore, questions the accuracy, speed, consistency, and transparency of data processing outcome. Therefore, in a well-equipped research facility that operates a range of mass spectrometery instruments and multiple data analysis programs, it becomes extremely challenging to maintain reproducibility without automating the highly sophisticated SOPs (standard operating procedures) of data analysis.

Essentially, as mentioned earlier, mass spectometery is a key multidisciplinary analytical technique commonly used under the domain of omics. To setup a centralized research infrastructure for various local and distributed mass spectometery labs is therefore vital. It should offer:

- Secure and highly scalable network of the researcher instruments.
- Storage of data into a common/public data repository.
- Compliance to the departmental SOPs for data processing and analytics.
- Centralized policy management and control to help enforce disciplinary regulations and policies in order to gain proper control and prevent uncontrolled distribution.

1.1 Current practices and challenges at Goettingen Proteomics Forum (GPF)

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 analysis of proteins by mass spectrometry. The ambition of the forum is the pooling of local proteomics expertise and tools to generate synergism and provide mutual support for the proteomics community in Göttingen.

At present, different working groups under GPF are running their mass spectometery labs. Table 1 lists down a range of mass spectometery instrumentations and software these state-of-the-art labs possess for experimentation. Based on its vendor lock-in technology and solutions, each lab follows a set of operations (i.e., SOP) for raw data collection, data conversions to meaningful formats, analysis and storage of data for research purposes. Whereas, GWDG being a common data processing center for GPF, supports the proteomics labs in the data analysis part. For this, GWDG hosts a proteomics analysis software, MASCOT server, which has become a standard for protein identification using mass spectometery data for many big vendors.

Apart from the above mentioned high-tech research facilities at GPF, all the processes and operations at proteomics labs and data processing cooperation currently following a strict manual routine, hence prone to reproducibility, productivity and transparency issues. Also in its current setup, the existing infrastructure supports no common and public research data repositories for an open exchange of data across different research settings. Furthermore, the lack of common workflow practices among the labs and non-adherence to open data standards may also hinder a systematic and spontaneous sharing of results among the settings.

Institute	Vendor MS Devices	MS Software

UMG (Institute of Clinical Chemistry)	• Thermo Scientific Q Exactive • Q-TOF Ultima Global mass spectrometer (Micromass) • MALDI MicroMX mass spectrometer (Micromass)	✓ Mascot (Matrixscience) ✓ Scaffold (Proteome Software) ✓ MaxQuant with Perseus.
MPI of Experimental Medicine (Proteomics Group)	• Ultraflex I (Bruker) • Synapt GS 2(Water) • Acquity-QDa(Waters)	Details to be added
MPI Biophysical Chemistry	Orbitrap instrumentation + LTQ Orbitrap XL + Q Exactive Plus + Q Exactive HF + Tribrid Fusion + Tribrid Lumos Triple quadrupole instrumentation • TSQ Vantage • TSQ Quantiva	Details to be added

Table 1: A list of Mass Spectometery instruments available at GPF

1.2 Available Tools and Technologies for the development of ArCare

ArCare proposes a secure, robust and automated research platform for proteomics labs, which offers:

- The automation of different mass spectometery operations (from data collection till data storage and analysis) currently followed in proteomics labs.
- The support and automation of standardized workflows (defined by HUPO-PSI and other active communities) based on open data formats and open source tools and software.
- Integration of data into existing transregional repositories, such as NFDI (Nationalen Forschungsdateninfrastruktur), for a sustainable and open exchange of data across different research settings.

Figure 1 depicts a high level design of ArCare solution based on the existing off-the-shelf tools and technological testbeds developed at the Goettingen campus. The heart of the solution is the SDN (Software Defined networking) capabilities to improvise a sophisticated control and management plane to increase efficiency of big data applications, such as proteomics in this case. Figure 2, emphasizes some of the benefits for running big data application over SDN network.

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 OpenDaylight controller, which allows the 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 SDN control plane and non-SDN data plane. For instance, as depicted in Figure 1, in order to provision a centralized control and automation of mass spectometery big data labs, at first level we need to utilize the strength of SDN agents to facilitate the communication between the ArCare controller and the network switches. At the second level, we need to introduce device agents to collect and transfer the mass spectometery data from the vendor devices.

MS Proteowizard: details need to be added

MS Applications: details need to be added

OpenMS: details need to be added

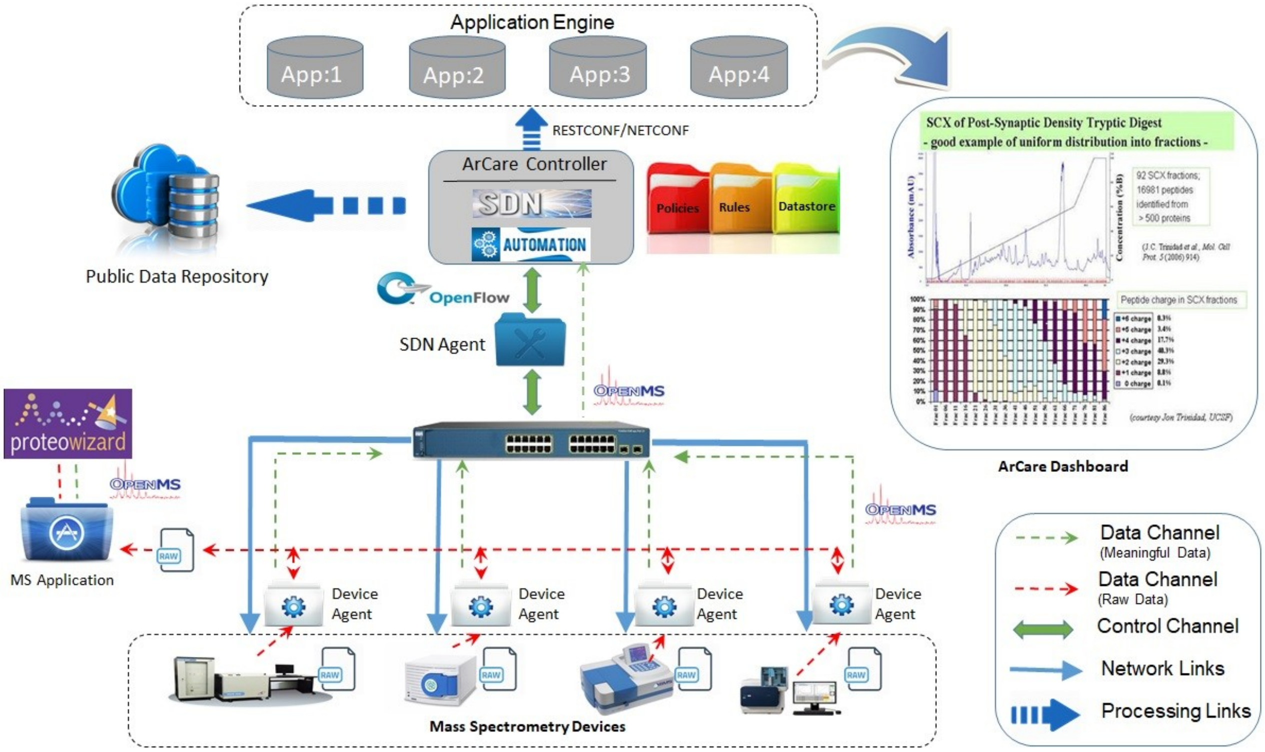

Figure 1: Overview of ArCare solution

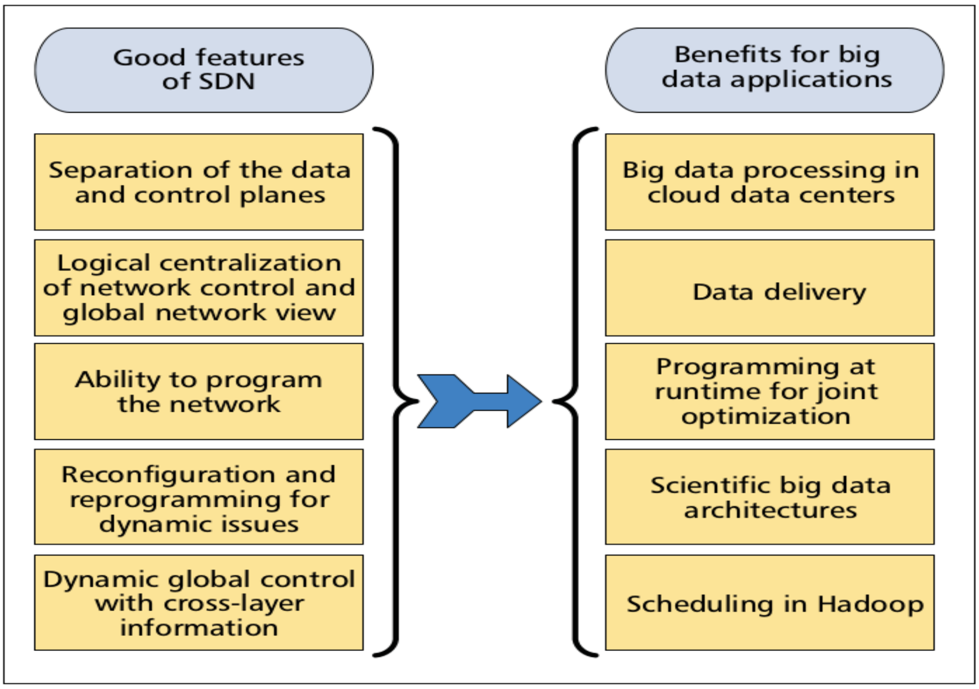
Good features of SDN
Separation of the data and control planes
Logical centralization of network control and global network view
Ability to program the network
Reconfiguration and reprogramming for dynamic issues
Dynamic global control with cross-layer information
Benefits for big data applications
Big data processing in cloud data centers
Data delivery
Programming at runtime for joint optimization
Scientific big data architectures
Scheduling in Hadoop

Figure 2: SDN capabilities to tackle big data applications

1.3 Comparison to already available alternative solutions

Scientific research is extremely data driven. Many scientific research devices produce tremendous sum of data. Furthermore, considering political and technical aspects, some scientific big data cannot be processed and analyzed at the identical placement where they are created. The big data should be sent and stored in different data centers and researchers at different universities or institutes for analysis. Nonetheless, today campus networks are not able to manage such a vast amount of data. As the amount of data increases, there is a greater need for unsophisticated, scalable end-to-end network architectures and implementations that enable applications to use the network most efficiently [1].

In the area of Healthcare data, as one of the straight example of big data, is increasing to a impressive rate— over 40% yearly. By 2020, Electronic Healthcare Record (EHR) is anticipated to need over 2000 exabytes—that's two billion terrabytes. [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 f 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].

Unfortunately, healthcare IT spending is not only unsuccessful to preserve with this rate of growth, but is awaited to fall. And today's data center infrastructures situation 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. Healthcare IT decision makers requires new opportunities to scale quicker and more cost effectively [4].

Healthcare organizations are using virtualization to integrate health IT infrastructure solutions to achieve more easily manage processes and overcome the pressure on network hardware. 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 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 healthcare and Bioinformatics are one of the best context to deploy SDN technology. As healthcare organizations reconcile to electronic health records (EHRs) and cloud computing, networks require to be capable of manage the increased amount of traffic [6]. A Software-Defined Infrastructure can assist healthcare facilities remove challenges and recognize fresh possibilities by reorganizing legacy networks, as shown above, into a more flexible SDN. As an outcome, IT can provide application-aware network, gain agility and flexibility for applications and services, and optimize cycle management [7].

1.1 Project-related publications

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

[Text]

1.1.2 Other publications

[Text]

1.1.1 Patents

1.1.2.1 Pending

[Text]

1.1.2.2 Issued

[Text]

2 Objectives and work programme

2.1 Anticipated total duration of the project

[Text]

2.2 Objectives

[Text]

2.3 Work programme incl. proposed research methods

2.3.1 WP1 - Formation of SDN Infrastructure and Integration with Proteomics Lab

Work Package 1 Overview			
WP Lead			
Partner			
PM pro Partner			
PM Gesamt			

2.3.2 WP2 - Automation and Reproducibility

Work Package 2 Overview			
WP Lead			
Partner			
PM pro Partner			
PM Gesamt			

2.3.3 WP3 - Integration with Data Repository

Work Package 3 Overview			
WP Lead			
Partner			
PM pro Partner			
PM Gesamt			

2.3.4 WP4 - Data Analytics Portal

Work Package 4 Overview			
WP Lead			
Partner			

PM pro Partner			
PM Gesamt			

2.3.5 WP5 - Project Coordination

Work Package 5 Overview			
WP Lead	GWDG		
Partner			
PM pro Partner			
PM Gesamt			

Project Plan (months)	**1-6**	**7-12**	**13-18**	**19-24**	**25-30**
WP1:					
WP2:					
WP3:					
WP4:					

[Text]

2.4 Data handling

[Text]

2.5 Other information

Please use this section for any additional information you feel is relevant which has not been provided elsewhere.

[Text]

2.6 Descriptions of proposed investigations involving experiments on humans, human materials or animals

[Text]

2.7 Information on scientific and financial involvement of international cooperation partners

[Text]

3 Bibliography

[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]http://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pd

[3] https://www.nature.com/articles/498255a

[4] https://www.intel.com/content/dam/www/public/us/en/documents/case-studies/transforming-healthcare-it-through-software-define-infrastructure.pdf

[5] 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

[6]https://hitinfrastructure.com/news/software-defined-networking-advances-health-it-connectivity

[7] https://www.intel.com/content/dam/www/public/us/en/documents/case-studies/transforming-healthcare-it-through-software-define-infrastructure.pdf

4 Requested modules/funds

Explain each item for each applicant (stating last name, first name). Follow the outline given in the relevant programme and module guidelines.

[Text]

5 Project requirements

5.1 Employment status information

For each applicant, state the last name, first name, and employment status (including duration of contract and funding body, if on a fixed-term contract).

[Text]

5.2 First-time proposal data

Only if applicable: Last name, first name of first-time applicant.

[Text]

5.3 Composition of the project group

List only those individuals who will work on the project but will not be paid out of the project funds. State each person's name, academic title, employment status, and type of funding.

[Text]

5.4 Cooperation with other researchers

5.4.1 Researchers with whom you have agreed to cooperate on this project

[Text]

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

[Text]

5.5 Scientific equipment

List larger instruments that will be available to you for the project. These may include large computer facilities if computing capacity will be needed.

[Text]

5.6 Project-relevant cooperation with commercial enterprises

If applicable, please note the EU guidelines on state aid or contact your research institution in this regard.

[Text]

5.7 Project-relevant participation in commercial enterprises

Information on connections between the project and the production branch of the enterprise.

[Text]

6 Additional information

If applicable, please list proposals requesting major instrumentation and/or those previously submitted to a third party here.

[Text]