COMMIT

PROJECTPLAN

WORKPACKAGES

DELIVERABLES

BUDGET

E-BIOBANKING WITH IMAGING FOR HEALTHCARE (P24)

Projectleider Prof.dr. Joost Kok, Leids Universitair Medisch Centrum

1. Background

Research in e-Science over the last years has provided us with new concepts, an e-Science ICT research infrastructure for the basis sciences and a variety of new tools. The benefits of e-Science research are evident, particularly for research areas in the basic sciences. The time is now ripe to apply the benefits of e-Science research beyond research prototypes.

The nature of life sciences research has been affected by the development of automated analytical instrumentation, which now enables large volumes of data to be generated, collected in databases and analyzed with computer programs. New high throughput screening capabilities, improved spectral or spatial resolution and new physicochemical parameters that can be monitored, have resulted in an almost explosive growth in data generation. Scientists will greatly benefit when all present-day bioinformatics knowledge related to their research will be combined in a system that simplifies mutation analysis and prioritizes candidate genes for their relevance in the genotype-phenotype relation. The bioinformatics methods harvest and integrate the data from a large series of existing systems including, but not limited to: genome annotation, proteome annotation, known SNPs and haplotypes, gene-disease association, metabolic networks, protein-protein interaction networks, splice variants, transcription factors/transcription factor binding sites, gene expression databases, literature, molecular sequence, structure, and additional information systems. For human sequence data legal aspects are also involved.

Multidisciplinary systems approaches are becoming much more prevalent in many sciences, particularly so in the life sciences. Often, life science researchers work together based on common collections of data. Typically, collaborations face issues concerning data storage and distribution of the data among the collaborators, e.g. the mass spectrometry experimental datasets grow very large, making them difficult to handle. Various initiatives have been started to deal with the problems concerning storage and sharing of experiment data. Several repositories and databases for storing and sharing measurement data, mass spectral fingerprints, peak lists, and proteomics, genomics, metabolomics, and lipidomics information have been created; e.g. Pride, Pedro, Metlin , Expasy, Mascot, BioCyc, LipidMaps, Lipid Bank, Golm metabolome database, PeptideAtlas, Global Proteome Machine Humane Protein Reference Database and the Tranche project.

Consortia are established to perform large cohort studies. Much effort is devoted to the careful collection of data, annotating the data and storing it in a secure way. It is now common for life-science researchers from different backgrounds and disciplines to work together – or in other words it is now widely recognized that the complexity of many

current scientific problems in the life sciences has outgrown the grasp of any single discipline illustrating that system level sciences has also entered this field. For example, a present-day research program might attempt to analyze linked data from in-vivo imaging, proteomics and metabolomics approaches.

The knowledge and resources of a single research group are no longer sufficient to utilize the latest developments. Consequently, the life sciences are increasingly collaborative where each collaborator contributes his or her specific expertise and/or data to the project. These collaborations are heterogeneous in nature and often widely distributed geographically. Currently it takes a lot of effort, persistence and good faith to make a collaboration work and achieve synergy. Multidisciplinary scientific collaborations will benefit from an e-Science platform by allowing them to interact efficiently and effectively. Such a platform practically facilitates the collaboration by having access to shared data, exchanging the knowledge and methods among the collaborators.

A successful multidisciplinary scientific project depends on four key ingredients: 1) high quality people, 2) efficient interaction between the people, 3) high quality experiment design and infrastructure, and 4) environments that facilitate data interpretation. In this project we intend to translate the developments in e-Science research to the life-science domain where they are integrated into existing experimental infrastructure to facilitate their application, and builds onto the existing VL-e fundaments. The focus in this project will be on e-biobanks. A biobank is a, for scientific purposes, assembled collection of (images of) human materials, coupled with medical and/or genetic and/or genealogical and/or other information about the donors. e-Biobanks are e-Science environments that facilitate biobank data interpretation, in particular for population-derived data stemming from different researchers and different modalities facilitating a population-based systems approach. It is the intention to work at an abstraction level that is domain-generic: it identifies common approaches multiple life science applications and enhances these approaches using the e-Science approach. A special focus is on imaging.

The partners in the project originate from academic medical centers, universities and knowledge institutions that are all stakeholder in the domain specific developments in e-Science. Linking e-Science expertise with real problems in the life sciences will result in an enhanced infrastructure that will economically benefit both industry and education in the Netherlands through enhanced efficiency in research and training.

2. Problem description

In the life sciences domain the adoption of high performance computational (HPC) and high performance storage (HPS) infrastructures, such as grids, will become essential to utilize the latest developments in the analytical techniques and to integrate the data in a multidisciplinary approach. It will be necessary to cope with the increasing complexity, variety, size and sources of data in the near future. Although grids are becoming available on a production basis (e.g., BIGGrid, EGEE), their exploitation for Science (=e-Science) remains currently limited to domains with a long tradition in HPC and HPS and large application development and support teams, such as in Physics.

Currently the front-end interfaces to grid-computing infrastructures are unsuitable for most life-science researchers, mostly due to an all but "user-friendly" interface. Life-sciences can only benefit from the research in e-Science when the life-scientists can directly, autonomously, and comfortably exploring the potential of these infrastructures. Molecular imaging with fMRI, molecular histology with mass spectrometry, high throughput drug discovery and high throughput genome and proteome research are key examples where in the life sciences these data-driven access problems are encountered, and where e-biobanking can provide badly needed solutions.

Life sciences are increasingly multidisciplinary and collaborative. Scientists work together in genomic, proteomic and metabolomic collaborations of all shapes and sizes to take on contemporary challenges for the biomedical sciences. For example consider scientists working on breast cancer research. Collaboration A) focusing on a mouse model of breast cancer can include a pathologist, a biochemist, a magnetic resonance imaging group and a proteomics group concentrating on imaging mass spectrometry. Collaboration B) focusing on human biopsy material can include a pathologist, a clinical oncologist, a tissue microarray group and a proteomics group including imaging mass spectrometry. In both cases different multi-modal imaging capabilities are combined. To work effectively the four groups from both collaborations, located at various facilities, work together within a collaboration on different parts of the same problem. To solve the problem they need to be able to share their data and knowledge. We need to rethink the way the data is stored. Current methods need to adapted and new methods need to be developed to efficiently deal with such large and heterogeneous data collections (annotation, linking, fusion, visualization, analysis, data and knowledge management, knowledge discovery). The best representation for *storing* data is not necessarily the best representation for *mining* them.

e-Biobanking approaches will be required for the integration of all of these data and for generating knowledge from the data. Analytical life sciences and pharmaceutical sciences are the key areas driving e-Biobanking. The evaluation, selection and deployment of generic frameworks for e-Biobanking services.(e.g. services that manage access to computational and data resources to services that deal with knowledge) are a pre-requisite for successful research activities in these translational research domains. A new generation of life scientists needs to be familiarized with the research results to ensure embedding of e-science. This requires a dedicated effort to make available the Dutch e-science research infrastructure for teaching and training in remote experimentation.

The core issues that are addressed in "e-Biobanking with Imaging for Healthcare" are related to data management, and distributed storage, access, analysis and mining of large scale life science data sets. The seamless access to the national distributed GRID infrastructure for the life sciences with a focus on imaging is a specific area that will be addressed.

3. Objectives

Project's goal

Developments in life sciences research have resulted in an almost explosive growth in data generation. The main problem with this data explosion is not so much the storage of the data; it is the efficient access. Moreover, the data becomes only valuable in the context of other data and knowledge. Currently no integrated systems exist that brings the ICT developments in life-science applications using e-Biobanking tools that can handle the extremely large datasets, complex computational requirements, information management, and advanced collaboration demands all necessary to open the high potential of multimodal data, including imaging, to clinical studies. The research performed in this project will help to put the ingredients together in an e-BioBanks platform in which images are first-class citizens, at service for life scientists (end-users) to generate new knowledge that will translate into better healthcare.

Planning of all dimensions

Overall objective of the project "e-Biobanking with Imaging for Healthcare" is to provide e-science instrumentation for BioBanks:

- Cooperative management, visualization, annotation, interpretation and enrichment of large collections of heterogeneous BioBank data at possible disparate locations,

- Tools for data-intensive scientific discovery, in particular analysis and mining across levels, and
- Support of end-users, to enable scientific discovery for healthcare.

The project is divided in three subprojects: 1) Molecular Histology, 2) Image Instrumentation, and 3) Integrated BioBanks, focusing on different aspects, but all working towards the overall objective. The subprojects are clusters of work packages. All e-Science tools will be employed in at least two of the three subprojects.
We want to establish an e-Biobank application platform that generates a high degree of transparency and dissemination within different collaborations based on shared data in the life sciences. As such it facilitates shared usage of data and it will add value to collaborations by combining all their aspects, going beyond existing data- and resource sharing platforms, in particular by also facilitating knowledge discovery.

We target to show the pragmatism and versatility of such a platform for molecular data including images in healthcare research from efficient organization of presentation logistics to in-depth scientific analysis of – and knowledge discovery in experimental data. The project should result in a e-Biobank application and analysis platform that is accessible for healthcare researchers in the Netherlands.

Results

- Subproject title: Molecular histology
  Approach: Application of molecular imaging using mass spectrometry in molecular histology.
  Goals: Creation of a reference biobank, discovery of biomarkers in tissue.
  Focus Data: Mass spectrometry, MRI.
  Biobank: Virtual Tissue Bank.
  E-Science Tools: Knowledge management, Annotation, Virtual Microscopy.
  Healthcare: Analysis of Tissue (Breast Cancer Tissue, Soft-tissue Sarcoma)

- Subproject title: Image Instrumentation
  Approach: Develop instrumentation for analysis of (medical) images.
  Goal: Analysis of (medical) images by end users. Discovery of patterns across biological levels.
  Focus Data: MRI, fMRI
  e-Science Tools: Front-end for e-infra, Functional Data Analysis
  BioBank: ID1000 study.

Healthcare: Analysis of Medical Images (Neuroimaging, Mammography)

- Subproject title: Integrated Biobanks
  Approach: Develop instrumentation for analysis and mining across different types of data.
  Goal: Discovery of subtypes and biomarkers in diseases using a systems approach.
  Focus Data: Integral data in Cohort Studies and Population Studies.
  E-Science Tools: e-infra for distributed data, secure data access, databases for query dominant applications, data mining algorithms, interoperability
  Biobank: IBD Biobank (Parelsnoer), BBMRI, Carema, Leiden 85+
  Healthcare: Inflammatory Bowel Disease, Coronary heart disease, healthy ageing

Deliverable Impact and Valorization

The development of an e-biobank platform with imaging will have a tremendous impact on healthcare research, which requires integration of data from multiple disciplines, integration of data from multiple sources, collection and analysis of large amounts of data, data collected at multiple locations (hospitals, primary healthcare organizations, homecare) and data exchange/sharing. Key ingredients for success in this field already exist in The Netherlands: 1) availability of data; 2) availability of methods; 3) availability of molecular imaging instrumentation; and 4) availability of e-Science infrastructure to cope with the scale of data, experiments and the collaborative aspect of research. The research performed in this project will help harness this exceptional basis by adopting e-Science expertise and techniques to put all the ingredients together at service of the researchers (end-users) to generate new knowledge that will translate into better healthcare. The participation of medical centers and a public/private partnership in the evaluation and establishment of several e-Biobanking frameworks will improve the (inter-)national competitive position of industrial and academic research. The dissemination of e-Biobanking and e-science research results to academic training programs will reinforce the already existing strength knowledge potential and workforce in the Netherlands.

Deliverable Dissemination

In addition to regular scientific publications, conference contribution and remote experimentation demonstrators we intend to develop and employ e-Biobanking with imaging environment as a teaching tool in the academic curricula in this domain. This will allow comprehensible student access to a unique infrastructure and will allow direct transfer of knowledge into the educational system.

International Embedding

Currently no integrated systems exist that brings the ICT developments in healthcare applications using e-Biobanking. Tools that can handle the extremely large datasets, complex computational requirements, information management, and advanced collaboration demands are all necessary to open the high potential of multimodal data, including imaging, to clinical studies. When completed the results from this e-Biobanking research would significantly increase the possibilities in healthcare research, pharmaceutical sciences, drug pipe-lines and basic chemical sciences.

Deliverable Synergy

The project team will work together with KNOWLEDGE MANAGEMENT IN SCIENCE and E-SCIENCE IN AGRIFOOD on information and knowledge management and with E-INFRASTRUCTURE VIRTUALIZATION FOR E-SCIENCE APPLICATIONS for the high performance computing needs. For data management we will work together with SPATIOTEMPORAL DATA WAREHOUSES FOR TRAVELERS

4. Economic and social relevance

The development of an e-biobank platform with imaging will have a tremendous impact on healthcare research, which requires integration of data from multiple disciplines, integration of data from multiple sources, collection and analysis of large amounts of data, data collected at multiple locations (hospitals, primary healthcare organizations, homecare) and data exchange/sharing. Key ingredients for success in this field already exist in The Netherlands: 1) availability of data; 2) availability of methods; 3) availability of molecular imaging instrumentation; and 4) availability of e-Science infrastructure to cope with the scale of data, experiments and the collaborative aspect of research. The research performed in this project will help harness this exceptional basis by adopting e-Science expertise and techniques to put all the ingredients together at service of the researchers (end-users) to generate new knowledge that will translate into better healthcare. The participation of medical centers and a public/private partnership in the evaluation and establishment of several e-biobanking frameworks will improve the (inter-)national competitive position of industrial and academic research. The dissemination of e-biobanking and e-science research results to academic training programs will reinforce the already existing strength knowledge potential and workforce in the Netherlands.

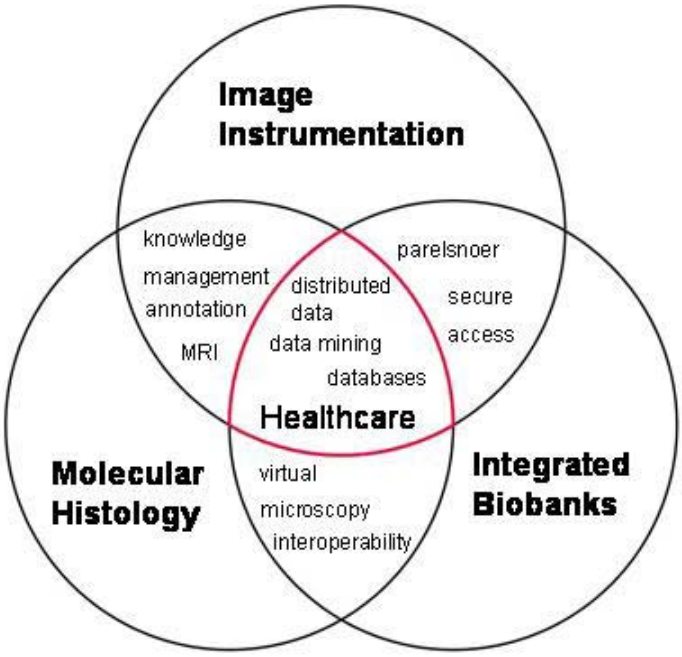
5. Consortium

The consortium consists of Amolf, LUMC, UvA, AMC, UMCN/MSD and TI Coast. It is a purposely designed mix of university hospitals, a university, a research institute and a

public/private partnership. This mix is needed to deal with the diverse e-biobanking aspects within imaging, analytical sciences and pharmaceutical research at the highest level: one needs to have the technology, the medical data and industrial drivers to render e-Biobanking environments that are really used. It also ensures the availability of qualified personnel. The project team will work together with KNOWLEDGE MANAGEMENT IN SCIENCE and E-SCIENCE IN AGRIFOOD on information and knowledge management and with E-INFRASTRUCTURE VIRTUALIZATION FOR E-SCIENCE APPLICATIONS for the high performance computing needs. For data management we will work together with SPATIOTEMPORAL DATA WAREHOUSES FOR TRAVELERS and The e-Science aspects of Biobanking will be addressed in close collaboration with the Netherlands Bioinformatics Center (NBIC). The collaboration with NBIC will ensure the necessary bioinformatics support. Collaboration with the Cyttron (imaging), Parelsnoer and BBMRI (biobanks) projects is foreseen.

6. Workplan

This project is split up in three subprojects, each consisting of work packets that cover different aspects of e-Biobanking with imaging for healthcare. The number of work packets is nine. The overall structure is as follows:



The disclosure of distributed data resources and the corresponding metadata will be realized in close collaboration with projects KNOWLEDGE MANAGEMENT IN SCIENCE and E-SCIENCE IN AGRIFOOD. The data reduction and knowledge management will employ the

high performance computational resources as discussed in project E-INFRASTRUCTURE VIRTUALIZATION FOR E-SCIENCE APPLICATIONS. Clear links with the visualization research program are made within all work packets as the e-biobank platform will be a web-based resource.

WORKPACKAGES

Project number P24	
WP title&acronym	e-BioCognition
WP leader	Sennay, Ghebreab, UvA

Objectives

Floods of data are being generated from the cognitive to the molecular levels with the aim of advancing our understanding of the human brain and human behavior. Yet, the data have hardly paved the way to new breakthroughs crossing relevant levels of biological organization. One reason for this is the absence of tools that allow researchers to collectively analyze neural, physiological and behavioral data. The objective of this work package is to develop a data-analytic framework that will make it possible to identify common patterns in a bank of noisy biosignals (neural, physiological and behavioral). Central in this framework is Functional Data Analysis (FDA). With FDA we will represent biosignals as curves and analyze them collectively in search of modality independent patterns. As FDA transforms all data into a uniform format it also facilitates the integration of data from geographically distributed sites. We will participate in the ID1000 experiment of Prof. Lamme and dr. Scholte (Brain and Cognition, UvA) in which 1000 subjects will be monitored while performing cognitive tasks such as, solving mathematical questions and watching movies. Monitoring includes acquisition of MRI, fMRI, heartbeat, and blood pressure. Blood samples will also be taken from each subject for DNA analysis at a later stage. We will apply FDA to these data to uncover meaningful patterns across biological levels. The tools will be deployed on top of the VL-e e-science fundament to solve the computational complexity of data analysis.

Project number P24	
WP title & acronym	Molecular imaging and knowledge management for molecular histology
WP leader	Ron Heeren, Amolf

Objectives

The aim of this work packet is to establish an environment to process and evaluate large datasets originating from molecular histology using imaging mass spectrometry. A concise statistical analysis for the extraction of biomarker patterns from the molecular images is targeted to enable an innovative diagnosis approach in a biomedical imaging environment. Protocols will be established and employed to generated detailed mass spectrometry based molecular image datasets. The extracted molecular images together with their corresponding descriptive metadata will be stored in the Netherlands Virtual Tissue Bank. In close collaboration with TI-COAST we will establish a workflow based knowledge management system that will build upon the KnowEx system developed in VL-e. Distributed processing of the large image datasets is considered using the IBIS cockpit system developed at the VU computer science group. The applications will be tested on breast cancer sections collected from the medical partners in this project. The virtual tissue bank will be filled with the molecular mass spectrometric images generated in the framework of this program from these sections to act as a reference database for medical professionals. The tissue microarray data of these diseases will be linked with the histological images when available.

Project number P24	
WP title & acronym	Front-end for Biomedical Data Analysis on e-infrastructures
WP leader	Sílvia, Delgado Olabarriaga, AMC

Objectives:

The objective in this WP is to research, design, develop and evaluate state-of-the-art virtual environments towards an e-infrastructure for biomedical research involving large data collections or complex/long data analysis. Using these environments it will be easier and more efficient for end-users (clinicians, radiologists, medical physicists, neuroscientists, bio-informaticians) to perform collaborative data analysis and management on research infrastructures such as grids, clouds, desktop grids and advanced instruments. In addition, collaboration, sharing and reuse of methodology and data will be supported and stimulated by these environments. The target virtual environments will provide improved (generic) front-ends to grid infrastructures that (1) are friendly to end-users, adopting modern interactive graphical interfaces and web protocols; (2) take into account realistic usage scenarios of selected biomedical research applications, in particular imaging; (3) are flexible to operate with heterogeneous grid middleware and distributed infrastructures; and (4) can be extended and tailored to specific needs of other life science applications .

Project number P24	
WP title & acronym	Multiplex imaging of tissues
WP leader	Liam McDonnell, LUMC

Objectives

The maturation and automation of histo-chemical techniques, in which antibodies are used to investigate the expression and distribution of specific protein biomarkers across many patient biopsies, has led to the establishment of databases that contain the distribution of biomarker proteins for many pathologies. The analysis of patient biopsies with antibodies for validated biomarkers are now established tools for patient diagnosis and prognosis, and remains one of the key technologies for validating new candidate biomarkers.

Mass spectrometry based analysis of proteins now allows the quantitative analysis of the protein content of tissue and body fluids, simultaneously determining the levels of thousands of proteins, including protein isoforms. Many studies have now established how changes in the total protein profile can be more effective 'biomarkers' because it provides a more complete representation of the biochemical system. Clinical tissue analyses would benefit from such a parallel analysis of multiple proteins, distinguishing protein isoforms and relative protein stoichiometry.

We have shown how mass-spectrometry based methods can be directly applied to tissue to measure its molecular composition and protein distributions. Imaging mass spectrometry has rapidly progressed, current technologies allow the parallel analysis of 100's of proteins with high sensitivity and selectivity. A low resolution (250 μm) imaging mass spectrometry analysis of a small tissue microarray has demonstrated how the protein signatures from each tissue biopsy can be used to classify their pathological state. In this case the 'biomarkers' are based on the interplay of multiple proteins.

In this work package we will develop new tools that for imaging mass spectrometry analysis of clinical tissues to address some of the cutting-edge topics in modern cancer research, namely the interaction between tumor cells and neighboring non-tumor cells (tumor interface zones). Grid based computing will enable the large datasets from these analyses (30-100 Gb) to be integrated with immunohistochemical analyses of adjacent sections, and the resulting multiplex imaging data classified according to its morphological-biochemical state. New data analysis and visualization tools will be developed that allow the multiplex imaging data to be interrogated by all members of the collaboration, thus enabling the expertise of each member of the consortium to be fully exploited. All metadata and processing steps will be recorded in the workflow based knowledge management system created in P24.2 and model datasets and the results of the analyses uploaded into the Netherlands Virtual Tissue Bank.

These tools will be used to investigate if changes in the expression patterns of large panels of proteins can provide improved diagnostic/prognostic capabilities. The data analysis capabilities provided through VLe++ are crucial to the development of multiplex imaging of clinical patient tissues, and to establish if a system-wide approach can provide diagnostic/prognostic tools for currently difficult-to-diagnose/distinguish cancers.

Project number P24	
WP title & acronym	Data generation and metadata management for Biobanking with MRI and IMS
WP leader	Ron Heeren , AMOLF

Objectives

The aim of this work package is to establish content for a data management system and a web service to access the virtual tissue BioBanks. Through online collaboration 2d and 3D IMS datasets will be annotated and aligned. World-class pathologists will guide this data fusion process. The merged datasets will be made available through a virtual microscopy web service that will allow the online exploration and investigation of large image datasets originating from various image resources in the Netherlands. New methodologies for image generation, both local and remote will be developed in close collaboration with TI-COAST using the e-Biobanking infrastructure developed in other work packages In addition existing metadata query tools will be employed that will allow feature based searches through the different image instances of the organism under study. It will establish the possibility of combined and correlated searches through different imaging modalities in 2 and 3 dimensions. For this purpose, new co-registration approaches using IMS compatible markers will be developed.

Project number P24	
WP title & acronym	Framework Selection for e-Biobanking: Identification and Evaluation of Core Services (BBMRI)
WP leader	Kai Ye, LUMC

Objectives

In the Biobanking and Bimolecular Resources Research Infrastructure (BBMRI) Biobanking initiative of The Netherlands all Dutch Biobanking groups are assembled which have collected GWAS data in their BioBanks. GWAS data are available from over 200.000 subjects participating in genetic studies for diverse diseases. The Netherlands is participating in numerous international consortia because of its large number of well documented an dwell phenotypes cohort studies. The development of tailor made software that allows for the sophisticated mining of this collective dataset would generate high ranking deliverables in terms of discovering pathways for different human diseases and profiles which may be used as tools in early diagnostics of disease. The current analysis of these datasets is based on comparisons between (diseased) cases and (population) controls of single genetic variants (SNPs) in datasets in which hundreds of thousands of such variants have been measured. This material lends itself extremely well for the application of novel mining tools. The goal is to generate data mining tools that allow the mining of existing genome wide scan datasets (GWAS) aimed at the identification of genetic factors and pathways involved in diverse traits and diseases.

Project number P24	
WP title & acronym	Framework Selection for e-Biobanking: Identification and Evaluation of Core Services (Merck)
WP leader	Jacob, de Vlieg, Merck

Objectives

The goal of this work package is to support the evaluation, selection and deployment of generic frameworks for e-Biobanking services.(e.g. services that manage access to computational and data resources to services that deal with knowledge) for the drug discovery and translational research domain.

In order to realize these goals we intend to use the experimentation environment that was started in VL-e and will be brought further in BigGrid. This will be combined with analysis services and knowledge resources that are available from the NBIC Biorange research program. For this project Merck will be provided access through her research network.

The approach taken is as follows: given a set of representative use-cases from the pharmaceutical sector, the adequacy of the services provided in the framework (or frameworks) under consideration are identified. These are used to define so-called core services which are abstractions within the drug discovery and translational medicine application domain. Based on this foundation domain specific application can be build more efficiently.

The key issue is to evaluate the services with respect to their potential of quickly building new applications on top of it. Also issues as the integration of various information/data services while respecting privacy regulations are part of such studies. In this way a systematic approach is taken to find and evaluate existing services on their adequacy to act as core services for this domain. Besides general services, developed in e-science, also specific services in for drug discovery and translational research relevant domains have to be taken into account.

Merck and the Computational Drug Discovery group will make at least two sets of use-cases available for this work package.

DELIVERABLES

Number of important journal paper
36

Number of important conference contributions
16

2c Products

1. Paper describing grid-based tools to classify multiplex imaging datasets. MALDI imaging MS of a patient series of soft tissue sarcomas will provide a detailed spatio-chemical map of each tissue, comprising hundreds of peptides and proteins. A histological analysis of each tissue, of adjacent sections, will then be used to provide a detailed morphological map of each tissue. Alignment of histological images and the MALDI imaging MS datasets will enable the peptide and protein profiles to be linked to their morphological origin. A classification algorithm will be built to identify individual biomarkers that distinguish between soft tissue sarcomas (based on their morphology) and validated against additional patient tissues. LC-MS/MS of tissue extracts will be used to identify the candidate biomarkers followed by immunohistochemistry for independent validation.
   - WP4, month 24

2. Paper describing grid-based multivariate tools for correlation analysis of mulitplex imaging datasets. Imaging MS can define experiments solely on the basis of their peptide and protein profiles, and thereby identify regions undergoing pathological change prior to morphological transformation, or discriminate between tumors with overlapping/identical morphologies. A series of multivariate tools will first be used to analyze the imaging MS data from a well differentiated soft tissue sarcoma patient series, to test the capabilities of such imaging MS-based molecular histology to correctly classify well differentiated tumors.
   - WP4, month 24

3. Paper on effectiveness of protein panels for distinguishing between different soft-tissue sarcomas that have overlapping morphologies (enabled by grid computing of large datasets). The ability of morphology-based classification and imaging MS-based molecular histology to differentiate soft-tissue sarcomas that have overlapping morphologies will then be compared, thus establishing the potential of imaging MS to

- WP4, month 48

4. Important conference contribution – ability of imaging mass spectrometry to differentiate between morphologically distinct soft-tissue sarcomas on the basis of their peptide and protein profiles and thereby identify new specific biomarkers. The work will be presented at two different conferences. At the American Society for Mass Spectrometry annual meeting, the principal meeting for mass spectrometry specialists, a methodological presentation will describe the implementation of GRID based computing for the classification analysis of a patient series of tissues, as well as present any future availability of source codes and example datasets. The improvements in diagnosis and insights into the biology of the soft tissue sarcomas generated by this classification analysis will be reported at the Human Proteome Organization meeting, the principal meeting for MS based proteome analysis of human diseases. Preliminary data may be presented at the specialised Imaging Mass Spectrometry conference to take place in Ourense, Spain, 2012 (co-organized by project leader).
   - WP4, month 24.

5. Important conference contribution – ability of imaging mass spectrometry based molecular histology to discriminate between well differentiated soft tissue sarcomas. The potential of imaging mass spectrometry to provide new molecular histology capabilities lies in its ability to detect changes in peptide and protein expression prior to (and independent of) morphological change. Consequently imaging mass spectrometry can detect regions of tissue that bear the molecular hallmarks of pathogenic transformation prior to histology. To detect these transformations requires data analysis capabilities that extend beyond morphological analysis. Two presentations will be prepared. A methodological presentation describing GRID based computing for performing multivariate analysis of multiplex imaging datasets, to be

- WP4, month 48.

6. PhD thesis. A full description of the implementation of the GRID based computational analysis of imaging mass spectrometry data will be provided, spanning automated feature recognition of the peptide and proteins detected in the imaging mass spectrometry datasets (a necessary prerequisite for the statistical analysis of these datasets), alignment and integration with histological analysis, and the statistical analysis. The thesis will be structured as an investigation of how imaging mass spectrometry analysis of entire patient series can complement current histological practice, but which requires the computational power of GRID based computing to analyze the very large number of variables and observations (e.g. 20k per tissue, 50 tissues, 500 variables). A series of quality control tests and performance metrics will be provided – describing the performance of the data analysis routines with an increasing number of nodes. A central tenet of the work will be to use the GRID based computational facilities to optimize the data analysis routines, such that they can then be run on lower power workstations (thus more valorizable). The thesis will then describe how imaging mass spectrometry can complement current histological practice by identifying new biomarkers and providing new diagnostic tools for differentiating between morphologically overlapping soft tissue sarcomas.
   - WP4, month 48.

2d Software

1. Biosignals
   Software for alignment and normalization of biosignals. A matlab processing pipeline will be created within the BIGGRID. This pipeline will serve as standard pre-

processing tool for separating out noise from signal in fMRI data, partially based on external physiological and behavioral measurements.
- WP 1 YP 2012

2. e-imaging MS processing software, WP2, month 24. Update to freely available imaging MS evaluation software, WP2, month 36.
   - WP 2 YP 2013

3. IBD e-biobank
   First prototype of the IBD e-biobank, year 2.
   Open source software, e-biobank tools, end of year 2.
   Open source software, e-biobank environment, end of year 3.
   Open source software, integrated e-biobank, end of year 4.
   Second prototype of IBD e-biobank including integration of images, year 3.
   -WP 6 YP 2014

4. GWAS data mining tools
   First prototype of the GWAS tool set, year 2.
   Open source software, GWAS data mining tools, end of year 2.
   Open source software, improved GWAS data mining tools, end of year 3.
   Open source software, integrated tool set GWAS data mining tools , end of year 4.
   Second prototype of the GWAS tool set, year 3.
   - WP 7 YP 2014

5. Data management software
   First prototype of the distributed biobank, year 2.
   Open source data management software, end of year 2.
   Open source data management software, including secure distribution, end of year 3.
   Open source data management software, integrated tool set, end of year 4.
   Second prototype of the distributed biobank, year 3.
   - WP 9 YP 2014

6. Description e-Front v 1.0 software published as open source, WP24.3, m14. e-Front v 2.0 software published as open source, WP24.3, m38
   - WP 3 YP 2013

2e User studies

1. User study evaluation of the method on a large group of subjects.
   Emotional human responses to video content will be predicted based on one fMRI and other biosignals. We will study the lower-envelope of biosignals needed for robust and reliable predictions.
   - WP 1 YP 2013

2. Next Generation Sequencing
   Grid scale alignment pipeline with provenance and transparent job failure handling. This system will be able to run on grid, cloud and super computer infrastructures. This will be implemented for the Illumina HighSeq and Complete Genomics data. m12 Algorithm for combining SNP and indel data. m24. Large scale alignment of genomics data using the Hadoop platform. m36
   - WP 7 YP 2014

2f Other results

1. Data collection of neural, physiological and behavioral signals of 500 subjects. The raw data is currently being acquired at the Spinoza Neuroimaging Center within the context of the ID1000 project. The data is stored in different formats and at different places. At m6, a uniform data set will be available and stored at a single site.
   Infrastructure, e-science environment for massive analysis of biosignals. A Matlab working environment will be created within the e-science platform BIGGRID. It will eventually serve as standard pre-processing environment for biosignals (neural, physiological, behavioral).
   Method functional data representation of heterogeneous biosignals. Biosignals will be represented as functions rather than discrete data and pre-processed to remove unwanted variation taking into account the heterogeneity of the data. At m12, a data pre-processing method will be available specifically developed to handle heterogeneous biosignal data.
   Data collection of signals for remaining 500 subjects. At m24, the ID1000 project will finish the acquisition of data from 1000 subjects. The raw data of the remaining 500 subjects will be transformed to the same format as and added to the set collected at m6.
   Method (UvA, m30), method for detection of patterns in heterogeneous biosignals Neural, physiological and behavioral human responses differ in time-scale at which potentially interesting patterns emerge. At m30, a method based on functional data analysis will be available to extract such reliable patterns across time-scale.

Data visual signals derived from movies presented to 1000 subjects. Movies shown to 1000 subjects in the ID1000 study will be automatically analyzed for low-level features and high-level concepts. This will result in time-varying perceptual, cognitive video annotations.

Method for combined analysis of heterogeneous biosignals and (visual) stimulus signals. At m36, a method will be available to match computer-extracted descriptions of video content to specific emotional responses of humans to the
- WP 1 YP 2014

2. In most clinical and cognitive studies, fMRI patterns with a physiological source such as heartbeat are considered to be irrelevant. To cancel out physiology-related patterns from the fMRI data, these data are increasingly being acquired together with heartbeat signals, ECG etc. Standard tools already exist to remove out intrinsic noise and unwanted artifacts such as head movement from fMRI the data. No tools exist, however, to efficiently and accurately remove physiology-related patterns from fMRI data. In WP24.1 we will create a new and advanced approach for separating out noise and irrelevant physiological patters caused by for example heartbeat from fMRI data. The data analytic framework to be used is functional data analysis, a branch of statistics dealing with curve (signal) data, not discrete data. The approach will be implemented in Matlab and will form an integral and crucial first part of the standard fMRI-processing pipeline of the Spinoza Neuroimaging Center and AMC (open downloads, UvA, m18).
   In marketing neural, physiological and behavioral measurements are increasingly being used to predict and steer success of branding, advertising and packaging. With fMRI, it is today possible to accurately and objectively identify the type of emotion a product, advertisement or commercial evokes. However, fMRI studies are costly and time-consuming. For these reasons, marketing companies are very much interested in limiting the use of fMRI and, instead, want to capitalize on easily measurable biosignals such as heartbeat and blood pressure. This requires tools to map (combinations of alternative) biosignals to emotional states. In WP24.1 we will develop a tool that uses a training set of fMRI and physiological responses to emotional stimuli, to learn what patterns in biosignals are predictive of what emotional state. The learned patterns can than be used (by marketing companies) to predict emotional responses to product, advertisement or commercial without the need to acquire fMRI data. UvA spinoff Neurensics is interested in WP24.1 results for neuromarketing purpose (product transfers, UvA, m30).

There are many ways to tell a story in video and many ways to budget a video. New scientific insight in how humans react to (emotional) videos neurally, physiologically and behaviorally, together with state-of the art technologies capable of specifying video content, are creating new ways for video production. In WP24.1 we will combine these two exciting new developments to explore new ways to tell a story in video and, hence, new ways to budget a video, for example a commercial. Results of WP24.1 will be disseminated to a broader public of scientist, movie makers, marketing companies through a popular paper on the relation between movie content and biosignals (popular paper, UvA, m36). In addition, we aim to build a demonstrator on video content prediction and production from biosignals (demonstrator, UvA, m48).

WP24.1 is internationally unique because of the combination of: 1) very large number of monitored subjects 2) the heterogeneity and diversity of biosignals recorded from these subjects and 3) integrative analysis of biosignals with stimulus (movie) signals. There are a few other groups in the world that integrate neural, physiological, behavioral and computational responses: prof. J. Gallant, Berkeley, prof. Alan Smeaton, Dublin University, prof. Shih-fu Chang Colombia University. None of these, however, collects data on our scale and uses sophisticated data analysis suitable for integrative analysis of heterogeneous signal data. There is contact with the last two research groups through WP1.10 (international cooperation, UvA, m48).

A direct cooperation exists with prof. Uri Hasson from Princeton University, who studies human brain responses to directed movies. In several neuro-cinematography studies he has collected fMRI, ECOG and intracranial recordings (from epilepsy patients) responses to movies from different genres. His interest is in automatic analysis of movie content and correlation of movie content to human brain responses, but lacks tools and methodology to do so. He therefore will provide us with movies from different genres that have been passively viewed by multiple subjects while fMRI and ECOG were recorded. In return we will computationally analyze these movies, extract spatio-temporal signals from them, and relate this signals to fMRI and ECOG with tools developed in WP24.1 (international cooperation, UvA, m30).

The main focus of WP24.1 is the analysis of biosignals: neural, physiological and behavioral responses of humans to video content. Included in this analysis will be the relation between biosignals and signals extracted from video content. To this end, in WP24.1 we will apply low-level feature detectors to video data, and use detector responses over time to characterize video content. Advanced content analysis of the movies will be done in cooperation with WP1.10 (synergy, UvA,m36). In WP1.10

emotional content will be predicted automatically from video data based on behavioural responses and the automatic detection of high-level concepts over time. This paves the way for unifying neural, physiological, behavioural and computational signatures of emotional states. The methodology by which this will be done is functional data analysis. Potentially, synergy between WP24.1 and WP1.10 will result in a system that predicts what video production leads to what emotion. UvA spinoff Neurensics has an interest in such a system for prediction of the impact of a commercial and/or storyboard.
- WP 1 YP 2014

3. Virtual tissue bank available for biomedical professionals, WP2, month 24. Update to virtual tissue bank available for biomedical professionals, WP2, month 48.
   Key of the virtual tissue biobank is its content. In collaboration with a number of medical partners will aim to evaluate existing tissue collections of various breast tumors with imaging MS and contributed the results to the Virtual tissue bank for exploration and annotation by collaborating pathologists. This will result in the PSE below.
   PSE for molecular histological imaging of cancer tissues, WP2, month 48.
   Popular paper describing Virtual Tissue Bank and its relevance for medical research, WP2, month 48.
   Netherlands virtual tissue bank available for biomedical professionals, WP2, month
   Update to Netherlands virtual tissue bank available for biomedical professionals, WP2, month 48.
   Training course on e-biobanking and molecular imaging, WP2, month 36.
   Distributed statistical analysis, part of European collaborative network, training and analysis. WP2, month 24.
   Distributed evaluation of imaging MS datasets of TMA's, part of European collaborative network, training and analysis. WP2, month 48.
   Workflow based knowledge management, part of European collaborative network, training and analysis. WP4, month 24.
   Virtual Tissue Bank, part of European collaborative network, training and analysis. WP4, month 48.

   Knowledge Management in Science and E-Infrastructure Virtualization for E-Science Applications – exploitation tools for data intensive distributed applications and distributed access to data sources developed under COMMIT, WP2, month 40.
   - WP 2 YP 2013

4. The front-end and services will be deployed as open services. This will enable to external end-users facilitated access to the BiGGrid infrastructure, and possibly also to EGI, to speed-up large-scale data analysis experiments, such as neuroscience studies and sequencing. This will help us contact a user base to collect requirements and evaluate the generated results.

   Release services for external users, WP24.3, m30

   The aim is to make the services (for example the scientific gateway for neuroimaging) available to external researchers in other institutions. This will facilitate dissemination to other communities.

   Results will be disseminated beyond the AMC, including end users from the Spinoza Brain Imaging Center Amsterdam (www.spinozacenter.nl/), in particular P24.1 (e-BioCognition),

   SILS (bioinformatics), NBIC BioAssist, NPC, NGS and Biobanking platforms (www.nbic.nl/research/research-projects/e-bioscience/), BBMRI-NL (biobanking), Parelsnoer (www.parelsnoer.org/), DigiBob (www.surfnet.nl/en/bestpractice/Pages/DigiBOB.aspx) and the European Life Science Grid Community (EGI virtual research community, http://wiki.healthgrid.org/LSVRC:Index). These users will benefit from the virtual environment services and contribute to this WP by providing requirements and evaluating the results.

   Activities will advertised at the COMMIT, AMC, UvA, NBIC, BBMRI-NL, GridForum, EGI and HealthGrid newsletters.

   The WP will attempt to reach the press with results of significant studies, e.g. GvNL.

   We will also participate in training sessions in NL and EU, and introduce a new course about the usage of these services in the AMC graduate school.

   Publicity paper, article on general public magazine tbd, WP24.3, m47

   The plan is to write a paper presenting the scientific gateway for neurosciences in the perspective of healthcare for psychiatric patients. This will be elaborated in conjunction with neuroscientists from the AMC.

   Demonstration in conference tbd, WP24.3, m20

   The neuroscience gateway will be demonstrated in a conference, for example HealthGrid or EGI User Forum.

   Demonstration in conference tbd, WP24.3, m36

   The gateway for the second use case (to be selected, deliverable 2e-2) will be demonstrated in a conference, for example HealthGrid or EGI User Forum.

   Training/education event in The Netherlands, WP24.3, m24

We will organize events to train new users to use the developed virtual environments. These will be presented in events organized by COMMIT, as well as external events, for example, linked to EGI, NBIC, BBMRI and other organizations.

Training/education event in The Netherlands, WP24.3, m40

e-Science course at the AMC Graduate School, WP24.3, m20

We will propose a course to AMC PhD students to disseminate concepts of e-science and e-infrastructure, with hands-on tutorials using the developed virtual environments.

e-bioscience course at the AMC Graduate School, WP24.3, m40

A large barrier for the adoption of e-infrastructures in Life Science research is the poor usability for the target end-users. The results obtained in this WP will contribute to lower this barrier, in particular by its dissemination (services, training and support) in the scope of the Life Science Grid Community under formation (EGI-Inspire project (www.egi.eu/projects/egi-inspire).

Release services for users of European Life Science Grid Community, WP24.3, m47

The Life Science Grid Community (LSGC) is the representative body for communication with the European Grid Initiative (EGI). It is a virtual organization that attempts to disseminate and organize users from the biomedical research community. Formally the HealthGrid association represents this community. The goal is to make the developed platform available for this community, which will provide large visibility to the COMMIT project.

Member of HealthGrid board, WP24.3, m1

S. Olabarriaga is member of the board of the HealthGrid association.

Member of Life Science Grid Community, WP24.3, m1

S. Olabarriaga is the contact person for the LSGC in The Netherlands, as well as the contact person with EGI for user support and dissemination.

Collaboration with SHIWA project (FP7-EU), WP24.3, m1

The SHIWA project aims at developing concepts and tools to enable interoperabilituy of grid workflow systems at various level (www.shiwa-workflow.eu, end July 2012). The AMC is partner in the project, and will search collaboration to become an early adopter of the SHIWA technology in the virtual environments.

Collaboration with SCI_BUS project (FP7-EU), WP24.3, m4

The SCI-BUS project aims at the development of scientific gateways based on LifeRay, P-Grade and SHIWA technologies (start October 2011).The AMC is partner in the project and will develop a scientific gateway for biomedical researchers at the AMC.

Organization of HealthGrid conference at AMC, WP24.3, m12

The aim is to organize the HealthGrid conference in Amsterdam in 2012 (normally June). This is the conference that gathers the members of the HealthGrid association, which covers the EU and US. The proposal has been informally discussed with the HealthGrid board and accepted. It would be excellent to have a stamp of COMMIT in it.

organization of workshop about front-ends for e-infrastructures, WP24.3, m45

The idea is to organize an international workshop as a satellite activity of a larger conference, for example, CCGrid. This is a very initial plan, since we need to see in which direction the "scientific gateway" wave will take us. It could become an edition of the current IWSG-Life: Workshop for Life Sciences – (https://sites.google.com/a/staff.westminster.ac.uk/iwsg-life2011/); or it could become a separate event.

Training in European event, WP24.3, m40

We wish to present some training/tutorial about the developed technology (generic framework used to build the virtual environments) or about the usage of the environment for a specific community (e.g., neurosciences). This will be chosen at a later stage.

We will collaborate closely with P20 (E-INFRASTRUCTURE VIRTUALIZATION FOR E-SCIENCE APPLICATIONS (workflow, security, architecture for data-intensive application) and observe the results generated in P14 (TRUSTED HEALTHCARE SERVICES, secure identity management). Results will be broadly disseminated inside and outside the COMMIT consortium

PS: adding more details here at this time is difficult because of the dependencies with other WPs and deliverables, therefore we only present the general approach.

In all cases below there will be contact between the parts, presentations from both sides about the complementary functionalities, a discussion about integration, adoption or interoperability options, and a follow-up plan. Although the opportunities of cross-fertilization inside the consortium are immense, the migration of solutions from prototype to production environments (as it is necessary for the virtual environment prototypes) is not trivial. For this reason we are not sure that all the deliverables below will indeed go beyond the "consideration" level. The delivery dates are tentative.

Consider functional data analysis software developed in WP24.1 (e-BioCognition) for integration into the virtual environment, WP24.3, m30

Consider the functional data and use cases collected in WP24.1(e-BioCognition) for evaluation of virtual environment, WP24.3, m30

Consider grid security methods and software developed in WP20.3 (SESI) for adoption in virtual environment, WP24.3, m30

Consider collaborative workflow system models and prototype generated in WP20.8 (WOPMOM) for integration into virtual environment, WP24.3, m30

Consider workflow services developed in WP20.9 (WSAR) for integration into virtual environment, WP24.3, m30

Consider component integration solutions developed in WP20.10 (WACI) for implementation of the virtual environment framework, WP24.3, m30

- WP 3 YP 2014

5. Tools and methods for alignment and integration of MRI and IMS, WP5, month 30.

   The generation of co-registered molecular image data is crucial to the success of a virtual tissue biobank. Automated feature recognition and alignment tools will result from this work package. A novel method using fiducial markers that have molecular signatures as well as optical signatures will be employed for post-analysis alignment and reconstruction of 3D molecular volumes. After alignment statistical correlation analysis will be a new indicator for co-registered data quality. Some of the markers found will be related to tissue degradation and hence can be used for orthogonal quality assessment of the molecular image data entered in the database. Protocols for data quality assessment are expected results from this work package.

   Sample multimodal IMS datasets of breast cancer tissues, WP5, months 36 and 48 (update).

   In collaboration with TI-COAST and the national cancer institute an extensive multimodal datasets on triple negative xenografted breast tumour sections. The newly generated dataset will contain secondary ion mass spectrometric data, MALDI-imaging MS analysis of peptides and proteins, lipid profile data, haematoxylin and eosin high resolution MIRAX data, immunohistochemical data targeting specific known and newly discovered cancer protein markers and RNA expression data taken from previous studies. In addition a literature collection on tumour origin and patient studies will be added and made available to the data collection.

   Virtual microscopy browser for interrogating exploring aligned multimodal imaging (this case histology and IMS), WP5, month 36.

   This browser will be evaluated and validated by medical professional in different medical institutions in the Netherlands and beyond. Initially focusing on the collaborative partners within P24 and its work packages. For large scale processing

distributed computing opportunities will be explored in collaboration with the group of Henri Bal.

Presentation Sample histology and IMS datasets of breast cancer tissues, WP5, months 36 and 48 (update).

Demonstrator of multimodal imaging datasets of breast cancer tissues, WP5, month 48.

Training course on virtual microscopy of multimodal imaging, WP5, month 36.

histology 2D/3D-IMS alignment methodology, part of European collaborative network, training and analysis. WP5, month 24.

Virtual microscopy browser, part of European collaborative network, training and analysis. WP5, month 48.

Public multimodal imaging datasets of breast cancer, part of European collaborative network, training and analysis. WP5, month 48.

Dissemination of multimodal visualization and processing tool to PNNL-EMSL molecular imaging facility. WP5, month 36-48

Parallel metadata generation on molecular image datasets in collaboration with P22, WP5, month 48.

Integration of web based knowledge management tools in collaboration with P23, WP5, month 48

- WP 5 YP 2014

6. Report on Mining of e-biobank, year 2

   Report on Mining of extended e-biobank, year 4

   Identification of subgroups of IBD patients, based on IBD e-biobank, end of year 2.

   Identification of subgroups of IBD patients, based on extended IBD e-biobank, end of year 4.

   Participation in international computer science conference with presentation and demos, year 1.

   Participation in international medical conference with presentation and demos, year 2.

   Participation in international computer science conference with presentation and demos, year 3.

   Participation in international medical conference with presentation and demos, year 4.

   Joint (together with CWI database experts in P19) design of the IBD e-biobank.

   Report at the end of year 2.

   - WP 6 YP 2014

7. Report on Mining of GWAS studies, year 2
   Report on Mining of extended GWAS studies, year 4
   Identification of subgroups of patients in large cohort study, end of year 2.
   Identification of subgroups of patients combining multiple large cohort studies, end of year 4.
   Participation in international computer science conference with presentation and demos, year 1.
   Participation in international medical conference with presentation and demos, year 2.
   Participation in international computer science conference with presentation and demos, year 3.
   Participation in international medical conference with presentation and demos, year 4.
   - WP 7 YP 2014

8. Report on Mining of the distributed biobank, year 2
   Report on Mining of the secure distributed biobank, year 4
   Identification of subgroups and biomarkers, end of year 2.
   Further Identification of subgroups and biomarkers, end of year 4.
   Participation in international computer science conference with presentation and demos, year 1.
   Participation in international medical conference with presentation and demos, year 2.
   Participation in international computer science conference with presentation and demos, year 3.
   Participation in international medical conference with presentation and demos, year 4.
   - WP 9 YP 2014

9. Downloads/Dissimination
   Download of example datasets of multiplex imaging of clinical tissues, WP4, month 48. To enable the further development of imaging mass spectrometry based analyses of diagnostically challenging tumors the imaging mass spectrometry data (non feature selected), as well as the reduced data (feature selected), and aligned histological analyses of the patient series of soft tissue sarcomas will be made available. The

data will be separated according to tumor type, histological grade and clinical stage and will be completely anonymous (further clinical data can be made available on request). A histopathological summary of the data will also be provided, clearly indicating which tissues are morphologically distinct and which are morphologically overlapping.

To enable the further development of the statistical tools crucial to imaging mass spectrometry based analyses of clinical tissues the GRID based data analysis algorithms will be made available as either a free source download or a licensed download. A set of instructions will be provided, including a summary of the results that can be obtained by their application to the example datasets of soft tissue sarcomas. Automated feature detection (of the peptides and proteins detected in the imaging mass spectrometry dataset) and alignment with histological analyses are crucial to these multiplex analyses. Accordingly the data reduction and alignment tools will also be made available (free source download or licensed download) and instructions provided, WP4, month 48.

Popular paper describing grid computing as an enabling technology for multiplex molecular imaging of clinical tissues, WP4, month 48. Imaging mass spectrometry generates very large datasets, containing the distributions of hundreds of peptides and proteins. The analysis of patient series of tissues is crucial for its clinical application. However just 50 tissues, each analyzed with 20k pixels, can generate a total data set of 1M observations of ▫500 features. The computational power of GRID based computing is necessary to perform statistical analyses of these large, multidimensional datasets, and thus establish the potential of imaging mass spectrometry to provide new diagnostic capabilities. In this article the implementation of GRID based computing for clinical imaging mass spectrometry will be described, followed by a demonstration of it capability to discriminate between morphologically overlapping soft tissue sarcomas. An explicit comparison between the accepted morphological grading, clinical staging and imaging mass spectrometry results will establish the potential of imaging mass spectrometry to complement current histopathological practice.

Knowledge transfer - imaging mass spectrometry training course about GRID based data analysis in imaging mass spectrometry, WP4, month 42. Detailed imaging mass spectrometry training courses, free to PhD students and Post-Doctoral researchers, are organized through a European imaging mass spectrometry consortium (currently funded via Nordforsk, follow-up COST Action submitted). The training courses cycle through the different aspects of the experiment, including data analysis and validation (Turku, Finland, Dec. 2009), and provide the ideal setting to communicate

the unique data analysis capabilities provided by GRID based computing, namely simultaneous analysis of entire patient series. An initial lecture will focus on the methodological background of the analysis. A follow-up practical will then explicitly demonstrate the implementation and performance of the data analysis routines, WP4, month 48.

Publicly available datasets and demonstrators of grid-computing, WP4, month 48. The soft tissue sarcoma datasets, aligned with a histological analysis of adjacent tissue sections, as well as the data analysis routines will be made publicly available (free download or licensed). A detailed set of instructions, complemented by imaging mass spectrometry training courses, which cover the GRID based analyses and the statistical background, will ensure the dissemination of the data analysis routines to the wider mass spectrometry and healthcare communities. The demonstrators will also include an explicit work-flow example of the alignment and integration with histology, and data reduction via automated feature detection and extraction of the imaging mass spectrometry datasets (both crucial steps in the data analysis workflow). All results of the analyses will be uploaded into the Netherlands Virtual Tissue Bank (P24.2).

European collaborative network, training and analysis. WP4, month 24. The research will join an existing imaging mass spectrometry collaborative network called Nordic signals, funded by Nordforsk, and involving research groups from Sweden, Finland, the Netherlands, Germany, France, United Kingdom and Switzerland. An application has been made for expanding this network as a COST Action (coordinator Liam McDonnell), whose explicit goal is develop standardized methodologies for healthcare research through an extensive exchange program and training. The unique capabilities provided by the GRID based analysis of imaging mass spectrometry datasets can thus be investigated for a range of pathologies (through the exchange program of Nordic Signals and COST), for both biomarker discovery (morphology driven analysis) and improved diagnosis (molecular histology).

The intensive data analysis routines developed under this imaging mass spectrometry analysis of clinical patient series, and the requirement of multiple image modalities (mass spectrometry and histology) that may be stored in different locations, may benefit from the tools for data intensive distributed applications and distributed access to data sources developed under COMMIT WP4 (Knowledge Management in Science and E-Infrastructure Virtualization for E-Science Applications). An audit of the compatibility of the data analysis and image integration routines will be performed to assess compatibility and where performance gains can be expected

(both in terms of increased processing and implementation) the data analysis algorithms will be adapted. M48.

- WP 4 YP 2014