Deliverable No. 2.1

State of the art review of the p-medicine environment

Grant Agreement No.:	270089
Deliverable No.:	D2.1
Deliverable Name:	State of the art review of the p-medicine environment
Contractual Submission Date:	30/09/2011
Actual Submission Date:	30/09/2011

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	***p-medicine***
Project Full Name:	From data sharing and integration via VPH models to personalized medicine
Deliverable No.:	D 2.1
Document name:	State of the art review of the p-medicine environment
Nature (R, P, D, O)[1]	R
Dissemination Level (PU, PP, RE, CO)[2]	PU
Version:	1
Actual Submission Date:	30/09/2011
Editor: Institution: E-Mail:	Holger Stenzhorn USAAR holger.stenzhorn@uks.eu

ABSTRACT:

This deliverable presents an overview of the current state-of-the art as found in the various areas of research dealt within the p-medicine project. It touches upon the variety of topics, such clinical trial process and standards, system design and architecture, clinical decision support systems, VPH modelling and the integrated Oncosimulator, high performance and cloud computing, semantic mediation and data integration, user-interfaces for complex systems as well as bioinformatics and personalized medicine.

KEYWORD LIST: State of the art review

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 270089.

[1] **R**=Report, **P**=Prototype, **D**=Demonstrator, **O**=Other

[2] **PU**=Public, **PP**=Restricted to other programme participants (including the Commission Services), **RE**=Restricted to a group specified by the consortium (including the Commission Services), **CO**=Confidential, only for members of the consortium (including the Commission Services)

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0 alpha	11/09/2011	Draft	Holger Stenzhorn
1.0 beta	19/09/2011	Draft	Holger Stenzhorn
1.0 prefinal	27/09/2011	Draft	Holger Stenzhorn
1.0 final	30/09/2011	Final	Holger Stenzhorn

List of contributors

- Alberto Anguita, UPM
- Danny Burke, ecancer
- Marie-Luise Christ-Neumann, FhG-IAIS
- Ruslan David, USAAR
- Dimitra Dionysiou, ICCS-NTUA
- Norbert Graf, USAAR
- Benjamin Jefferys, UCL
- Lefteris Koumakis, FORTH
- Wolfgang Kuchinke, UDUS
- Aisan Maghsoodi, Philips
- Juliusz Pukacki, PSNC
- Simona Rossi, SIB
- Stelios Sfakianakis, FORTH
- Georgios Stamatakos, ICCS-NTUA
- Holger Stenzhorn, USAAR
- Marian Taylor, UOXF
- Giorgos Zacharioudakis, FORTH

Contents

1 Introduction

The application of the latest high-performance computing methodologies – which have seen a tremendous advancement over the last couple of years – upon the gigantic amounts of data gained in current post-genomic research and clinico-genomic trials offer the enormous chance for novel, individualized treatment regimens for cancer patient that vastly advance their prognosis and outcome. But in order to make reaching this goal possible, a common infrastructure needs to be set-up for both (basic) researchers and clinicians. Such would enable them to efficiently share, link and analyse the data encompassing a multitude of granularity and modality levels, such as patients' clinical information, bio-molecular findings or imaging studies. In order to be accepted by its potential users, this infrastructure must also obey the multiplicity of needs and requirements stemming from their daily work routine incorporating many aspects which include, for example, the capability of interfacing with existing systems or the usability of specific software tools.

Consequentially, it is sensible – before actually defining the needs and requirements – to thoroughly analyse and describe the present approaches and tools found in the prospective users' work environments as well as existing technologies and their applicability to the envisaged infrastructure. To make this overall goal more concrete, the task at hand is to provide reviews on current decision supporting systems, tools and software for the seamless integration of clinical care and basic research data, clinical trial guidelines, repositories of clinical, bio-molecular, and medication information, and so on. And as a European project, this task necessarily incorporates the latest progress and achievements gained in other past and on-going European projects as well.

The overall reviewing task has been partitioned following the specific expertise of each of the partners participating in this task which has resulted in the following natural distribution of work and sections within this deliverable:

- USAAR, UDUS and IEO: Clinical trial process and standards
- FORTH: System design and architecture
- Philips: Clinical decision support systems
- UCL and ICCS: VPH modelling and the integrated Oncosimulator
- PSNC: High performance and cloud computing
- UPM: Semantic mediation and data integration
- FhG-IAIS: User-interfaces (for complex systems)
- SIB: Bioinformatics and personalized medicine

The results of this deliverable have to be seen in close connection with the ones described within the deliverable of task 2.2 where the needs and requirements of the prospective users are described in a scenario-based fashion. Both deliverables D2.1 and D2.2 together provide both the architects of the p-medicine platform and tools as well as its implementers with the required information to create a system that is based and fulfils both on the actual needs of the people in their regular daily work and also does not "reinvent the wheel" by incorporating existing standards, technologies and methodologies.

(Note: All trademarks and web addresses that are mentioned in this deliverable are the property of their respective owners.)

2 Clinical Trials Process

Clinical trials are conducted in three main phases [1]:

1. Planning and preparation of clinical trial
2. Trial management and clinical data collection
3. Analysis and reporting of results

The recent developments and improvements of the state of the art procedures of clinical trials are distributed unevenly between these parts. Whereas the first phase is still hampered by many requirements for the need for ethical and regulatory approval [2,3], the data collection phase has seen an increasing use of electronic data capture techniques.

2.1 Planning and Preparation of Clinical Trials

The trial process for medicinal products requires a number of important preparatory steps before the first patient can be enrolled. The most important preparatory step is the creation of the clinical trial protocol describing in detail the objectives, research design, methodology, statistical considerations and the participation and organisation of the clinical trial. The content and structure of the trial protocol is determined by guideline ICH GCP E6 [2]. Trial protocols can differ considerably according to the medical areas concerned. Especially protocols of oncological trials can differ considerably. The protocol heterogeneity impedes the standardisation and reuse of protocol elements for protocol implementation. Before the protocol is written, protocol feasibility has to be explored to ensure that only a trial that is feasible and therefore with a high chance of success will be conducted. Still, too many clinical trials show difficulties to recruit the planned patients quota and are running late [3]. Proper feasibility analysis can identify the existence of a suitable patient population. The current recruitment rate of adult patients into clinical trials is low. On average, about less than 3 to 5% of newly diagnosed cancer patients are enrolled in clinical trials [4]. The demands for the personalization of diagnosis and treatments will have consequences for the trial design, especially smaller patient populations have be considered. Personalized drugs and translational medicine may result in the need for higher trial flexibility, more adaptive trials as well as smaller trials with high complexity.

Prior to the enrolment of patients in a trial (for medicinal products or devices) the approval of the Competent Authority (CA) and a positive vote of an Ethics Committee (EC) must be obtained. To obtain approval, documents must be prepared and submitted, including patient information, informed consent, and patient insurance information. Any clinical trial on a medicinal product requires a Clinical Trial Authorisation (CTA) from the CA in the EU member states in which the trial is being carried out. The management of these requirements is a considerable burden especially in case of international academic trials.

A Trial Master File (TMF) and an Investigator Site File (ISF) have to be prepared and to be gradually filled with study documents during the course of the trial. An Investigator's Brochure (IB) contains the efficacy and safety details the investigators and other clinical staff should know before administering the test product to humans. For data collection a special form, the Case Report Form (CRF), has to be designed. Additionally, the storage, distribution and management of the medicinal product have to be planned. Procedures for the import of laboratory data and a process for dealing with laboratory values must be in place. Standard Operating Procedures (SOP) ensure the quality of the trial conduct. They should be authorised, reviewed at regular intervals and staff should be trained in using relevant SOPs.

The EC is an independent body of healthcare professionals and non-medical members, whose responsibility is to protect the rights, safety and well-being of human trial subjects.

The EC gives an opinion about the quality of the trial protocol, the suitability of investigators and the adequacy of site facilities. The CTA application form, accompanying guidance documents and the EudraCT number can be obtained from the EMA website (https://eudract.ema.europa.eu). Insurance coverage for all trial subjects must be obtained. In international trials insurance certificates must be provided from each participating country.

2.2 Managing the Clinical Trial Process and Clinical Data Collection

To support the conduct of clinical trials a framework consisting of recruitment, randomisation/blinding, data management, adverse event reporting, and monitoring must be in place. In the first step patients are screened by the investigator for participation and checked against inclusion and exclusion criteria. It is an important aspect for a successful clinical trial to achieve the planned number of patients, to meet the required sample size to achieve a convincing statistical power. Before participating in a trial the patient has to give an informed consent. It is guaranteed that all trial subjects entering the trial have their human rights guaranteed, their personal data is protected and patients can withdraw their informed consent at any time without any consequence. During randomisation the patient that meets the inclusion criteria is randomly assigned to a particular treatment, to either the new medicine, a medicine that is considered standard therapy, or a placebo.

Patient data is collected at investigator sites (trials centres) by using CRFs and a Clinical Data Management System (CDMS). Traditional paper-based data collection methods are time consuming, expensive and can result in incomplete or invalid data. The adoption of Electronic Data Capture (EDC) has grown rapidly over the past years and has reached about 50% of all clinical trials [5]. EDC is a technique for collecting clinical trial data in such a way that they are delivered to the sponsor in electronic form instead of paper. This includes mostly the following scenario: Information that is first recorded on paper by the investigator or the patient and is subsequently entered into a computer at the investigator's site, and is then delivered electronically to the sponsor. The computerized system into which the investigator enters the clinical trial data is generally provided and maintained by the sponsor or a third party vendor, for example an EDC provider or a Clinical/Contract Research Organisation (CRO). It is customized for each trial and may include data entry support mechanisms to validate the data as the data are being entered, thus resulting in cleaner data compared to paper CRFs. In addition, clinical laboratory data are transmitted to the sponsor electronically and batch-loaded into the sponsor's database. Data can be captured directly by electronic patient reported outcome (ePRO): the patient enters information on an electronic device. Finally, collected data is queried, cleaned, stored and analysed with the CDMS.

The European Clinical Research Infrastructure Network (ECRIN) [6] is a current EU-FP7 funded project to support European academic clinical trials. ECRIN consists of integrating national clinical research facilities into a Europe-wide network, able to provide support to clinical research in any medical field, and for any type of clinical research through information and consulting, and through a set of flexible services for the conduct of multinational clinical studies. ECRIN links national networks of Clinical Research Centres (CRC) and Clinical Trials Units (CTU) in twelve countries. To support its trials ECRIN will employ an IT framework, using data management systems located in dedicated and GCP-certified ECRIN data centres. An ECRIN wide survey was carried out to determine the types of CDMS in use in European academic research centres that gives an overview of the state of the art of clinical data management in Europe [7]. In general, the vast majority of centres conduct data management: In nearly 80 to 90% of centres with own data management a CDMS is in routine use. Many different types of CDMS are employed with the focus on commercial products (50 to 60%) and proprietary solutions (30 to 40%). Altogether twenty different commercial CDMS products, seven different open-source solutions and 17/18 proprietary solutions are in use. Of the commercial products the most widely used ones are MACRO and

Capture Systems, followed by solutions that are employed in at least three centres, namely eResearch Network, CleanWeb, GCP Base and SAS.

Of the CDMS in use, data collection (over 90%), query management (about 90%) and reporting (about 70%) are the most widely used functions in ECRIN centres. About 70% of CDMS are using eCRFs to collect data at investigator sites. A considerable number of ECRIN centres that employ special software or use a unique concept for clinical data collection exist. For example, the Copenhagen Trial Unit is using the groupware platform Lotus Notes/Domino from IBM for clinical trial data management. Another ECRIN centre uses the "Clinical International Trial Management System" CITMAS for patient tracking, enrolment, randomisation, data capture and reporting [8,9].

2.3 Examples of CTMS Used in Several ECRIN Centres

2.3.1 MACRO

MACRO from Infermed (http://www.infermed.com) is an EDC solution that offers both offline and online data collection and drag-and-drop study design. It is a solution that is used by many academic customers, for example at the Diabetes Trial Unit of the University of Oxford, National Blood Service (UK), Institut Gustave Roussy (Paris), Institut Curie (Paris), University of Vienna, several Study Coordination Centres (SCC) in Germany or the Netherlands Cancer Institute. MACRO employs drag-and-drop techniques to reduce study design time and provide control over the CRF layout. Multiple laboratories can be created for each study including detailed normal ranges and common toxicity criteria schemes. Commonly used questions and whole eCRFs can be stored in a MACRO library. In 2007 MACRO had been selected by the international leukaemia network (http://www.leukemia-net.org) as EDC tool for the clinical trials that are conducted within the network. The purchase of the commercial solution MACRO was a way to provide a validated GCP-compliant capture system for the network. In a separate approach, the network is involved in on-going collaborative projects to develop quality assurance methods for systems based on open-source software components [10]. As additional services, central randomisation and a PID-Generator were established. The web-based patient randomization service for multi-centre clinical trials permits patient randomization into one of two or more treatment arms. The PID-Generator generates a unique patient identifier for each study patient that is used as a pseudonym and can be used as input for a second pseudonymization. MACRO is used this way: After the trial protocol is finalized by the sponsor, trial structure and eCRFs are designed and the specification of the data dictionary is prepared, eCRFs are implemented and reviewed by the sponsor, after the eCRFs are validated, the user and site registration is performed, user training can begin.

2.3.2 Capture System

Capture System from Clinsight (http://www.clinsight.fr) is used in particular by many French investigators, for example by several units of the French Institut National de la Santé et de la Recherche Médicale (INSERM), Institut National du Cancer (INCa), Institut Régional Fédératif du Cancer, Centre Hospitalier Universitaire/Hopitaux de Rouen, Centre Hospitalier Universitaire de Nice, Centre d'Investigation Clinique - Epidémiologie Clinique Antilles Guyane (CIE 802 INSERM). It is a comprehensive solution for clinical data management, and monitoring with main modules:

- CSDesigner: electronic CRF design
- CSEntry: data entry from paper CRFs (single data entry, double data entry)
- CSOnline: web interface for investigators to add patients, to randomize them and to fill in the electronic case report forms with patient data

- CSTest: consistency tests and query creation
- CSCoder: medical coding (MEDDRA, WHODRUG, VEDDRA, ATCVET)
- CSExport: data export to statistical software solutions (SAS, SPSS) and in EXCEL, ASCII, HTML and XML formats
- CSMonitor: management of a clinical trial and monitoring

Capture System allows on-line data entry by investigator. It is compatible with the main browsers (Internet Explorer, Netscape, Firefox). Special functions offered are an entry guide, an on-line help during data entry, a data entry tracking table, history of data modifications (audit trail) as well as the ability to print CRFs in PDF format. During data entry into eCRFs, data entry controls cover dates, times, integer/real number formats, value intervals for numbers, pull-down list for libraries of values and the triggering of warning messages.

The Capture System Import module allows the import of data as ASCII files into the eCRFs. The import follows a four-step process:

1. Configuration of the structure of the import file from a template file and definition of links with the patient code and CRF fields
2. Recording the files to import by sets (file import in sets: creation of a set and associated configuration; add, delete the fields to import)
3. Import of data into temporary ORACLE tables
4. Validation and transfer of data to the electronic CRF

Electronic monitoring is still used only rarely. CSMonitor supports monitoring (investigator centres, patients, SAEs, deviations, monitoring visit reports, phone contacts, etc.). It shows the state of trial progress and documentation on site (CV collection dates, protocol signature dates, potential protocol amendments and their signature dates, local ethical committees).

2.3.3 eResearch Network, TrialMaster and eClinical Suite

The data management system of eResearch Technology, Inc. is called eResearch Network. The main component of eResearch Network is the module eData Management, an internet-based tool for collecting, editing, and managing clinical trial data. The component eData Entry is an EDC system allowing the use of an Internet browser to input data into a centralized database in an online or offline environment. eResearch Technology, Inc sold its EDC division to OmniComm who absorbed eResearch Technology and markets it's own EDC system TrialMaster. After integration eResearch Network became eClinical Suite consisting of several modules:

- eClinical Portal (dashboard for study metrics)
- eData Capture (data capture system, monitoring tool)
- eData Management (data management system, CRF design, data archive)
- ad-hoc reporting (creation of reports and data lists) and others

2.3.4 CleanWEB

The Java-based CleanWEB by Telemedicine Technologies is an integrated solution with designer, connector and data collection by web browser. It offers different types of randomisation and on-line monitoring. CleanWEB can be licensed either for an autonomous exploitation by the customer (Integral licence), or on a per trial basis (Single Trials Licence); an additional partnership programme gives a flexible alternative to the Integral licence

scheme. Data capture is performed directly on the computer at the site, using a web browser. An integrated cache system optimizes data transmission during delays especially in case of the transfer of source data (X-Ray imagery data, ECG recordings). The design and implementation eCRF is supported by the availability of an enriched set of data entry fields (scores & grades, ICD, thesaurus) with on-line help and automated edit checks, and native multilingual user interface. Many types of randomisation mechanisms and automated alerts and reminders linked to the study timetable can be implemented.

2.3.5 ClinInfo

ClinInfo, formerly developed for the Clinical Pharmacology Service at University Claude Bernard Lyon, has already been used for international studies of large populations. ClinInfo is a scalable and easily adaptable system that can be adapted to all kind of studies to manage data from several dozen to tens of thousands patients. Users are for example: Centre Hospitalier Universitaire de Saint-Etienne, Ciba, Duphar, Genentech, Institut Henri Beaufour, Novartis and Synthélabo.

2.3.6 Other Systems Mainly Used in a Single Country

2.3.6.1 e-MedSolution

e-MedSolution by International System House Ltd. Budapest is a health care information system, used by Hungarian clinical trial centres. ECTrial is a clinical data base software.

2.3.6.2 ECTrial

ECTrial is a clinical data base software.

2.3.6.3 SINATRAS

SINATRAS is an EDC system developed by SAKK for the Clinical Trial Unit at the Inselspital in Bern (Switzerland). As a web-based solution, it is suitable for multi-centre studies. Validity checks on the eCRF forms can detect errors or outliers so that these can be corrected immediately. In addition, SINATRAS contains integrated monitoring functionalities with query administration functionalities. Data entry for the first trial using SINATRAS launched on May 1st 2008. The system highlights any missing or non-plausible data during input at the site. Data is immediately accessible to all actors (study centers, data managers, investigators, study monitors). Additional functions covered are query management, traceability of changes (audit trail), user administration, archiving of electronic CRFs and data export for statistical analysis (http://www.studycoordinationcenter.ch/en/clinical-trial-unit/datamanagement).

2.4 Use of Data Analysis Software for Clinical Data Management

Some centres are using data analysis software (SAS, SPAD by Decisia) or database software (Microsoft Access) for their clinical data management processes. SAS PheedIT is a web technology/SAS-based integrated solution with modules for study set-up, data entry, report generation, validation and data export. It allows for easy study design and set up for data entry. PheedIT may be linked to any web based information page for informing investigators. An export feature is used to transfer clinical study data into an analysis-friendly structure (analysis database). This database is then used by a report engine to generate standardized reports, analysis tables, graphs, etc. PheedIT has three operating modes. First, the Development Mode (study set up and test of any modifications to a study). An

administration tool is used to transfer metadata (such as modules, variables, etc.) into the Production Mode. In Production Mode the audit trail and different logs are automatically active and all activities carried out may be reported.

Even data mining software like SPAD, a suite for exploratory and predictive analysis, is used for data collection and data management in clinical trials. One centre is using Microsoft InfoPath in connection with Windows SharePoint Services to generate surveys by XML forms and distributes them over a network.

2.5 Open-Source Solutions for Clinical Data Management

Open-source solutions for clinical data management are of special interest in the academic community. The volatile market for EDC solutions, where software systems can suddenly disappear, are bought up to increase market share by a competitor, or are not developed further, let open-source software appear to be an alternative to commercial solutions that may be more suitable for academic research centres. Several open-source CDMS have been developed and are used in clinical research, though they make up only a small part of all solutions in use. Both commercial solutions as well as open-source solutions have to be system validated for GCP compliance before being employed for clinical trials.

2.5.1 PsyGrid/openCDMS

PsyGrid was developed for the data management of large trials of complex interventions in mental health and has been developed further for all sorts of trials. It has been renamed openCDMS (http://www.opencdms.org) and is now available under a free licence (LGPLv3).

The data management component features scheduled e-mail reminders for when data has to be collected in longitudinal studies, scheduled report generation and delivery by e-mail, an configurable data “review and approve” workflow for data management, and on-line or off-line data entry with automatic resynchronisation. The study definition component features a full study lifecycle management via easy to use graphical tools, including study versioning, publication and resynchronisation, a fully customisable data set definition including data elements, validation rules and scheduling, special derived responses that allow a value to be generated automatically by performing a calculation using previous responses as inputs, support document workflow, custom data validation rules and a flexible consent model that allows consent for each document to be configured in groups. A randomisation system with generic, configurable Stratified Random Permuted Block of Random Block Length algorithm is integrated, including the generation of randomizer statistics, and trial feasibility planning. System interoperability is based on modular web service architecture and even includes online randomisation with SMS/e-mail notification.

openCDMS is used by the UK Mental Health Research Network, the UK Diabetes Research Network and the National Institute of Health Research (UK).

2.5.2 GCP BASE

GCP BASE is a web-based tool for remote data capture for clinical trials developed at Mario Negri Institute for Pharmacological Research (Italy) and released as free software (GPL).

2.5.3 EpiData

EpiData consists of different modules: “Entry” can be used for data collection and “Analysis” performs basic statistical analysis and data management. EpiData software can be easily

installed. For example, it can be run from a USB stick. Development of EpiData Software is decided in consultation with an international group of persons and released by the (non-profit) EpiData Association in Denmark. Data export is possible in following formats: to Stata, SPSS and SAS with labels and missing value definitions, DBF, CSV (http://www.epidata.dk/).

2.5.4 OpenClinica

OpenClinica from Akaza Research LLC (https://community.openclinica.com) is a widely used open-source software for web-based EDC and data management. It facilitates study protocol configuration, data collection with eCRFs, supports 21 CFR Part 11 and other regulatory guidelines. OpenClinica has been installed by several ECRIN centres, for example by KKS Düsseldorf, KKS Essen and KKS Leipzig. The parameterisation and system validation turned out to be difficult, because it has to be performed without the support of a company. Also, the Centre for Sepsis Control and Care (CSCC) in Jena uses OpenClinica for clinical trials and for the Sepsis register for clinical studies [11]. The centre decided to use OpenClinica because the data model is conform to CDISC ODM (http://www.cdisc.org/odm), the software exhibits a complex right management and audit trail, CRFs are designed by using Excel templates, and query functionalities are available. Further lighlights of the product are the possibility to import clinical hospital data from a HIS as ODM or as web services and the linking to the CCTS Suite of caBIG (https://cabig.nci.nih.gov/adopt/CTCF) or the PID Generator of TMF e.V.. It was described negatively that the Patient Reported Outcome support is missing and no coding with MedDRA and visit planning is available.

The advantages of open-source solutions to support data collection in clinical trials have been well described [12]. The authors suggest that research organisations and funders should combine efforts to produce open-source solutions for trial data management. In this way, a shared platform could be easily established and could bring additional benefits like electronic submission to regulators, automated sharing of data, and automatic contributions to public databases such as pharmacovigilance and drug monitoring registries. Such an open-source system would have the potential to save money by eliminating the reliance of academic centres on using expensive database software systems and their administrators.

2.5.5 CliniTraq

CliniTraq is a new open-source, web-based clinical trial software system that provides comprehensive multi-study subject visit and specimen tracking. CliniTraq's electronic data capture is protocol-driven and can adapt to protocol modifications.

2.5.6 Commercial Systems and the CDMS Landscape

In many ECRIN centres clinical data systems are already in use for many years and ECRIN members have gathered experience using them. Often these systems are different from the ones pharma industry is using. It must be considered, that solutions used by pharma industry might not be optimal solutions for the support of international trials by academic centres. Nonetheless, a market analysis and an evaluation of future trends in the area of clinical data management solutions was conducted by ECRIN and the KKS Network [13]. Many systems were evaluated, amongst others Medidata (RAVE-Platform), PhaseForward (InForm), DataTRAK (Eclinical Suite), Formedix Ltd. (Origin, Transform, Express), Entimo AG (DARE, ePRO), Adept Scientific GmbH (StudyBuilder), ARC Seibersdorf Research GmbH (Research Network), COMMEDIA-Group GmbH & Co.KG (Profiler-RES), SecuTrial (iAS), XClinical GmbH (MARVIN), Majoro Infosystems (ClinAccess), Clinipace (TEMPO), EclinForce (SmartStudy), TranSenda (Clinical Trial Manager), ClickTrials (ClickFind), ClinSource NV (TrialXS), DataTrial Inc. (NowEDC). However for academic institutions in general, these

systems are often prohibitively expensive to implement. This may be the case in particular in the early phase trial setting, where the ratio of cost of set-up per patient is proportionally much higher since the accrual number is low.

The CDMS market consists of about fifty commercial solutions. Although, competition exists for industry clients, even small solution providers can find their niche, by offering innovative or specialised solutions. Life Science Insight [14] published an evaluation of the clinical data management system landscape and concluded that "trial sponsors are increasingly looking to replace home-grown systems" to be able to add much-needed functionality to their drug development IT systems. An inspection of the CS market yielded following main findings:

1. Companies placed in the leadership portion include: Phase Forward, Oracle Clinical, Nextrials, and Medidata, followed by Siebel and SAS
2. Technology users should determine the functionality that gives them the best results while carefully watching the financial and operational viability of chosen vendors (because many CDMS companies have already disappeared from the market)
3. Current technology leaders must assess their position and identify their weaknesses
4. System integrators and partners need to forge key alliances that will help them implement and customize solutions efficiently while meeting the business needs and budgets of a diverse group of pharma and biotech companies

Recently, more and more EDC providers have extended their systems of core functions of clinical data management with clinical trial and site management functionalities. Clinical trial management functionalities cover additional trial data such as documentation (protocols, case report forms, etc.), patient recruitment/enrolment, investigator relationship management (IRM), electronic monitoring, integrated reporting, site management, medication management and cost tracking, clinical data archiving, and adverse event reporting.

Vendors of CDMS develop their system mainly according to the requirements of their industry clients (e.g. product specific level for eCRF design) and only to a minor degree to the demands of academic research centres. Concerning research sites at academic and government institutions Hanover and Julian [15] have concluded that these institutions constitute only a secondary market for clinical trial software providers and because they lack economic importance. This may constitute an argument for a more prevalent use of open-source solutions in academic research.

2.5.7 State of the Art of Clinical Data Management Systems

Commercial CDMS are not only tailored mainly to the requirements of pharma industry, but the leading commercial products may cost academic centres about 200000 – 400000 Euro for the basic installation and additional 20% annual maintenance costs, resulting in additional 40000 Euro per year or more. However, one should consider, that maintenance by a provider means considerable support, including further development of the products, regular updates and a help desk. Often, the software provider will install the software at the centre and will conduct the first steps in system validation, including installation qualification and operational qualification. Users of open-source software must conduct these steps themselves.

Academic centres try to use clinical trials as a means to increase income and to build a steady portfolio of trials, but have to do this with small resources and with as little interruption in health care processes and research as possible. For academic centres, the priority must be to conduct clinical trials and not to maintain a computer centre. Hosted solutions and the Application Service Providing (ASP) concept might be a solution to this constraint. In addition, academic staff will often change employment, the person knowledgeable of the database or the CDMS may leave the centre, jeopardising the entire infrastructure. This detrimental effect may only be absorbed by an academic user community offering mutual

support and being able to set up a common knowledge base. Such a user community should be built around the CDMS solutions supporting its use by knowledge exchange and training.

In future, interoperability of CDMS will be the key issue for the advancement of software solutions. The situation is still so that a multitude of available software platforms designed to manage various aspects of clinical trials, work independently of one another, unable to efficiently communicate or share information [16]. The next step is to allow records and data to be linked and shared, and eventually to improve efficiency, increase safety and reduce costs. The use of common data standards will be the most important way to achieve interoperability of disparate systems in medical research. Even with data transfer standards in use, with the multitude of different EMR systems in hospitals and the complexities of clinical trials, it is currently not yet feasible for clinical trial sites to create the source document for each clinical trial within an EMR, or to use the EMR to export all of the clinical trial data directly into the sponsor's database, while at the same time maintaining the complete source documentation in electronic format [17].

Academic trials tend to be more complex than industry trials, requiring tight cooperation between investigators and data management and thus an efficient integration of the CDMS in the research processes. Tighter and more efficient cooperation may be achieved by joint development of a CDMS solution specifically adapted to the needs and work processes of academic research centres and offered as open-source supported by a user community. This approach would have the important advantage that such a developed CDMS could support standards and enable interoperability with other systems (e.g. biobanks, imaging) to a degree that industrial solutions cannot offer or can achieve only with considerable additional costs. Still, any solution used in clinical trials routinely should be GCP compliant.

2.6 Need for GCP-Compliant Clinical Data Management Solutions

The use of CDMS in clinical trials is subject to regulations on data management (e.g. 21 CRF Part 11, EU GMP Guideline Vol. 4, Annex 11 Computerized Systems, GCP, data protection laws, e-signature requirements). For clinical centres employing CDMS it becomes necessary to implement best practices for CRF design, query resolution, and study start-up, including user acceptance testing, system validation, creation of a data management plan and training of investigators in the use of the application, causing considerable pressure on resources. To be GCP compliant the data management of clinical trials must be supported by a quality system, to protect patients and to ensure that the collected data are correct. This quality system requires independent audits to determine whether data management activities are conducted correctly according to study protocols, standard operating procedures (SOPs), GCP and relevant regulations. A quality management system for data management is in place in over 90 % of centres that perform data management. GCP studies can only be conducted with a validated data management system. Still over half of the centres still need to conduct a complete system validation of their data management system. The reason for this lack is that a complete system validation is a major effort that costs considerable time and resources. To enable the data management processes for GCP compliant international clinical trials ECRIN will create dedicated and certified “ECRIN Data Centres” that provide data management services. An ECRIN standard for GCP compliant data management [18] has been developed for the certification process that can be used by every centre interested in examining the status of GCP compliance of it’s data management (Additional file 1: Standard requirements for GCP compliant data management in multinational clinical trials. http://www.trialsjournal.com/content/12/1/85/additional). The ECRIN standard includes 115 IT requirements, 107 data management requirements and thirteen other requirements. The IT requirements cover basic IT infrastructure, validation and local software development and support clinical trial management. The data management and other requirements cover the implementation of a specific clinical data management application, data management of trials across the unit, international aspects and the competence of a trials unit's staff. The standard

is intended to provide an open and widely used set of requirements for GCP-compliant data management for clinical trials, particularly in academic trial units.

An additional required quality process is study monitoring that includes activities, like initiation visits, monitoring visits during the trial and close-out visits. The monitor focuses on those trial data and study information that are essential for an assessment of trial participant's safety, well-being and rights, as well as the quality of collected data. It is often overlooked, that clinical site monitoring is one of the most costly parts of a clinical trial. ECRIN members develop the monitoring process further by using a risk-based approach [19]. To prevent a high workload by monitoring each site with the same profundity, a structured procedure for risk analysis in clinical trials and strategies for on-site monitoring adapted to identified risks was developed. In clinical research, it is the source data (medical charts, e.g.) that are transcribed onto paper CRFs or into the EDC system. During the trial, it is the investigator's responsibility to secure and maintain the source data, and the sponsor's responsibility to ensure the reported trial data are accurate, complete and verifiable from source documents. To ensure the accuracy of this process, monitors visit the sites to carry out Source Data Verification (SDV), in which CRF records are manually compared to the corresponding source data in the charts.

Enabling efficient adverse events reporting in clinical trials is one of the most important tasks conducting a trial. Especially two events: serious adverse reactions and suspected unexpected serious adverse reactions need special attention and must be reported within a fixed period. In some cases already special adverse events reporting tools are used.

Although highly sophisticated, web-based EDC systems still show sometimes impairments [20], especially, systems are not robust enough to handle the workload at the investigator site (e.g. slow web page refreshing), systems support different versions of basic software (e.g. internet browsers), CRF pages are not displayed correctly (e.g. missing data field boxes) and systems often may generate unnecessary queries. Often monitors receive complaints by investigators about lack in performance of the CDMS interface. Therefore, quality management, validation, best practices and training are important support activities for the conduct of clinical trials and especially for the step of data collection. Unfortunately, the uses of CDMS that are able to streamline data processing for the sponsor impose considerable burdens on the research sites at the investigator's working place. Considering the necessary resources for recruiting investigators and the low retention rate, usability of CDMS solutions in the clinical working area of investigators has to be improved (for example: reduction of redundant data input) and more has to be done to make using electronic data capture tools more attractive to investigators.

Increasingly imaging biomarkers (a characteristic that is measured by an imaging technique) and surrogate endpoints have shown to facilitate the use of small group sizes in clinical trials, obtaining faster results with good statistical power. Imaging used in clinical trials is able to reveal subtle change that is indicative of the progression of therapy that is missed out by more traditional approaches. Even statistical bias may be reduced as after the image is taken with contact to the patient, the findings based on the image can be evaluated without direct patient contact. For example, results of the measurement of tumour shrinkage are a commonly used surrogate endpoint in solid tumour response studies to assess the effects of anticancer drugs. In spite of these advantages, an integrative solution with the connection of imaging software with the CDMS is still missing.

Quality of Life (QoL) assessment by the patient has become an often used outcome in many controlled trials and other studies. Recently electronic diaries are used to record patient's responses with small, portable electronic devices and mobile phones. Patient's responses are collected, stored and transmitted to a computer, where data from several devices is pooled and analyzed. Electronic diaries can automatically record date and time for each entry as it is made, and they can make it very simple to enter a data point. Yet, QoL data collection and CDMS often run in parallel processes and still few integrated solutions exist.

2.7 Trial Closing

After all patients have been examined, the clinical trial sites are closed and the data base is locked. This means that data cannot be added or updated anymore. The collected data is sent to the sponsor or leading investigator for analysis. The finalisation of the trial includes the creation of a statistical report with the analysis of all results according to the statistical analysis plan, and the preparation of the final trial report.

The end of the trial is reached with the archiving of the essential clinical study documents. Filled with documents during the conduct of the trial, the Trial Master File contains at the end all GCP-essential documents, including the study database, and must be archived by the sponsor. The Investigator Site Files in the different trial centres contain the essential documents necessary for the investigator to carry out the study, and is archived at the corresponding clinical centre. An archive for clinical study documents must be lockable, accessible only by authorised persons and must protect documents against water and fire. Electronic documents and data should be archived on durable media in an open standard format which is independent of specific operating systems, applications and special equipment. This requirement ensures that study documents can be kept readable for long periods without the original data generating system available.

For documents XML, TIFF and PDF and for the database CDISC ODM are such suitable formats. The ODM is designed to facilitate the interchange and archiving of the metadata and data for clinical research. The export of trial data in ODM format is especially suited for archiving clinical trials, because among other features ODM contains the entire clinical trial data (including metadata) and full audit trail information. Two recent developments in CDISC will affect the archival of studies: the further development of ODM in the area of “eCRF submission” and the use of “electronic Source Data” and thus the necessity to archive this source data [21]. Unfortunately, many CDMS cannot export the clinical trial database in ODM but use other or proprietary formats. For electronic archiving over longer storage periods, data migration strategies may be necessary to guarantee readability of trial data.

2.8 References

[1] Kuchinke W, Ohmann C (2009) A technical introduction to the basic requirements of clinical trials. EJHP Practice, 15, pp. 20-22

[2] ICH-GCP E6 (R1) Guideline for Good Clinical Practice. (CPMP/ICH/135/95). Online available: http://www.ema.europa.eu/pdfs/human/ich/013595en.pdf

[3] Ohmann C, Kuchinke W (2007) Meeting the challenges of patient recruitment. Int J Pharm Med, 21 (4): 263-270

[4] Corrie P, Shaw J, Harris R (2003) Rate limiting factors in recruitment of patients to reduce time to market in the pharmaceutical industry. BMJ; 327, (7410): 320-321

[5] El Emam K, Eng B, Jonker E, et al. (2009) The Use of Electronic Data Capture Tools in Clinical Trials: Web-Survey of 259 Canadian Trials. J Med Internet Res; 11(1):e8. Online available: http://www.jmir.org/2009/1/e8

[6] Demotes-Mainard J, Brachet R, Kubiak C (2009) Integrating clinical research in Europe: the European Clinical Research Infrastructures Network EMBnet.news. 15(2). Online available: http://journal.embnet.org/index.php/embnetnews/article/view/5/35

[7] Kuchinke W, Ohmann C, Yang Q, et al. (2010) Heterogeneity prevails: the state of clinical trial data management in Europe - results of a survey of ECRIN centres. Trials, 11:79. Online available: http://www.trialsjournal.com/content/11/1/79

[8] Nelausen K, Hansen H (2000) A clinical Internet trial management system – CITMAS Technology and Health Care - Special issue on abstracts of MEDNET 2000. 5th world congress on the Internet in medicine, Brussels, archive 8(3-4)

[9] Jarden M (2009) Hematopoietic stem cell transplantation. The Effect of a Multimodal Intervention on Physical Capacity and Functional Performance, Treatment-related Symptoms, and Quality of Life, Ph.D Dissertation at the Faculty of Health Sciences, University of Copenhagen

[10] Fischer A, Müller T (2007) Electronic data capture - MACRO, LeukemiaNet

[11] Löbe M, Meinecke F (2011) Open source clinical data management with OpenClinica. 56. GMDS Jahrestagung, Mainz, Abstract Band, Johannes Gutenberg-University, GMDS, Germany. pp. 532-533

[12] Fegan G, Lang T (2008) Could an open-source clinical trial data-management system be what we have all been looking dor? PLOS Medicine. 5(3):347-349

[13] ECRIN-TWG (2008) Deliverable D11. Identification, evaluation and prioritization of possible common or compatible GCP-compliant data management tools for multination. trials

[14] Hanover J (2005) Operating costs in clinical trials force CROs to work with EDC. Life Science Insights, IDC

[15] Hanover J, Julian E (2005) Competitive analysis U.S. Clinical Trial Management Systems 2004 Vendor Analysis. Leadership Grid and Market Shares

[16] Gaves P, Clewlow A (2009) Trends and Issues in an Electronic Clinical Data Management World. Touch Briefings 2009, eClinical visions, pp. 1-5

[17] Mitchel J, Kim Y, Choi J, Park G, Suciu L, Horn M (2010) The final eFrontier. Applied Clinical Trials

[18] Ohmann C, Kuchinke W, Canham S, et al. (2011) Standard requirements for GCP-compliant data management in multinational clinical trials. Trials, 12:85, online available: http://www.trialsjournal.com/content/12/1/85

[19] Brosteanu O, Houben P, Ihrig K et al. (2009) Risk analysis and risk adapted on-site monitoring in noncommercial clinical trials. Clinical Trials; 6: 585–596

[20] Borfitz D (2009) Investigative sites: the trouble with e-Clinical technologies, eCliniqua

[21] Kuchinke W, Aerts J, Semler S, Ohmann C (2009) CDISC standard-based electronic archiving of clinical trials. Methods Inf Med. 48(5):408-13

3 Clinical Research Standards

3.1 Clinical Research Standards Harmonisation Recommendations

The International Conference on Harmonisation (ICH) (http://www.ich.org) Multi-disciplinary Group 2 (M2) Expert Working Group (EWG) was established during the ICH meeting 1994 in Brussels to facilitate international electronic communication by evaluating and recommending open and non-proprietary Electronic Standards for the Transfer of Regulatory Information (ESTRI). Some activities of the EWG resulted in recommendations, the summary of them (last updated on 30.06.2011) are presented on the table bellow and can be downloaded at http://estri.ich.org/recommendations.

Category	Title	Version	Date Endorsed	Implemented			
				EU	Japan	US	Canada
General	Procedure	2.0	November, 2005	Yes	Yes	Yes	Yes
General	ESTRI Gateway	2.0	November, 2005	Yes	Yes	Yes	Yes
File Format	PDF	2.0	April, 2011	Yes	Yes	Yes	Yes
File Format	XML	1.0	November, 2005	Yes	Yes	Yes	Yes
Information Transfer	EDIINT AS1/AS2	2.2	June, 2010	Yes	Yes	Yes	Yes
File Integrity	MD5	1.0	June, 2010	Yes	Yes	Yes	Yes

One of the additional requirements identified by ICH is the availability of “an adequate number of qualified staff and adequate facilities for the foreseen duration of the trial to conduct the trial properly and safely”. To assure patient safety and high quality in clinical research programs, the Medical Executive Committee of the NIH Clinical Center (USA) has developed the following essential standards for performing clinical research (http://www.cc.nih.gov/ccc/clinicalresearch).

3.1.1 Clinical Informatics, Data Management and Protocol Tracking

“Each institute sponsoring clinical research should develop a central clinical investigations database that maintains all data specified to be collected in the clinical study (either intervention or natural history). The clinical research information system being continually developed by the clinical center interfaces with and supports each institute’s clinical research needs. A confederated database will enable information exchange, enabling access to and sharing of clinical and research information among all institutes. The institutes require data-management infrastructures to maintain their central data registries, to enhance existing databases, to provide eligibility checklists, to record patient randomization and entry into their protocols, to provide report generation, data warehousing, and data entry forms, and to monitor data collection.”

3.1.2 Biostatistics Support

A qualified biostatistician must review all clinical protocols before approval implementation.

3.1.3 Quality Assurance and Quality Improvement

Each institute must have access to a quality assurance/improvement program with infrastructure that ensures that clinical trials are monitored adequately and centrally. The institute should determine the appropriate extent and nature of monitoring. This determination should be based on considerations of the study objectives, purpose, design, complexity, blinding, size, and endpoints, and should include the following:

- Onsite protocol monitoring during clinical trials. Statistically controlled sampling is an acceptable method for selecting the data to be verified. For interventional trials, the institutes should demonstrate a capacity to review a minimum of 10% of patient records on selected clinical trials to assure data accuracy, protocol compliance, and adherence to regulatory requirements.
- Access to an independent Data Safety and Monitoring Board (DSMB) for at least a semi-annual overview of all randomized blinded studies.

3.1.4 Protocol Review

Each institute must provide a scientific review by a written protocol review process and infrastructure (e.g. administrative staff) to support an appropriately constituted Institutional Review Board (IRB).

3.1.5 Human Resources and Physical Plant

Necessary personnel, office space proximal to patient care areas, and accompanying resources should be available to support the clinical research infrastructure.

3.1.6 Training and Education

All clinical investigators (PIs/AIs) are required to take a training course, or equivalent, on the roles and responsibilities of clinical investigators. The clinical centre provides this course.

All IRB chairs and members (including lay members) receive orientation materials and are required to take specialized training modules provided by the clinical centre/Office of Human Subjects Research (OHSR). Continuing education will be provided for IRB members

3.1.7 Research Participants

The organization will provide participants (and communities) with appropriate educational materials about the clinical research process (clinical centre patient education materials; protocol specific patient education materials, Patient Recruitment and Public Liaison).

Each informed consent document will provide participants with information about how to voice concerns or discuss problems related to their protocol participation (Patient Representative and study team members).

The organization periodically assesses participants' perceptions of the clinical research experience and uses this data to drive improvement activities (patient surveys and portal).

3.2 HITSP Clinical Research Interoperability Specification

By following the focus on international interoperability specifications for p-medicine project would be recommended the Healthcare Information Technology Standards Panel (HITSP) (http://www.hitsp.org) state of the art specifications. HITSP is a cooperative partnership between the public and private sectors from the United States. The Panel was formed for the purpose of harmonizing and integrating standards that will meet clinical and business needs for sharing information among organizations and systems.

The Clinical Research Interoperability Specification (IS 158) covers clinical research in all its forms as it interoperates with healthcare systems. The specification spans two industries, healthcare and clinical research, and incorporates standards from healthcare (HL7 and IHE) and research (CDISC). The design leverages existing HITSP constructs and communication methodologies where applicable, and lays out new constructs as needed. The design also leverages the current players in the clinical research industry such as Electronic Data Capture (EDC) systems and research registries. Close to the detailed clinical research case studies descriptions, HITSP presents the information exchange requirements and the design specification. All above specifications are available for download on the HITSP web site.

3.3 eSource Data Interchange (eSDI) Document

The eSDI document (http://www.cdisc.org/esdi-document) is the product of the CDISC eSDI Initiative, the purpose of which was “to investigate the use of electronic technology in the context of existing regulations for the collection of eSource data (including that from eDiaries, EHR, EDC) in clinical trials for regulatory submission by leveraging the power of the CDISC standards, in particular the Operational Data Model (ODM)”.

The overarching goals are to make it easier for physicians to conduct clinical research, collecting data only once in an industry standard format for multiple downstream uses, and thereby to improve data quality and patient safety. The eSDI Document includes:

- An extensive review and analysis of the relevant existing regulations
- Twelve requirements for conducting regulated clinical research using eSource data collection in the context of existing regulations
- Five potential scenarios, three of which include the use of electronic health record systems and associated benefits of standards
- An appendix on responsibilities of each of the various functional groups conducting clinical research
- A template for evaluating an eSource data collection process versus the requirements
- A good practices checklist for investigators

As response to eSDI document, European Medicines Agency (EMA) published an important paper [1], named “Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials” which refers to / describes

- The requirements of Directive 95/46/EEC [2] on the protection of individuals regarding the processing of personal data and the free movement of such data and the considerations set out in this paper should be followed
- Clear requirements for electronic source and transcribed data need to be stated in a way that the processes can be used and accepted with confidence when such requirements are complied with, and the benefits of the systems can be fully realised
- The current expectations of GCP inspectors (any departure from this paper would need to be justified)

3.4 Good Clinical Practice Compliance

According to EMA website, Good Clinical Practice (GCP) is an international ethical and scientific quality standard for designing, recording and reporting trials involving human subject participation. Compliance with this standard provides public assurance that

- The rights, safety and wellbeing of trial subjects are protected
- The clinical trial data are credible

The protection of clinical trial subjects is consistent with the principles set out in the Declaration of Helsinki (http://www.wma.net/en/30publications/10policies/b3). This is a statement of ethical principles developed by the World Medical Association. Requirements for the conduct of clinical trials in the European Union (EU), including GCP and good manufacturing practice (GMP) and GCP or GMP inspections, are implemented in:

- The Clinical Trial Directive (Directive 2001/20/EC [3])
- The GCP Directive (Directive 2005/28/EC [4])

The basic principles of GCP [5] are mentioned bellow:

1. Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki and that are consistent with GCP and the applicable regulatory requirement(s).
2. Before a trial is initiated, foreseeable risks and inconveniences should be weighed against the anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks.
3. The rights, safety and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society.
4. The available nonclinical and clinical information on an investigational product should be adequate to support the proposed clinical trial.
5. Clinical trials should be scientifically sound and described in a clear, detailed protocol.
6. A trial should be conducted in compliance with the protocol that has received prior Institutional Review Board (IRB)/Independent Ethics Committee (IEC) approval/favourable opinion.
7. The medical care given to, and medical decisions made on behalf of, subjects should always be the responsibility of a qualified physician or, when appropriate, of a qualified dentist.
8. Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).
9. Freely given informed consent should be obtained from every subject prior to clinical trial participation.
10. All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification.
11. The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).
12. Investigational products should be manufactured, handled, and stored in accordance with applicable Good Manufacturing Practice (GMP). They should be used in accordance with the approved protocol.
13. Systems with procedures that assure the quality of every aspect of the trial should be implemented.

3.5 State of the Art Related to Clinical Research Data Standards

The top discussion forums for moving toward clinical research data standards that support applied uses are the Clinical Data Standards Interchange Consortium (CDISC) and the Regulated Clinical Research (RCRIM) Technical Committee of Health Level Seven (HL7) [6].

CDISC and RCRIM will be in details presented in further deliverables, at this stage we would mention that above clinical research data standards are heterogeneous due to differences in terminologies, message/information structure, data types, etc. Nevertheless, an effort to harmonize the CDISC and HL7 models has been identified. The Biomedical Research Integrated Domain Group (BRIDG) Model (http://www.bridgmodel.org) is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC), the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), the National Cancer Institute (NCI) and its Cancer Biomedical Informatics Grid (caBIG), and the US Food and Drug Administration (FDA). The BRIDG model is an instance of a Domain Analysis Model (DAM). The goal of the BRIDG Model is to produce a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artefacts. This domain of interest is further defined as: Protocol-driven research and its associated regulatory artefacts: i.e. the data, organization, resources, rules, and processes involved in the formal assessment of the utility, impact, or other pharmacological, physiological, or psychological effects of a drug, procedure, process, or device on a human, animal, or other subject or substance plus all associated regulatory artefacts required for or derived from this effort, including data specifically associated with post-marketing adverse event reporting

3.6 References

[1] GCP Inspectors Working Group (GCP IWG) (2010) Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials, 09 June 2010, EMA/INS/GCP/454280/2010

[2] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23.11.1995, pp. 31–50

[3] Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001, Official Journal of the European Communities, 2001

[4] Commission Directive 2005/28/EC of 8 April 2005, Official Journal of the European Union, 2005

[5] ICH Topic E 6 (R1) Guideline for Good Clinical Practice, Note for guidance on clinical practice, EMA, July 2002, CPMP/ICH/135/95

[6] Richesson R, Krischer J (2007) Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions, J Am Med Inform Assoc. 14(6): 687–696.

4 VPH Modelling and the Integrated Oncosimulator

This chapter consists of two parts. The first part (4.1) provides a general overview of the VPH NoE (Network of Excellence on the Virtual Physiological Human) that aims, inter alia, at ensuring compatibility among various VPH oriented research projects that are partly funded by the European Commission. The second part (4.2) provides a comprehensive review of the literature pertaining to the mathematical and computational modelling of tumour growth and treatment response and more broadly to multiscale cancer dynamics. Special emphasis lies on in silico oncology as this constitutes the context for the Oncosimulator development and several clinically oriented cancer biomechanism models in the frame of p-medicine project.

4.1 General Overview of the VPH NoE Project

4.1.1 Background

The VPH NoE project started in June 2008 and funding currently continues until November 2012. It is coordinated at University College London (UCL), with twelve partners from the UK, France, Germany, Spain, Belgium, Sweden and New Zealand. Its aims range from the development of a VPH Toolkit and associated infrastructural resources, integration of models and data across the various relevant levels of physiological structure and functional organisation, through to VPH community building, training activities and support.

4.1.2 Philosophy

One of the key challenges in the development of quantitative, integrative and predictive models describing human physiology is to provide the necessary research infrastructure. This includes methodologies, databases and computational tools to allow scientists working in different scientific fields (at various physiological levels and scales) to communicate, exchange data and technologies in a standardised manner. The scale of data to be generated, processed, and exchanged requires software tools and massive computer storage currently not widely available. Dissemination is another key challenge, as the VPH NoE scope is by definition multidisciplinary and only a very limited number of journals currently accept physiome-related papers. Scientists able to deal with multidisciplinary topics are required, necessitating training of multidisciplinary individuals and VPH specialists.

4.1.3 Objectives

The main objectives of the network are:

- Inter-institution and interdisciplinary research projects
- Development of the VPH Toolkit, a shared and mutually accessible resource
- Facilitation of development of horizontal and vertical model/data integration
- Development of interdisciplinary training activities and VPH careers
- Establishing a core set of VPH-related dissemination and networking activities
- Implementation of key working groups to pursue VPH integration research worldwide
- Creation of industrial, clinical and scientific advisory boards for consultation

4.1.4 Work Packages

The NoE consists of five work packages. One manages and coordinates the network, and the remaining four are described below.

4.1.4.1 Exemplar Projects

The NoE develops Exemplar Projects (EP) working towards integration amongst VPH researchers to address specific research challenges. The aim is to provide solid examples of horizontal and vertical model/data integration, that may only be achieved through integrating disparate knowledge and research infrastructures. This “infrastructure” is provided through development of the VPH Toolkit, detailed in the next section. Nine integrative interdisciplinary EPs have been chosen which make use of, or contribute to, the VPH Toolkit, through use or development of modelling, simulation and visualisation-related tools, data or methods:

1. A multi-organ Core Model of arterial pressure and body fluids homeostasis
2. Integrated multi-level modelling of the musculoskeletal system
3. Fighting aneurysmal disease (FAD)
4. Multi-scale simulation and prediction of the drug safety problems related with hERG
5. Digital Patient Working Group: Modelling and visualising brain function and pathophysiology
6. Establishing ontology-based methods for the VPH ToolKit to improve interoperability between data and models: the Guyton case study
7. Integrating genetic theory/genomic data with multiscale models in population context
8. The NoE, Infrastructure and the Challenge of Call 6
9. Execution of medical image simulation workflows on DEISA through workflow interoperability between the Virtual Imaging Platform and the VPH toolkit

4.1.4.2 Toolkit

The VPH NoE aims to develop, evolve and promote standardised markup languages which permit interoperability of models and, where this is appropriate, interoperable codes which may be coupled both horizontally and vertically. Standards developed need to be suitable for, adopted by, and adhered to not only within the European VPH initiative, but also on a global basis, for example via interaction with the international Physiome Project. The VPH Toolkit provides a means to ensure that all VPH-funded projects are able to work towards this aim.

Particular technological foci for development are:

- Open markup language (XML) standards for describing data and models at spatial scales that range from proteins to the human organ
- Application programming interfaces/libraries for implementing these VPH standards
- Workflows that use existing middleware for facilitating grid-enabled VPH research
- Web-accessible repositories for data, models and workflows based on the VPH standards and including annotation and tutorials for non-expert biologist users
- Libraries of open-source computational routines and graphical user interfaces (GUIs) that, via the APIs, can access the data and model repositories

The ToolKit will be developed through the creation, accumulation, and curation of VPH research-related “capacities” – the integration of existing work, and its further development

towards greater interoperability. A companion website, the VPH ToolKit portal (http://toolkit.vph-noe.eu), plays a central role in this effort. It provides a knowledge base of the "capacities" available, whether these be specific tools, methods for conducting VPH research in an integrative fashion, or services available to researchers. It thus enables researchers to find technologies easily that may be of relevance to them, rather than re-inventing the wheel. It also provides a structure that can help to place individual activities in their correct context within the VPH initiative as a whole.

Guideline documents are now available (http://toolkit.vph-noe.eu/toolkit-guidelines) to assist groups preparing content for submission to the ToolKit in ensuring that their submissions are of the highest quality. The eight key topics cover

- Optimising the submission of tools, models and data (separately)
- Understanding and respecting ethical constraints, approaches to licensing
- Improving interoperability, maximising usability, and the use of ontological annotation

All eight documents were released in March 2011 (update with examples in early 2012).

4.1.4.3 Training and Career Development

The NoE addresses training and career development for both early and in-career VPH researchers. Activities also pay special attention to the outcomes generated from other VPH-related projects and existing EC-initiatives (e.g. Marie Curie) to ensure complementarity with existing activities. The goal of this task is the design and implementation of actions directed at the development of VPH research education and careers. Key tasks are:

- VPH Integrative Study Plan - assessment of NoE partner course/teaching capabilities and requirements
- VPH Industrial and Clinical Careers Assessment
- VPH Training scheme
- VPH Study Groups
- Analysis/promotion of mobility schemes for VPH researchers
- Formal Strategy Document advising on VPH training and careers
- VPH Educational Materials, including VPH Textbook

The training objectives of the VPH NoE are two-fold. The first one is to provide specific training on VPH-related tools, data, data-handling and modelling. The second is to foster institutional support within VPH NoE member-institutions to create an interdisciplinary European-wide study programme. Such activities form an important part of a wider, pan-European process directed towards introducing systematic educational activities with the aim to ensure that academia, medicine and industry in Europe have a workforce equipped appropriately to meet the possibilities offered by this new and important discipline. Training activities happened in Nottingham June/July 2009 and Barcelona September/October 2010.

4.1.4.4 Spreading Excellence

VPH NoE adopts an innovative approach towards enabling VPH research efforts both within and beyond the European research community. The impact of VPH NoE initiatives relating to VPH Exemplars, the VPH ToolKit, and interdisciplinary training is to be maximised. Further, emphasis is placed on developing clear, consistent lines of communication and information dissemination within and beyond the VPH NoE itself, crucial to the ongoing success of the VPH initiative as a whole. The main dissemination activities in the first year are:

- Construction of the VPH NoE public web site
- Elaboration of initial dissemination material (presentations, flyer, poster)
- Organisation of VPH NoE events
- Creation and animation of the external Advisory Boards (clinical and industrial)
- Creation and governance of the network's Editorial Board
- Publication of the scientific print media, including VPH special issues
- Identification of the VPH NoE Working Groups
- Production of the VPH Roadmap

4.1.5 Formal Interactions Between p-medicine and the VPH NoE

The VPH Toolkit and VPH-Share, another VPH project, will interact with p-medicine in several key areas, with the following deliverables:

- Requirements for enhancing VPH models for clinical decision support (1/2012)
- Data Warehouse stores ontologically annotated clinical, patient and simulation data, sharing cloud-based solutions with VPH-Share (report 9/2011, integration 9/2014)
- Workbench contributes tools to and use tools from the VPH Toolkit and set a collaboration exchange mechanism (Specification 1/2012)
- Clinical Trials uses and validates VPH tools and adapt them for clinical use (9/2013)
- VPH Modelling and Integrated Oncosimulator models which satisfy major VPH compatibility requirements (9/2014)
- Patient empowerment tool to monitor and implement donors' wishes (1/2013) and an interactive tool to support empowerment (7/2013)
- Education/training tutorials and eLearning tools submitted to the VPH Toolkit (1/2013)

p-medicine may draw heavily upon the models developed in the NoE, and aims to make them available to clinicians. p-medicine will also share strategies for use of cloud technologies with other VPH projects, most notably VPH-Share. Tools developed during the p-medicine project should aim to adhere to guidelines set by the NoE, and they should be made available to the community through the VPH toolkit website, where appropriate.

4.1.6 Conclusions

The primary resources offered by the VPH NoE that might be useful to p-medicine are:

Resource	**Status**
VPH Toolkit Portal	Mature, fully featured and well-used, 57 tools are registered with the most popular getting over 300 "hits".
VPH Toolkit guidelines	First published in an unfinished form in March 2011, due to be finished early 2012. The documents are being continuously revised and feedback is sought. Guidelines might be followed selectively.
Industrial,	These boards have been combined due to the complexity of convening

Clinical and Scientific Advisory Boards for consultation	multiple boards with many eminent authorities. Peter Kohl from Oxford University was originally managed the board, but a recent move to Imperial College reduced his involvement. Since the board is in flux, it is unlikely that it might be consulted in the short term, but in the longer term this might be a useful resource for p-medicine.
Training activities	Several training activities related to VPH have taken place in the past. Future activities might incorporate p-medicine training to extend reach of p-medicine.
VPHI Institute	Newly established and thus currently not a mature resource, however it will help expand the reach of p-medicine and ensure the longevity of its products.

4.2 Multiscale Cancer Modelling and the Oncosimulator

4.2.1 Brief Generic Literature Review

The last few decades have witnessed an increased interest of the scientific community into the development of computational models for simulating tumour growth and response to treatment. At the beginning of the era of personalized medicine, sophisticated multiscale models yield valuable quantitative insights into complex mechanisms involved in cancer and may ultimately contribute to patient-specific therapy optimization. Comprehensive literature reviews are available inter alia in [1-7].

From the mathematical standpoint the major cancer modelling approaches can be distinguished into predominantly continuous and predominantly discrete although sometimes the boundaries of such a distinction are not very clear. Thus the character of several modelling approaches could (also) be considered hybrid.

Predominantly continuous models rely primarily on differential equations to describe processes such as diffusion of molecules, changes in tumour cell density and invasion of tumour cells into the surrounding tissue [8-13].

Predominantly discrete modelling considers several discrete states in which cells may be found and possible transitions between them, governed by “decision calculators”, such as cytokinetic diagrams and agent-based techniques [1,2,4,14-29]. Discrete models are usually represented by cellular automata of several forms and variable complexity (grids of cells or groups of cells, in which a finite number of states and a set of evolution and interaction rules are defined). Due to the hypercomplexity of cancer-related topics, each modelling approach is intrinsically able to satisfactorily address only some of the aspects of this multifaceted problem. As far as clinically-oriented cancer simulation models are concerned, their ultimate goal is their eventual translation into clinical practice, which entails

- Thorough sensitivity analysis to both comprehend and validate their behaviour and also gain further insight into the simulated mechanisms, in a more quantitative way
- An adaptation and validation process based on real clinical data

4.2.2 Discrete Entity-Based Cancer Simulation Technique (DEBCaST)

Most cancer modelling techniques developed up to now adopt the straightforward bottom-up approach focusing on a better understanding and quantification of rather microscopic tumour dynamics mechanisms and the investigation of crucial biological entity interdependences including, inter alia, tumour response to treatment in the generic investigational context. To this end several combinations of mathematical concepts, entities and techniques have been developed and/or recruited and appropriately adapted. They include, inter alia, population

dynamics models [30-33], diffusion related continuous and finite mathematics treatments [8,9,34-45], cellular automata and hybrid techniques [14-16,46-55], agent-based techniques [56,57] etc. Additionally, a number of bulky clinical tumour models focusing mainly on invasion and tumour growth morphology rather than on tumour response to concrete therapeutic schemes as administered in the clinical setting have appeared. Finite difference and finite element-based solutions of the diffusion and classical mechanics equations constitute the core working tools of the corresponding techniques [34,35,58].

However, a number of concrete and pragmatic clinical questions of importance cannot be dealt with neither by the bottom-up approach nor by the morphology oriented bulky tumour growth models in a direct and efficient way. Two examples of such questions are the following [59,60]: Can the response of the local tumour and the metastases to a given treatment be predicted in size and shape over time?, What is the best treatment schedule for a patient regarding drugs, surgery, irradiation and their combination, dosage, time schedule and duration? A promising modelling method designed with the primary aim of answering such questions is the Discrete Event-Based Cancer Simulation Technique (DEBCaST) [1,2,17-26,61-73]. DEBCaST is basically a top-down biomodelling approach in the sense that macroscopic data, including inter alia anatomic and metabolic tomographic images of the tumour, provide the framework for the integration of available and clinically trusted biological information pertaining to lower and lower biocomplexity levels such as clinically approved histological and molecular markers. However, DEBCaST also provides a powerful framework for the investigation of multiscale tumour biology in the generic investigational context.

From the mathematical standpoint DEBCaST is primarily a discrete mathematics method, although continuous mathematics (continuous functions, differential equations) are used in order to tackle specific aspects of the models such as pharmacokinetics and cell survival probabilities based on pharmacodynamical and radiobiological models. Adoption of the discrete approach as the core mathematical strategy of DEBCaST has been dictated by the obvious fact that from the cancer treatment perspective it is the discrete (i.e. the integer) number of the usually few tumour cells surviving treatment and their discrete mitotic potential categorization (stem cells, progenitor cells of various mitotic potential levels and differentiated cells) that really matters. These discrete entities and quantities in conjunction with their complex interdependences may give rise to tumour relapse or to ensure tumour control over a given time interval following completion of the treatment course. Cell cycle phases have a clearly discrete character too. Also, the properties of the different cell phases may vary immensely from the clinical significance perspective. A classic example is the lack of effect of cell cycle specific drugs on living tumour cells residing in the quiescent G0 phase.

It is noted that complex interdependencies of microscopic factors in the surrounding milieu of the cells such as oxygenation, nutrient supply and molecular signals emitted by other cells play a critical role in the mitotic fate of tumour cells. Their effect is taken into account in DEBCaST through the local mean values of the corresponding model parameters. To this end imaging, histological and molecular data is exploited as will be described further down.

Due to the numerical character of the method a careful and realistically thorough numerical analysis concerning consistency, convergence and sensitivity/stability issues is absolutely necessary before any application is envisaged.

Tumour neovascularisation is taken into account in an indirect yet pragmatic way by exploiting grey level and/or colour information contained within slices of tomographic imaging modalities sensitive to blood perfusion and/or the metabolic status of the tumour. [1,2,16,17,25,74]. The reason for adopting the above mentioned strategy rather than developing or integrating detailed tumour angiogenesis models is that no microscopic information regarding the exact mesh of the neovascularization capillaries throughout the tumour can be currently extracted from clinically utilized imaging modalities. Nevertheless, the microscopic functional capillary density distribution over the tumour can be grossly

estimated based on various imaging modalities such as T1 gadolinium enhanced MRI in the case of glioblastoma multiform (GBM) and arterial spin labelling (ASL) MRI.

Precursors of DEBCaST can be traced in the well-established and clinically applicable disciplines of pharmacology and radiobiology. Integration of molecular biology in DEBCaST may be viewed as the introduction of a perturbator or adaptor of the cellular and higher biocomplexity level parameters. In such a way in vivo measurable clinical manifestation of tumour dynamics is placed in the foreground. This is one of the reasons why DEBCaST is gaining wider and wider acceptance within the clinical and the industrial environment including the emergent domain of in silico oncology [1,4,59,64,75-78]. Both the large scale European Commission (EC) funded research and development (R&D) project ACGT (http://www.eu-acgt.org) and ContraCancrum (http://www.contracancrum.eu) have adopted DEBCaST as their core cancer simulation method. It is worth noting that in both projects the role of clinicians is prominent. A biomedical engineering concept and construct tightly associated with DEBCaST, the Oncosimulator, which is currently under clinical adaptation, optimization and validation is sketched below.

In [22], in order to convey the core philosophy of the method to the reader in a concise way a symbolic mathematical formulation of DEBCaST in terms of a hypermatrix and discrete operators is presented. Two specific models of tumour response to chemotherapeutic and radiotherapeutic schemes are briefly outlined to exemplify DEBCaST's application potential. The above article/chapter concludes by discussing several critical aspects including numerical analysis, massive parallel code execution, associated technologies, extensions and validation in the frame of clinico-genomic trials and future challenges and perspectives.

An encouraging fact as far as industrial and eventually clinical translation of the method is concerned is that both DEBCaST and the Oncosimulator have been selected and endorsed by a worldwide leading medical technology company (Philips Research) and now constitute modules of their research and development line (ContraCancrum project). One of the envisaged final products of this endeavour is a radiotherapy treatment planning system based on both physical and multiscale biological optimization of the spatiotemporal dose administration scheme. A clinical trial-based validation process for the system is currently at the stage of its detailed formulation [79].

4.2.3 Oncosimulator

The Oncosimulator is at the same time a concept of multilevel integrative cancer biology, a complex algorithmic construct, a biomedical engineering system and eventually in the future a clinical tool which primarily aims at supporting the clinician in the process of optimizing cancer treatment in the patient individualized context through conducting experiments in silico, i.e. on the computer. Additionally it is a platform for simulating, investigating, better understanding and exploring the natural phenomenon of cancer, supporting the design and interpretation of clinicogenomic trials and finally training doctors, researchers and interested patients alike [4,19,62]. The notion, core architecture and several implementations of the Oncosimulator have emerged within the In Silico Oncology Group, Institute of Communication and Computer Systems, National Technical University of Athens (http://www.in-silico-oncology.iccs.ntua.gr).

A synoptic outline of the clinical use of a specific Oncosimulator version, as envisaged to happen following an eventually successful completion of its clinical adaptation, optimization and validation process is provided in the form of the following seven steps (see Figure 1):

1. Obtain patient's individual multiscale and inhomogeneous data. Data sets to be collected for each patient include: clinical data (age, sex, weight etc.), eventual previous anti-tumour treatment history, imaging data (e.g. MRI, CT, PET etc images), histopathological data (e.g. detailed identification of the tumour type, grade and

stage, histopathology slide images whenever biopsy is allowed and feasible etc.), molecular data (DNA array data, selected molecular marker values or statuses, serum markers etc.). It is noted that the last two data categories are extracted from biopsy material and/or body fluids.

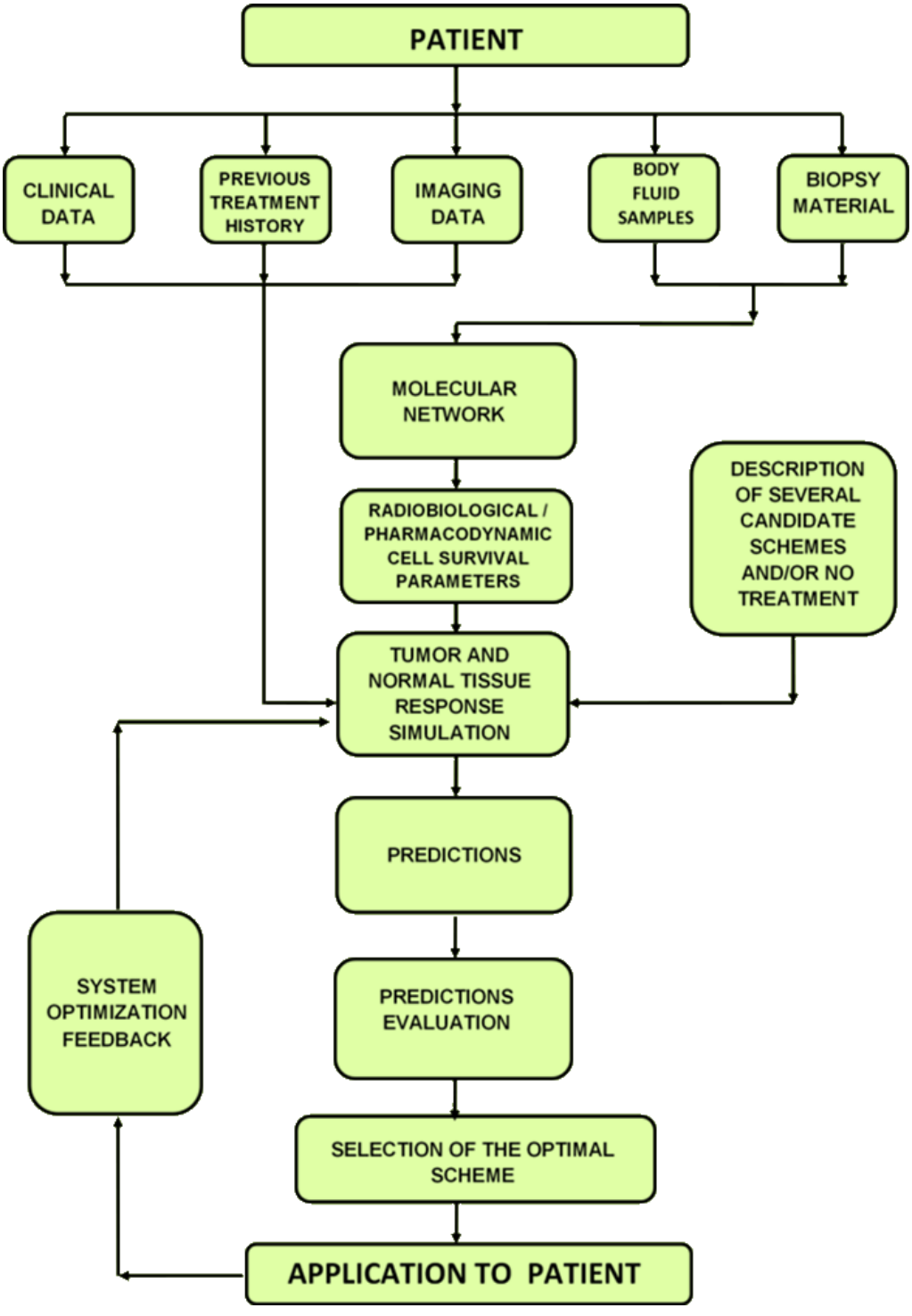


Figure 1: Oncosimulator: a synoptic workflow diagram of one of the system versions (adapted from [22]).

2. Preprocess patient's data. The data collected are preprocessed in order to take an adequate form allowing its introduction into the "Tumour and Normal Tissue Response Simulation" module of the Oncosimulator. For example the imaging data

are segmented, interpolated, eventually fused and subsequently the anatomic entity or entities of interest is or are three dimensionally reconstructed. This reconstruction will provide the framework for the integration of the rest of data and the execution of the simulation. In parallel the molecular data is processed via molecular interaction networks so as to perturb and individualize the average pharmacodynamic or radiobiological cell survival parameters.

3. Describe one or more candidate therapeutic scheme(s) and/or schedule(s). The clinician describes a number of candidate therapeutic schemes and/or schedules and/or no treatment (obviously leading to free, i.e. non-inhibited, tumour growth), to be simulated in-silico i.e. on the computer.
4. Run the simulation. The computer code of tumour growth and treatment response is massively executed on distributed grid or cluster computing resources so that several candidate treatment schemes and/or schedules are simulated for numerous combinations of possible tumour parameter values in parallel. Predictions concerning the toxicological compatibility of each candidate treatment scheme are also produced.
5. Visualize the predictions. The expected reaction of the tumour as well as toxicologically relevant side effect estimates for all scenarios simulated are visualized using several techniques ranging from simple graph plotting to four-dimensional virtual reality rendering.
6. Evaluate the predictions and decide on the optimal scheme or schedule to be administered to the patient. The Oncosimulator's predictions are carefully evaluated by the clinician by making use of their logic, medical education and even qualitative experience. If no serious discrepancies are detected, the predictions support the clinician in taking their final and expectedly optimal decision regarding the actual treatment to be administered to the patient.
7. Apply the theoretically optimal therapeutic scheme or schedule and further optimize the Oncosimulator. The expectedly optimal therapeutic scheme or schedule is administered to the patient. Subsequently, the predictions regarding the finally adopted/applied scheme or schedule are compared with the actual tumour course and a negative feedback signal is generated and used to optimize the Oncosimulator.

4.3 References

[1] Stamatakos G, Dionysiou D, Zacharaki E, Mouravliansky N, Nikita K, Uzunoglu N (2002) In silico radiation oncology: combining novel simulation algorithms with current visualization techniques. Proc IEEE. Special Issue on Bioinformatics: Advances and Challenges 90(11):1764-1777

[2] Dionysiou D, Stamatakos G, Uzunoglu N, Nikita K, Marioli A (2004) A four-dimensional simulation model of tumour response to radiotherapy in vivo: parametric validation considering radiosensitivity, genetic profile and fractionation. J Theor Biol 230: 1–20

[3] Anderson A, Quaranta V (2008) Integrative mathematical oncology. Nat Rev Cancer 8: 227-234

[4] Graf N, Hoppe A, Georgiadi E, Belleman R, Desmedt C, Dionysiou D, Erdt M, Jacques J, Kolokotroni, Lunzer A, Tsiknakis M, Stamatakos G (2009) "In silico oncology" for clinical decision making in the context of nephroblastoma. Klin Paediatr 221: 141-149

[5] Deisboeck T, Zhang L, Yoon J, Costa J (2009) In silico cancer modeling: is it ready for prime time? Nat Clin Pract Oncol 6(1): 34-42

[6] Ventura A, Jackson T, Merajver S (2009) On the role of cell signaling models in cancer research. Cancer Res 69(2): 400-402

[7] Deisboeck T, Stamatakos G (Eds) (2011) Multiscale Cancer Modeling. CRC Press. Print ISBN: 978-1-4398-1440-6, eBook ISBN: 978-1-4398-1442-0

[8] Frieboes H, Zheng X, Sun C, Tromberg B, Gatenby R, Cristini V (2006) An integrated computational/experimental model of tumour invasion. Cancer Res 66(3):1597-604

[9] Enderling H, Chaplain M, Anderson A, Vaidya J (2007) A mathematical model of breast cancer development, local treatment and recurrence. J Theor Biol 246: 245–259

[10] Powathil G, Kohandel M, Sivaloganathan S, Oza A, Milosevic M (2007) Mathematical modeling of brain tumours: effects of radiotherapy and chemotherapy. Phys Med Biol 52: 3291-3306

[11] Rockne R, Alvord E, Szeto M, Gu S, Chakraborty G, et al. (2008) Modeling diffusively invading brain tumours: an individualized approach to quantifying glioma evolution and response to therapy. In: Belomo N, Chaplain M, Angelis E (eds.) Selected topics in cancer modeling. Birkhaeuser. pp. 207-221

[12] Castorina P, Carcò D, Guiot C, Deisboeck T (2009) Tumour growth instability and its implications for chemotherapy. Cancer Res 69(21): 8507–15

[13] Sottoriva A, Verhoeff J, Borovski T, McWeeney S, Naumov L, et al. (2010) Cancer stem cell tumour model reveals invasive morphology and increased phenotypical heterogeneity. Cancer Res 70(1): 46-56

[14] Duechting W, Ulmer W, Lehrig R, Ginsberg T, Dedeleit E (1992) Computer simulation and modeling of tumour spheroids growth and their relevance to optimization of fractionated radiotherapy. Strahlenther Onkol 168(6): 354-360

[15] Kansal A, Torquato S, Harsh G, Chiocca E, Deisboeck T (2000) Simulated brain tumour growth dynamics using a three-dimensional cellular automaton. J Theor Biol 203:367-82

[16] Stamatakos G, Zacharaki E, Makropoulou M, Mouravliansky N, Marsh A, Nikita K, Uzunoglu N (2001) Modeling tumour growth and irradiation response in vitro- a combination of high-performance computing and web based technologies including VRML visualization. IEEE Trans Inform Technol Biomed 5(4): 279-289

[17] Stamatakos G, Antipas V, Uzunoglu N, Dale R (2006) A four dimensional computer simulation model of the in vivo response to radiotherapy of glioblastoma multiforme: studies on the effect of clonogenic cell density. Br J Radiol 79: 389-400

[18] Stamatakos G, Antipas V, Uzunoglu N (2006) A spatiotemporal, patient individualized simulation model of solid tumour response to chemotherapy in vivo: the paradigm of glioblastoma multiforme treated by temozolomide. IEEE Trans Biomed Eng 53: 1467- 1477

[19] Stamatakos G, Dionysiou D, Graf N, Sofra N, Desmedt C, Hoppe A, Uzunoglu N, Tsiknakis M (2007) The Oncosimulator: a multilevel, clinically oriented simulation system of tumour growth and organism response to therapeutic schemes. Towards the clinical evaluation of in silico oncology. Proc 29th Annual Intern Conf IEEE EMBS. pp. 6628-6631

[20] Stamatakos G, Kolokotroni E, Dionysiou D, Georgiadi E, Desmedt C (2010) An advanced discrete state-discrete event multiscale simulation model of the response of a solid tumour to chemotherapy: Mimicking a clinical study. J Theor Biol 266(1): 124-139

[21] Stamatakos G, Georgiadi E, Graf N, Kolokotroni E, Dionysiou D (2011) Exploiting clinical trial data drastically narrows the window of possible solutions to the problem of clinical adaptation of a multiscale cancer model. PLOS ONE 6(3), e17594

[22] Stamatakos G (2011) In Silico Oncology Part I: Clinically Oriented Cancer Multilevel Modeling Based on Discrete Event Simulation. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 407–436. DOI: 10.1201/b10407-19

[23] Dionysiou D, Stamatakos G, Uzunoglu N, Nikita K (2006) A computer simulation of in vivo tumour growth and response to radiotherapy: New algorithms and parametric results. Comput Biol Med 36:448–464

[24] Dionysiou D, Stamatakos G (2006) Applying a 4D multiscale in vivo tumour growth model to the exploration of radiotherapy scheduling: the effects of weekend treatment gaps and p53 gene status on the response of fast growing solid tumours. Cancer Informatics 2:113-121

[25] Dionysiou D, Stamatakos G, Marias K (2007) Simulating cancer radiotherapy on a multi-level basis: biology, oncology and image processing. Lect Notes Comp Sci 4561: 569-575

[26] Dionysiou D, Stamatakos G, Gintides D, Uzunoglu N, Kyriaki K (2008) Critical parameters determining standard radiotherapy treatment outcome for glioblastoma multiforme: a computer simulation. Open Biomed Eng J 2: 43-51

[27] Anderson A, Weaver A, Cummings P, Quaranta V (2006) Tumour morphology and phenotypic evolution driven by selective pressure from the Microenvironment. Cell 127: 905-915

[28] Ubezio P, Cameron D (2008) Cell killing and resistance in pre-operative breast cancer chemotherapy. BMC Cancer 8

[29] Titz B, Jeraj R (2008) An imaging-based tumour growth and treatment response model: investigating the effect of tumour oxygenation on radiation therapy response. Phys Med Biol 53: 4471-4488

[30] Guiot C, Delsanto P, Carpinteri A, Pugno N, Mansury Y, Deisboeck T (2006). The dynamic evolution of the power exponent in a universal growth model of tumours. J Theor Biol 240(3): 459-63

[31] Guiot C, Delsanto P, Gliozzi A (2011) Do tumour invasion strategies follow basic physical laws? In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press, pp. 237–252. DOI: 10.1201/b10407-13

[32] Komarova N (2011) Building stochastic models for cancer growth and treatment. ”In Multiscale Cancer Modeling. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 339–358. DOI: 10.1201/b10407-16

[33] Maley C, Lewis W, Reid B (2011) Has cancer sculpted the genome? Modeling linkage and the role of tetraploidy in neoplastic progression In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 45–66

[34] Murray J (2003) Mathematical biology II: spatial models and biomedical applications. (third edition). Springer-Verlag, Heidelberg, pp. 543-546

[35] Swanson K, Alvord E, Murray J (2002) Virtual brain tumours (gliomas) enhance the reality of medical imaging and highlight inadequacies of current therapy. Br J Cancer 86:14-18

[36] Breward C, Byrne H, Lewis C (2003) A multiphase model describing vascular tumour growth. Bull Math Biol 65(4): 609-40

[37] Cristini V, Frieboes H, Gatenby R, Caserta S, Ferrari M, Sinek J (2005). Morphological instability and cancer invasion. Clin Cancer Res 11: 6772–6779

[38] D’Onofrio A (2005) A general framework for modeling tumour-immune system competition and immunotherapy: analysis and medical inferences. Physica D 208: 220-235.

[39] Ramis-Conde I, Chaplain M, Anderson A (2008) Mathematical modelling of cancer cell invasion of tissue. Math Comput Model 47: 533-545.

[40] Bergdorf M, Milde F, Koumoutsakos P (2011) Continuum models of mesenchymal cell migration and sprouting angiogenesis. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 213–235. DOI: 10.1201/b10407-12

[41] Chakraborty G, Sodt R, Massey S, Gu S, Rockne R, Ellsworth C, Swanson K (2011) Bridging from multiscale modeling to practical clinical applications in the study of human gliomas. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 359–383. DOI: 10.1201/b10407-17

[42] Chaplain M, Macklin P, McDougall S, Anderson A, Cristini V, Lowengrub J (2011) Multiscale mathematical modeling of vascular tumour growth: an exercise in transatlantic cooperation. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 253–308. DOI: 10.1201/b10407-14

[43] Fletcher A, Mirams G, Murray P, Walter A, Kang J, Cho K, Maini P, Byrne H (2011) Multiscale modeling of colonic crypts and early colorectal cancer. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 111–134. DOI: 10.1201/b10407-7

[44] Hirsch S, Lloyd B, Szczerba D, Székely G (2011) A multiscale simulation framework for modeling solid tumour growth with an explicit vessel network. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 309–337. DOI: 10.1201/b10407-15

[45] Konukoglu E, Clatz O, Delingette H, Ayache N (2011) Personalization of reaction-diffusion tumour growth models in MR images; application to brain gliomas characterization and radiotherapy planning. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 385–406. DOI: 10.1201/b10407-18

[46] Duechting W, Vogelsaenger T (1981) Three-dimensional pattern generation applied to spheroidal tumour growth in a nutrient medium. Int. J. Biomed. Comput., 12 (5), 377–392.

[47] Ginsberg T, Ulmer W, Duechting W (1993) Computer simulation of fractionated radiotherapy: further results and their relevance to percutaneous irradiation and brachytherapy. Strahlenther Onkol 169: 304-310

[48] Stamatakos G, Zacharaki E, Uzunoglu N, Nikita K (2001) Tumour growth and response to irradiation in vitro: a technologically advanced simulation model. Int. J. Radiat Oncol Biol Phys 51(3) Sup.1: 240-241

[49] Zacharaki E, Stamatakos G, Nikita K, Uzunoglu N (2004) Simulating growth dynamics and radiation response of avascular tumour spheroid model validation in the case of an EMT6/Ro multicellular spheroid. Comput Methods Programs Biomed 76:193-206

[50] Anderson A, Basanta D, Gerlee P, Rejniak K (2011) Evolution, regulation and disruption of homeostatis and its role in carcinogenesis. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 1–30. DOI: 10.1201/b10407-2

[51] Gatenby R (2011) The physical microenvironment in somatic evolution of cancer. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press 2011. pp. 135–155. DOI: 10.1201/b10407-8

[52] Harjanto D, Zaman M (2011) Multiscale modeling of cell motion in three-dimensional environments. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 157–172. DOI: 10.1201/b10407-9

[53] Jean L, Luebeck E (2011) A stochastic multiscale model framework for colonic stem cell homeostatis. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 87–109. DOI: 10.1201/b10407-6

[54] Purvis J, Shih A, Liu Y, Radhakrishnan R (2011) Cancer cell: linking oncogenic signalling to molecular structure. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 31–44. DOI: 10.1201/b10407-3

[55] Solé R (2011) Catastrophes and complex networks in genomically unstable tumourigenesis. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 67–86. DOI: 10.1201/b10407-5

[56] Mansury Y, Deisboeck T (2003) The impact of “search precision” in an agent-based tumour model. J Theor Biol 224(3): 325-337

[57] Wang Z, Bordas V, Sagotsky J, Deisboeck T (2011) Simulating cancer growth with agent-based models. In: Deisboeck T, Stamatakos G (eds.) Multiscale Cancer Modeling. CRC Press. pp. 173–192. DOI: 10.1201/b10407-10

[58] Clatz O, Sermesant M, Bondiau P, Delingette H, Warfield S, Malandain G, Ayache N. (2005). Realistic simulation of the 3-D growth of brain tumours in MR images coupling diffusion with biomechanical deformation. IEEE Trans Med Imaging 24(10):1334-46

[59] Graf N, Hoppe A (2006) What are the expectations of a clinician from in silico oncology? Proc. 2nd International Advanced Research Workshop on In Silico Oncology, Kolympari, Chania, Greece, Sept. 25-26, 2006. Marias K, Stamatakos G (eds.) 36-38

[60] Graf N, Desmedt C, Buffa F, Kafetzopoulos D, Forgo N, Kollek R, Hoppe A, Stamatakos G, Tsiknakis M (2008) Post-genomic clinical trials - the perspective of ACGT. Ecancermedicalscience 2

[61] Stamatakos G, Dionysiou D, Nikita K, Zamboglou N, Baltas D, Pissakas G, Uzunoglu N (2001) In vivo tumour growth and response to radiation therapy: a novel algorithmic description. Int. J. Radiat. Oncol. Biol. Phys 51(3) Sup.1: 240

[62] Stamatakos G, Uzunoglu N (2006) Computer simulation of tumour response to therapy. In S. Nagl Ed. Cancer Bioinformatics: from therapy design to treatment. John Wiley & Sons Ltd, Chichester, UK. pp.109-125

[63] Stamatakos G, Antipas V, Uzunoglu N (2006) Simulating chemotherapeutic schemes in the individualized treatment context: the paradigm of glioblastoma multiforme treated by temozolomide in vivo. Comput Biol Med. 36(11): 1216-34

[64] Stamatakos G, Dionysiou D, Uzunoglu N (2007) In silico radiation oncology: a platform for understanding cancer behaviour and optimizing radiation therapy treatment. In: Akay M (Ed.) Genomics and Proteomics Engineering in Medicine and Biology. Wiley-IEEE Press, Hoboken, NJ. pp.131-156

[65] Antipas V, Stamatakos G, Uzunoglu N, Dionysiou D, Dale R (2004) A spatiotemporal simulation model of the response of solid tumours to radiotherapy in vivo: parametric validation concerning oxygen enhancement ratio and cell cycle duration. Phys Med Biol 49: 1-20

[66] Antipas V, Stamatakos G, Uzunoglu N (2007) A patient-specific in vivo tumour and normal tissue model for prediction of the response to radiotherapy: a computer simulation approach. Meth Inf Med 46: 367-375

[67] Kolokotroni E, Stamatakos G, Dionysiou D, Georgiadi E, Desmedt C, Graf N (2008) Translating multiscale cancer models into clinical trials: simulating breast cancer tumour dynamics within the framework of the “Trial of Principle” clinical trial and the ACGT project. Proc. 8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2008), Athens, Greece, 8-10 Oct. 2008. IEEE Catalog Number: CFP08266, ISBN: 978-1-4244-2845-8, Library of Congress: 2008907441, Paper No. BE-2.1.1

[68] Kolokotroni E, Dionysiou D, Uzunoglu N, Stamatakos G (2011) Studying the growth kinetics of untreated clinical tumours by using an advanced discrete simulation model. Mathematical and Computer Modelling, in press.

[69] Georgiadi E, Stamatakos G, Graf N, Kolokotroni E, Dionysiou D, et al. (2008) Multilevel cancer modeling in the clinical environment: simulating the behaviour of Wilms tumour in the

context of the SIOP 2001/GPOH clinical trial and the ACGT project. In Proc of 8th IEEE International Conference on Bioinformatics and Bioengineering. Athens, Greece

[70] Stamatakos G, Dionysiou D (2009) Introduction of hypermatrix and operator notation into a discrete mathematics simulation model of malignant tumour response to therapeutic schemes in vivo. Some operator properties. Cancer Informatics 2009:7 239–251

[71] Lunzer A, Belleman R, Melis P, Pukacki J, Spychala P, Stamatakos G (2010) Validating the ACGT Oncosimulator with a grid-supported visualisation environment. In: Stamatakos G, Dionysiou D (eds.) Proc. 4th Int. Adv. Res. Workshop on In Silico Oncology and Cancer Investigation (4th IARWISOCI) - The ContraCancrum Workshop, Athens, Greece, Sept. 8-9, 2010 (www.4th-iarwisoci.iccs.ntua.gr), pp. 93-95

[72] Marias K, Dionysiou D, Sakkalis V, Graf N, Bohle R, Coveney P, Wan S, Folarin A, Büchler P, Reyes M, Clapworthy G, Liu E, Sabczynski J, Bily T, Roniotis A, Tsiknakis M, Kolokotroni E, Giatili S, Veith C, Messe E, Stenzhorn H, Kim Y, Zasada S, Haidar A, May C, Bauer S, Wang T, Zhao Y, Karasek M, Grewer R, Franz A, Stamatakos G (2011) Clinically driven design of multi-scale cancer models: the ContraCancrum project paradigm. Interface Focus, Published online on 30 March 2011, doi: 10.1098 / rsfs.2010.0037

[73] May C, Kolokotroni E, Stamatakos G, Buechler P (2011) Coupling biomechanics to a cellular level model: an approach to patient-specific image driven multi-scale and multi-physics tumour simulation. Progress in Biophysics and Molecular Biology, in press

[74] Marias K, Dionysiou D, Stamatakos G, Zacharopoulou F, Georgiadi E, Maris T, Tollis I (2007) Multi-level analysis and information extraction considerations for validating 4D models of human function. Lect Notes Comput Sci 4561, pp. 703-709

[75] Stamatakos G (2006). Spotlight on cancer informatics. Cancer Informatics 2: 83-86

[76] Stamatakos G (2008) In silico oncology: a paradigm for clinically oriented living matter engineering. Proc. 3rd International Advanced Research Workshop on In Silico Oncology, Istanbul, Turkey, Sept. 23-24, 2008. Stamatakos G, Dionysiou D (eds.) pp.7-9

[77] Stamatakos G, Kolokotroni E, Dionysiou D, Georgiadi E, Giatili S (2009) In silico oncology: a top-down multiscale simulator of cancer dynamics. Studying the effect of symmetric stem cell division on the cellular constitution of a tumour. In Dössel O, Schlegel W (eds.) WC 2009, IFMBE Proc 25/IV, pp. 1830–1833, 2009

[78] Graf N, Desmedt C, Hoppe A, Tsiknakis M, Dionysiou D, Stamatakos G (2007). Clinical requirements of “in silico oncology” as part of the integrated project ACGT (Advancing Clinico-Genomic Trials on Cancer). Eur J Cancer Suppl 5(4): 83

[79] Graf N (2011) In silico oncology part II: clinical requirements regarding in cilico oncology Deisboeck T, Stamatakos G (eds.) CRC Press. pp. 437–446. DOI: 10.1201/b10407-20

5 Bioinformatics and Personalized Medicine

Bioinformatics has a crucial role in personalized medicine, as in modelling complex systems (like the human body), implementation of methods and tools for making "omics" data meaningful (prognostic and predictive models, study of diagnostic, prognostic and predictive biomarkers) and role in translation of biological knowledge into clinical practice [1].

Communication plays an important role in this process. The study design as well as the iterative process that allows the formulation of the correct hypotheses until the extraction of results, their interpretation and the translation into clinical practice, are clearly achievable only in dialogue between clinical, bench, computational and scientific research (Figure 1).

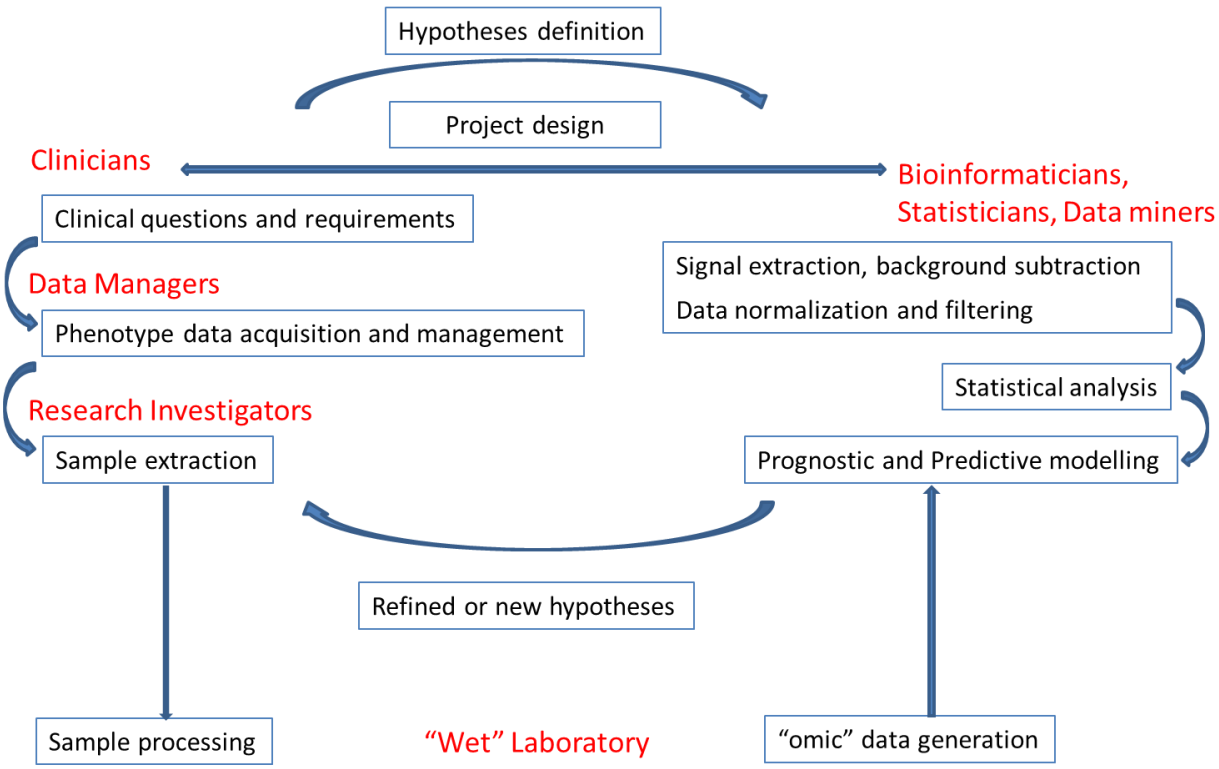


Figure 1: The study design and the execution/research process as a dialogue between clinical, laboratory and computational research environment in the context of translational biomedical research (adapted from schema in [2]).

Further, data availability through different technological platforms enhanced the possibility to better understand genes and protein regulation along with new computer resources, VPH technologies [2,3] and new, expanded personalized medicine concepts (see Figure 2) [4,5].

5.1 Biomedical Ontologies, Terminologies and Databases

Bioinformatics interacts with many resources in their study of the "omics" data; we can classify them in three categories: ontologies, terminologies and databases (OTDs). These three resources are fundamental for the correct interpretation and communication of results.

OTDs are also used for life-science data integration, patient status description, and drug delivery information provision in the domain of oncology. Specific features of these OTDs make them relevant for clinical practice in oncology and for oncology-related biomedical research and therefore to bioinformatics. In the next paragraphs we list the main OTDs used.

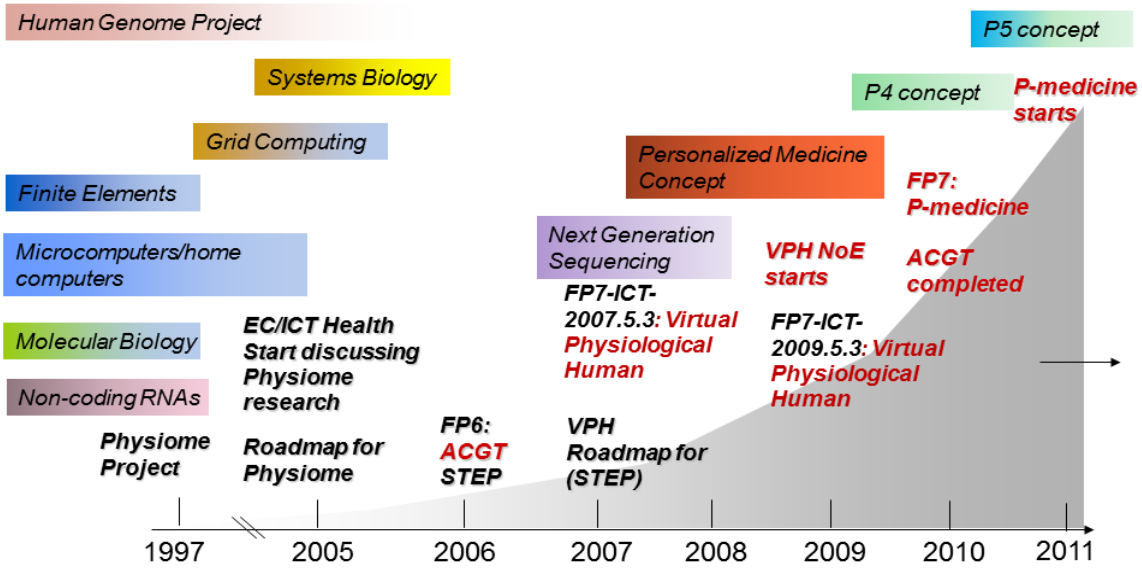


Figure 2: Adapted version of the slide number 14 of the VPH NoE slide set 2011 presentation for dissemination, [6]. It illustrates the evolution toward VPH models and personalized medicine concept and research in the last 15 years.

5.1.1 Ontologies

In general, we can formally define ontology by using the following definition by Wikipedia:

> "Ontology (from the Greek ὤν, genitive ὄντος: "of that which is", and -λογία, -logia: science, study, theory) is the philosophical study of the nature of being, existence or reality as such, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences."

In term of life sciences, we can find several useful ontologies online like:

5.1.1.1 Gene Ontology (GO) and Gene Ontology Annotation (GOA)

GO is developed by the Gene Ontology Cnsortium. GOA@EBI (European Bioinformatics Institute, http://www.ebi.ac.uk/GOA) as one partner in this consortium, develops GOA.

The GO (http://www.geneontology.org) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO project has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. As of May 2011, GO contains 34173 terms of which 100% have definitions with 20771 belonging to the biological process axis, 2833 to the cellular component axis and 9010 to the molecular function. GOA provides assignments of gene products to the Gene Ontology (GO) resource.

Every GO term is associated with an evidence code belonging to one of the categories:

- Experimental evidence codes: EXP, IDA, IPI, IMP, IGI, IEP

- Computational analysis evidence codes: ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA
- Author statement evidence codes: TAS and NAS
- Curator statement evidence codes: IC and ND
- Automatically assigned evidence codes: IEA
- Obsolete evidence codes: NR

More details can be found in http://www.geneontology.org/GO.evidence.shtml. GO and GOA are available for free use within the terms of license; the archive can be downloaded at http://www.geneontology.org/GO.downloads.ftp.cvs.shtml.

Many tools have been built around GO, some by the GO consortium and many out the consortium, they can be classified as:

- Ontology or annotation: Browse, search engine, visualization,editor
- Database or data warehouse
- Software library
- Statistical analysis
- Slimmer-type tool
- Term enrichment
- Text mining
- Protein interactions
- Functional similarity
- Semantic similarity

A complete list of the available tools per category can be retrieved at http://www.geneontology.org/GO.tools.shtml.

Several genome browsers and functional tools use GO to conduct enrichment analysis or they simple report them as associated to the respective genes:

- http://www.ensembl.org
- http://david.abcc.ncifcrf.gov

GO and GOA provide annotations to various gene products which are directly associated with carcinomas. The mapping of those gene products to entities within Uniprot (http://www.uniprot.org) and pathway databases (http://www.genome.jp/kegg) and that to OMIM (http://www.ncbi.nlm.nih.gov/omim) further close the loop by which the various functions and effects of those gene products can be queried. GO terms themselves provide a rather primitive collection of relations between the classes. However the annotations to those terms and their relationships help provide certain kinds of inferences.

5.1.1.2 Open Biomedical Ontologies (OBO)

The OBO has created controlled vocabularies for shared use across different biological and medical domains. It is part of the resources of the U.S. National Center for Biomedical Ontology (http://www.bioontology.org) [7]. It has been designed and maintained to follow the principles of interoperability, improvement of quality and formal rigor. Related projects are:

- Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup)
- Gene Ontology Consortium (http://en.wikipedia.org/wiki/Gene_Ontology)

- Sequence Ontology (http://www.sequenceontology.org)
- Generic Model Organism Databases (http://en.wikipedia.org/wiki/Generic_Model_Organism_Database)
- Standards and Ontologies for Functional Genomics (http://www.sofg.org)
- Functional GEnomics Data (FGED) (http://en.wikipedia.org/wiki/FGED_Society)
- Ontology for Biomedical Investigations (http://en.wikipedia.org/wiki/Ontology_for_Biomedical_Investigations)
- Plant Ontology (http://www.plantontology.org)
- Phenoscape (https://www.phenoscape.org)

5.1.1.3 Foundational Model of Anatomy (FMA)

FMA (http://sig.biostr.washington.edu/projects/fm) is concerned with the representation of classes and relationships necessary for the symbolic representation of the structure of the human body in a form that is understandable to humans and is also navigable by computerized systems. Specifically, the FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. FMA has four interrelated components:

1. Anatomy taxonomy: classifies anatomical entities according to the characteristics they share and by which they can be distinguished from one another.
2. Anatomical Structural Abstraction: specifies the part-whole and spatial relationships that exist between the entities represented in the taxonomy
3. Anatomical Transformation Abstraction: specifies the morphological transformation of the entities represented in the taxonomy during prenatal development and the postnatal life cycle
4. Metaknowledge: specifies the principles, rules and definitions according to which classes and relationships in the other three components of FMA are represented. FMA contains approximately 72,000 classes, over 115,000 terms and over 2.1 million relationship instances from 168 relationship types.

FMA is very useful for representing anatomical entities in relevance to oncology. These include carcinoma staging, locations for radiotherapy and surgery, access routes for various procedures, locations for drug actions, and so on. The robust formalism allows to derivation of inferences, especially for staging of carcinomas.

FMA is available for free use but a contract must be signed and download access asked for.

5.1.2 Terminologies

5.1.2.1 Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT)

SNOMED CT (http://www.ihtsdo.org/snomed-ct) is a generic healthcare terminology together with various relations between it's over 300,000 concepts. There are about a million descriptions of those concepts and about a million semantic links between them. The SNOMED CT core content consists of:

- Concepts Table
- Descriptions Table
- Relationship Table

- History Table
- ICD Mapping

Since SNOMED CT covers the generic medical domain, there are many areas where there are overlaps with the domain of carcinomas. In particular, the classification of procedures, medications and diseases are useful. However the problems with the classifications and relationship formalisms in SNOMED CT may lead to some limitations in inference derivation.

SNOMED CT is available under a special license.

5.1.2.2 Medical Dictionary for Regulatory Activities (MedDRA)

MedDRA (http://www.meddramsso.com) is a terminology for drug and medical device side-effects and malfunctions. It emphasizes ease of use for data entry, retrieval, analysis, and display when dealing with registering, documenting, and safety monitoring of medical products. The top-level classification of MedDRA consists mainly of disorders classified according to various body systems: respiratory disorders, cardiac disorders, gastrointestinal disorders, immune system disorders, endocrine disorders, and so on.

MedDRA is used to code drug and medical device-side effects in all the medical domains and thus is also used for management of carcinomas.

An annual subscription fee is required for use.

5.1.2.3 Unified Medical Language System (UMLS)

UMLS (http://umlsinfo.nlm.nih.gov) consists of the Metathesaurus, Semantic Network, SPECIALIST Lexicon and MetamorphoSys:

- The Metathesaurus is the vocabulary database of over a million terms dealing with the content of biomedical literature and Electronic Health Records. When more than one meaning is assigned to a single vocabulary term then both meanings of the term are represented within the Metathesaurus with the reference to specific source vocabularies. The source vocabularies integrated with the Metathesaurus include the ICD, SNOMED CT, CPT codes, DSM, HUGO, MedDRA and NCI Thesaurus.
- The Semantic Network consists of Semantic Types that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus a set of Semantic Relations, which exist between Semantic Types.
- The SPECIALIST Lexicon provides the lexical information needed for the SPECIALIST Natural Language Processing (NLP) System.
- The MetamorphoSys is the UMLS installation wizard and Metathesaurus customization tool included in each UMLS release.

UMLS is a conglomerate where terms from over 100 OTDs can be queried. The Metathesaurus has been extensively used for text mining and natural language processing in biomedical domain and thus is relevant for carcinomas. The UMLS Semantic Network and the Metathesaurus are not formalized ontologies; however, recently efforts are being made to formalize the Semantic Network in a way that inferences can be made based on it. UMLS has also been used to for mutant protein term identification from the natural text, something that helps in a semiautomatic extension of the existing mutant protein databases.

UMLS is available under a special license.

5.1.2.4 International Classification of Diseases (ICD)

The ICD (http://www.who.int/classifications/icd) is designed to promote international comparability in the collection, processing, classification, and presentation of diagnostics in health epidemiology, health management and mortality statistics. These include the analysis of the general health situation of population groups and monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables such as the characteristics and circumstances of the individuals affected.

To a large extent, ICD provides a disease classification on the basis of anatomy. Although not all the diseases within ICD are classified according to anatomy, the neoplasms are more or less classified within the anatomical partition. Thus, an ontology of carcinomas, which follows the anatomical partition for classification of neoplasms and related diseases, can use portions of ICD more easily than other disease classifications. However there are issues of misclassifications within ICD and also terms that do not represent a real disease.

ICD is available for free use within the terms of its license.

5.1.2.5 Medical Subjects Headings (MeSH)

MeSH (http://www.nlm.nih.gov/mesh) is a controlled vocabulary thesaurus consisting of sets of terms-naming descriptors in a hierarchical structure that permits searching at various levels of specificity. The top-level classification includes: Anatomy, Organisms, Diseases, Chemicals and Drugs, Analytical, Diagnostic and Therapeutic Techniques and Equipment, Psychiatry and Psychology, Biological Sciences, and Physical Sciences. MeSH is used on MEDLINE to index bibliographic citations and author abstracts from over 4,000 journals.

MeSH is useful for the carcinoma domain due to its usage within PubMed (http://www.ncbi.nlm.nih.gov/pubmed). All major carcinoma literature is classified within PubMed and is available for retrieval using the MeSH coding.

MeSH is available for free use within the terms of its license.

5.1.2.6 National Drug Code Directory

The Drug Listing Act of 1972 requires registered drug establishments to provide the FDA with a current list of all drugs manufactured, prepared, propagated, compounded, or processed by it for commercial distribution. Drug products are identified and reported using a unique, three-segment number, called the National Drug Code (NDC) (http://www.fda.gov/cder/ndc) which is a universal product identifier for human drugs. FDA inputs the full NDC number and the information submitted as part of the listing process into a database known as the Drug Registration and Listing System (DRLS). Several times a year, FDA extracts some of the information from the DRLS database for publication in the NDC Directory.

The usage of NDC is mandatory for coding related to medications applying to all medical domains and thus is applicable to carcinomas. Although NDC usage is mandated only within the USA, many other countries base their requirements in line with what is proposed by NDC. Moreover, since most of the major Hospital Information Systems and Drug Databases are NDC compliant, these codes are embedded in systems used almost around the world.

NDC is available for free use within the license terms.

5.1.3 Databases

5.1.3.1 Online Mendelian Inheritance in Man (OMIM)

The OMIM (http://www.ncbi.nlm.nih.gov/omim) is a catalogue of human genes and genetic disorders together with textual information and references. It illustrates the genes which have been associated with a particular disease in literature. OMIM focuses primarily on inherited or heritable, genetic diseases. It is also considered to be a phenotypic companion to the human genome project and was originally based upon the book Mendelian Inheritance in Man. Each entry is given a unique six-digit number whose first digit indicates the mode of inheritance of the gene involved.

The connection between gene abnormalities and diseases is useful for almost all the diseases present within OMIM. However it is especially important for hereditary diseases and carcinomas. The large number of genetic abnormalities associated with carcinomas is the evidence that such associations are related to the various protein and pathway abnormalities forming a part of the pathologies within carcinomas [8,9].

OMIM is available for free use within the terms of its license.

5.1.3.2 Universal Protein Resource (UniProt)

The UniProt Consortium, which is comprised of the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR) developed UniProt (http://www.uniprot.org). It is a central repository of protein sequence and function and provides several tools:

- UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record to speed up searches
- UniProt Archive (UniParc) is a repository with the history of all protein sequences.

All the sources of UniProt provide mutant protein databases with annotations to the diseases they are associated with. The number of mutant proteins associated with carcinomas form one of the largest portion of mutant protein databases. UniProt plays an important role in bridging together the gap between biological and medical information related to carcinomas.

UniProt is available for free use within the terms of its license.

5.1.3.3 NIH Single Nucleotide Polymorphism (dbSNP) Database

Single Nucleotide Polymorphisms (SNPs) are the most common genetic variations, taking place once every 100 to 300 bases. A key aspect of research in genetics is the association of sequence variation with heritable phenotypes. It is expected that SNPs will accelerate the identification of disease genes by allowing researchers to look for associations between a disease and specific differences (SNPs) in a population. This differs from the more typical approach of pedigree analysis that tracks transmission of a disease through a family. It is much easier to obtain DNA samples from a random set of individuals in a population than it is to obtain them from every member of a family over several generations. Once discovered, additional laboratories, using the sequence information around the polymorphism and the specific experimental conditions, can use these polymorphisms. The Single Nucleotide Polymorphism database (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP) is a public-domain archive for a broad collection of simple genetic polymorphisms. As of May 2011, dbSNP contains over 19 million validated Human RefSNP clusters, over 6 million validated Mus musculus RefSNP clusters and over 3 million validated Gallus gallus clusters.

In the last few years, SNPs have gained a lot of importance in clinical research. The database information is compared to gene expression information of many carcinomas.

Multispecies database allows comparison across different species and also make results from animal models comparable to the human case. SNPs are widely used in chemotherapy drug development targeted against specific mutant proteins or protein complexes [10,11]. Recently SNPs have also been applied for clinical research in radiotherapy [12].

dbSNP is available for free usage within the terms of licensing.

5.1.3.4 Japanese Single Nucleotide Polymorphism (JSNP) Database

The JSNP (http://snp.ims.u-tokyo.ac.jp) is the database for DNA sequence variations, polymorphic markers to investigate genes susceptible to diseases or those related to drug responsiveness. The 38th data release consists of 197,195 SNPs and 84,651 SNPs with allele frequency. SNPs will also be deposited in the public dbSNP and GWAS central.

Similar to dbSNP, JSNP database information is useful for gene expression studies and drug development.

JSNP is available for free usage within the terms of licensing.

5.1.3.5 The microRNA database: miRBase

MicroRNAs (miRNAs) are a class of non-coding RNAs (ncRNAs, RNAs that do not codify for proteins); they are a large class of phylogenetically conserved single-stranded RNA molecules of 19 to 25 nucleotides that are involved in post-transcriptional gene silencing that were found to be involved in any type of analyzed human cancer [13]; miRNAs not only regulate various developmental and physiologic processes but also are involved in cancer development, diagnosis and progression [14-17]. miRNAs also play a role in other diseases, such as schizophrenia [18] and diabetes [19].

The stability of miRNAs in formalin-fixed, paraffin-embedded tissues and body fluids is advantageous for biomarker discovery and validation. In addition, miRNAs can be extracted from small biopsy speciments, which is a further advantage. Finally, miRNAs are potential therapeutic agents for personalized cancer management [20].

Other ncRNA classes, like long nc-RNAs (lincRNAs) have been found associated with cancer, especially metastasis [21], ultra conserved genes (UCGs) have been found deregulated in cancer [22] and it has been hypothesized that microRNAs may regulate the expression of ultraconserved regions in various cancers including CRC and CLL [23] (see Table 1). High-throughput technology for gene expression assays has led to the discovery that most human transcriptional units are ncRNAs [24,25].

The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download (http://www.mirbase.org/ftp.shtml)

miRBase is available for free use within the terms of its license.

ncRNA	Description	Approximate size (nt)
lncRNA	Long non-coding RNA	> 200

macroRNAs	Long expressed non-coding regions (ENORs)	> 10,000
miRNA	microRNA	~ 19-25
siRNA	Small interfering/silencing RNA	~ 20–25
piRNA	PIWI-interacting RNA	~ 24–30
snoRNA	Small nucleolar RNA	~ 70–240

Table 1: A list of several non-coding RNAs.

5.2 Tools for the Analysis of Biomedical Data

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This huge availability of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index data and for special tools to view and analyse them. The large interest in bioinformatics, a discipline at the intersection of molecular biology and computer science, is fuelled by the excitement surrounding the sequencing of the human genome and the promise of a new era in which genomic research dramatically improves the human condition.

Bioinformatics is a fertile recent area for interdisciplinary research as well as a source for innovative information science and technology development. It has already served as an inspiration for many biological metaphors in computing, and conversely, information and computation paradigms have become ubiquitous in molecular biology. Researchers at the frontiers of biology and informatics are developing and can be expected to increasingly develop very novel symbiotic forms of science and technology.

5.2.1 Microarrays Data Analysis

5.2.1.1 Introduction

Microarray data analysis is heavily dependent on Gene Expression Data Mining (GEDM) technology, and in the recent years a lot of research efforts are in progress. GEDM is used to identify intrinsic patterns and relationships in gene expression data. The identification of patterns in complex gene expression datasets provides two benefits:

- Generation of insight into gene transcription conditions.
- Characterization of multiple gene expression profiles in complex biological processes, e.g. pathological states.
- GEDM activities are based on two approaches:
    - Hypothesis testing: to investigate the induction or perturbation of a biological process that leads to predicted results
    - Knowledge Discovery: to detect internal structure in biological data [26-28].

5.2.1.1.1 Available Systems and Tools

- Affymetrix (http://www.affymetrix.com)
- Qiagen (http://www.qiagen.com)

- Agilent (http://www.agilent.com)

5.2.1.2 Data Analysis Phases

By measuring transcription levels of genes in an organism under various conditions or in different tissues we can build up 'gene expression profiles', which characterize the dynamic function of each gene. Microarray data are represented in a matrix with rows representing genes, columns representing samples and each cell containing a number characterizing the gene expression level in the particular sample, i.e., the gene expression matrix.

The analysis of microarray data can be summarized in several main steps:

1. Experimental design
2. Pre-processing
3. Normalization
4. Exploratory analysis
5. Statistical tests: class comparison, class prediction, survival analysis

Because there are many sources of noise and systematic variability in microarray experiments [29] data normalization and pre-processing are crucial in analysis [30]. Normalization includes those transformations that control systematic variability within a chip or across multiple chips [31]. The simplest way data normalization can be done is by dividing or subtracting all expression values by a representative value for the system or by a linear transformation to a fixed mean (i.e., 0.0) and unit variance (i.e., 1.0) (sometimes called "median polishing"). However, the linear response between the true expression level and measured fluorescent intensity may not be guaranteed [32,33], especially when dye biases depend on array spot intensity or multiple print tips are used in the microarray spotter [34].

Data pre-processing includes those transformations that prepare the data for the subsequent analysis with scaling and filtering as major steps. A low variation filter to exclude genes that did not significantly change across experiments has been successfully used in many studies [35]. Statistical significance testing, like variance analysis and multiple comparisons, can also be used to filter data showing no significant change across conditions when a sufficient number of repeated observations are available. It is highly recommended to scatter-plot the data whenever possible. The most straightforward approach to microarray data analysis is to find differentially expressed genes across different experimental conditions [36].

5.2.1.2.1 Academic (Free) Systems and Tools

- R/Bioconductor (http://www.bioconductor.org)
- BRB Array Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html)
- TM4 (TIGR MIDAS) (http://www.tm4.org/midas.html)

5.2.1.2.2 Commercial Systems and Tools

- Partek (http://www.partek.com)
- GeneSpring GX (http://www.agilent.com/chem/genespring)

5.2.1.3 Clustering and Gene Expression Profiling

Cluster analysis is currently the most used multivariate technique to analyse microarray data. Clusters can be developed using a variety of similarity or distance metrics, among them:

- Euclidean distance

- Manhattan distance
- Correlation

Hierarchical tree clustering joins similar objects together into successively larger clusters in a bottom-up manner (i.e., from the leaves to the root of the tree), by successively relaxing the threshold distance of joining objects or sets [37,38].

The relevance-networks approach takes the opposite strategy [39]. It starts with a completely connected graph with the vertices representing each object and the edges representing a measure of association, and then links are increasingly deleted to reveal "naturally emerging" clusters at a certain threshold.

Partitional clustering algorithms, such as K-means analysis and self-organizing maps [40] which minimize within-cluster scatter or maximize between-cluster scatter, were shown to be capable of finding meaningful clusters from functional genomic data [41,42].

The reliability and quality measures of clusters, as well as multilevel visualization for the evaluation of clustering solutions, should be addressed as well [43].

5.2.1.3.1 Academic (Free) Systems and Tools

- R/Bioconductor (http://www.bioconductor.org)
- BRB Array Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html) (Excel add-in)
- CLUSTER (Eisen Lab) (http://rana.lbl.gov/EisenSoftware.htm)
- dCHIP (http://biosun1.harvard.edu/complab/dchip)
- Expression Profiler (http://www.ebi.ac.uk/expressionprofiler)
- GenePattern (http://www.broadinstitute.org/cancer/software/genepattern)

5.2.1.3.2 Commercial Systems and Tools

- GeneSpring (http://www.agilent.com/chem/genespring)
- Partek (http://www.partek.com)

5.2.1.4 Classification & Gene Expression Profiling

Classification is a supervised intelligent data analysis approach. One goal of supervised expression data analysis is to construct classifiers, like decision trees, by using one of the following methods: support vector machines (SVM), compound covariate predictor, diagonal linear discriminant analysis, nearest neighbour predictor and nearest centroid predictor.

By comparing samples, we can find classification-archetypes (class descriptions) with which differentially expressed genes are combined to distinguish between the samples and "discriminant" genes might be identified. For indicative references about microarrays and gene expression classification refer to the ULR links and references below.

5.2.1.4.1 Academic Systems and Tools

- AFFYR PAGE (http://www.cbs.dtu.dk/staff/laurent/download/affyR)
- GenePattern (http://www.broad.mit.edu/cancer/software/genepattern)
- Boosting (http://stat.ethz.ch/~dettling/boosting.html)

5.2.2 SNP Detection

Sequence variations are caused when a single nucleotide base differs between different members of species or between two chromosomes of an individual.

SNPs are very important in and are associated with polygenetic disorders, such as breast cancer [44], colon [45], non-small cell lung cancer [46], gastric [47] and prostate [48] among the others and they can also modify miRNA binding sites [49].

5.2.2.1.1 SNP Detection Tools

- PbShort (http://bioinformatics.bc.edu/marthlab/PbShort)
- ssahaSNP (http://www.sanger.ac.uk/resources/software/ssahasnp)

5.2.2.1.2 Other Tools

- TagDust: Tool to identify/eliminate artefacts from next-generation sequencing data
- ShortRead: Package for input, quality assessment and exploration of high-throughput sequence data

5.2.3 Deep Sequencing-Based Expression Analysis

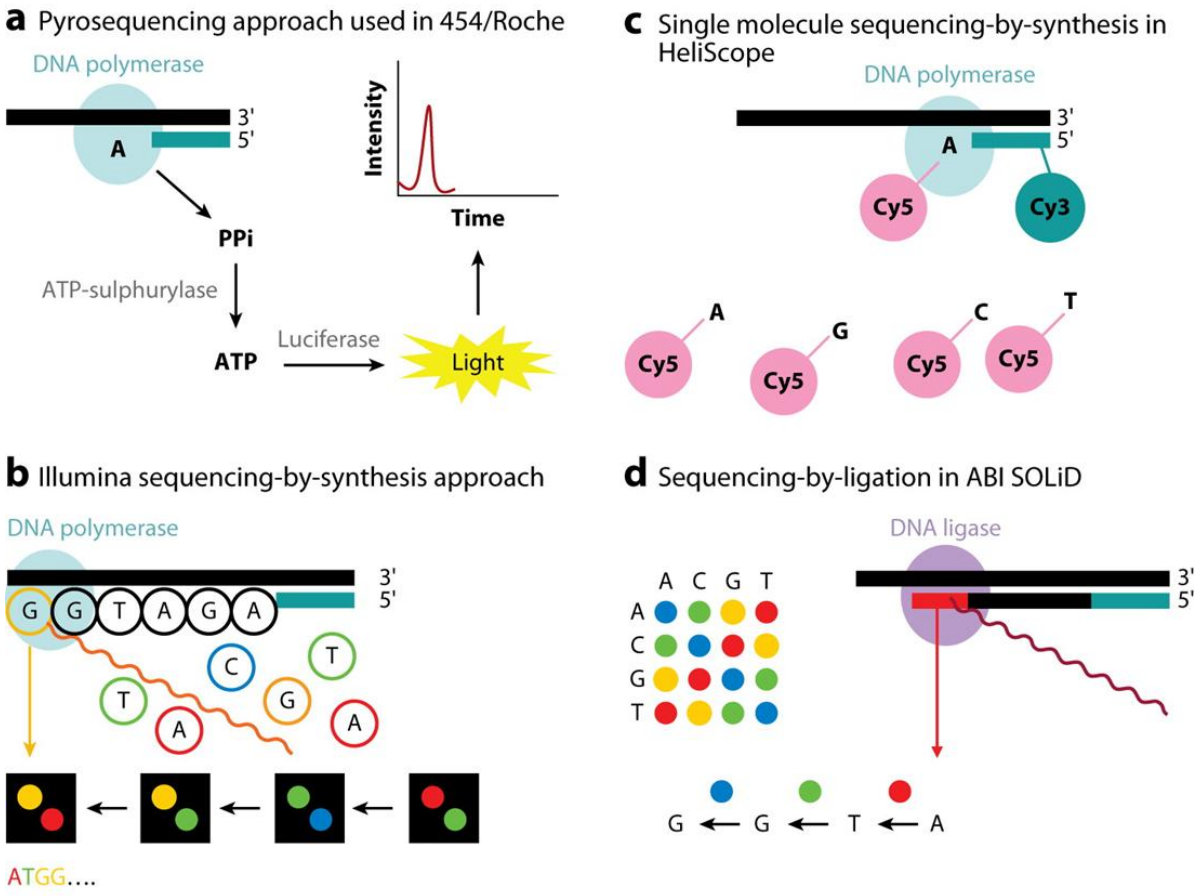
Sequencing technology has come a long way since Sanger first introduced sequencing and assembly as a methodology for sequencing entire genomes. Initially this technology was only applicable to small genomic sequences such as the genome of the bacteriophage and viruses, and bacterial artificial chromosomes (BACs), sequencing was expensive and required a great deal of manual labor in order to assemble the reads into the underlying sequence. Today, sequencing and assembly methodologies can be applied to entire mammalian genomes and most of the labor is automated. The next (or 3rd) generation sequencers came onto the scene in the early 2000's, their general characteristics include:

- Amplification of genetic material by PCR
- Ligation of amplified material to a solid surface
- Sequence of the target genetic material is determined using Sequence-by-Synthesis (using labelled nucleotides or pyrosequencing for detection) or Sequence by ligation
- Sequencing done in a massively parallel fashion and sequence information is captured by a computer

In the following table several platforms have been listed with sequencers outputs compared (for their different sequencing approaches of those platforms see the figure below):

Sequencing platform	ABI3730xl Genome Analyzer	Roche (454) FLX	Illumina Genome Analyzer	ABI SOLiD	HeliScope
Sequencing chemistry	Automated Sanger sequencing	Pyro-sequencing on solid support	Sequencing-by-synthesis with reversible	Sequencing by ligation	Sequencing-by-synthesis with virtual terminators

			terminators		
Template amplification method	In vivo amplification via cloning	Emulsion PCR	Bridge PCR	Emulsion PCR	None (single molecule)
Read length	700–900 bp	200–300 bp	32–40 bp	35 bp	25–35 bp
Sequencing throughput	0.03–0.07 Mb/h	13 Mb/h	25 Mb/h	21–28 Mb/h	83 Mb/h



The use of the deep sequencing technology is particular important in the field of Genomic Medicine, especially for:

- Implications in diagnosis, treatment and prevention
- Personalized medicine
- Low cost, ~$1000 genome

Many bioinformatics tools are available for Next Generation Sequencing (NGS) data analysis and usually they belong to one of the following categories:

- Alignment of reads to reference genome
- Assembly of de novo sequence
- Quality Control & Base Calling
- Polymorphism detection
- Genome browsing and annotation

5.2.3.1 Alignment Tools

- Cross_match is a general purpose application for comparing any two DNA sequence sets (http://www.phrap.org)
- ELAND: (from Illumina) includes ungapped alignment with a finite read length
- Exonerate is a generic tool for pairwise sequence comparison (http://www.ebi.ac.uk/~guy/exonerate)
- MAQ is a software that builds mapping assemblies from short reads generated by the next-generation sequencing machines (http://maq.sourceforge.net/maq-man.shtml)
- Mosaik pairwise aligns each read to a specified series of reference sequences (http://bioinformatics.bc.edu/marthlab/Mosaik)
- SHRiMP is a software package for aligning genomic reads against a target genome (http://compbio.cs.toronto.edu/shrimp)
- SOAP has been in evolution from a single alignment tool to a tool package that provides full solution to next generation sequencing data analysis (http://soap.genomics.org.cn) [50]
- Zoom is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis (http://www.bioinformaticssolutions.com/products/zoom)
- Novalign it is an aligner for single-ended and paired-end reads from the Illumina Genome Analyser (http://www.novocraft.com)

De novo sequencing involves assembling overlapping reads to form contiguous sequence of DNA.

5.2.3.2 Assembly Tools

- ABySS (http://www.bcgsc.ca/platform/bioinfo/software/abyss)
- ALLPATHS (http://www.broadinstitute.org/scientific-community/science/programs/genome-sequencing-and-analysis/computational-rd/computational-)
- Edena (http://www.genomic.ch/edena.php)
- Euler-SR (http://euler-assembler.ucsd.edu/portal)
- SHARCGS (http://sharcgs.molgen.mpg.de)
- SHRAP (available individually upon request)
- SSAKE (http://www.bcgsc.ca/platform/bioinfo/software/ssake)
- Velvet (http://www.ebi.ac.uk/~zerbino/velvet)

Tools to determine nucleotide base depending on signal on a sequencer produced trace file:

- PyroBayes (http://bioinformatics.bc.edu/marthlab/PyroBayes)
- Alta-Cyclic (http://hannonlab.cshl.edu/Alta-Cyclic/main.html)
- BayesCall (http://bayescall.sourceforge.net)

5.2.4 NGS and Diseases

Thanks to the development of NGS technologies, the human genome has been mapped in many individuals and with them has been growing the challenge and the opportunity to understand this large amount of data and to ultimately determine how changes in the genome lead to disease. In this era, issues and strategy related to data integration are emerging; it is expected that integrating numerous data sets and several omics data will provide more biological insights than using one data set alone (although highly informative) and only one omic, because integrating them together offers the chance to answer many questions that still remain open. Therefore, integrative analysis has become an essential part of experimental design in the era of next-generation genomics, as well as the co-operation among clinicians, computer scientists, research scientists and bench scientists. The scientific community has still to work on creating and agreeing in standard tools for next-generation data visualization, manipulation and analysis [51-54].

5.3 Pathway and Interaction Analysis

An ambitious direction is to attempt to model and infer regulatory networks globally, or along more specific subcomponents such as a pathway or a set of co-regulated genes. A major obstacle is that our knowledge of transcription and other critical molecular level mechanisms remains incomplete, especially referring to in-vivo perturbations or “noise” at various stages of regulation in molecular processes which could mark the difference between changes, often epigenetic, which may significantly affect other processes, versus those which do not.

5.3.1 Gene and miRNA Regulatory Networks

On the theoretical side, several mathematical formalisms have been applied to model genetic networks. These range from discrete models, such as Boolean networks, as in the pioneering work of Kauffman, to continuous models based on differential equations, such as continuous recurrent neural networks or power-law formalism, probabilistic graphical models and Bayesian networks. None of these formalisms appears to capture all the dimensions of gene regulation and most of the work in this field is still very preliminary. The manual inference of pathway information as it occurs e.g. in the interpretation of gene expression data [55] is assisted with the use of pre-complied protein interaction databases, like those available from Ingenuity (http://www.ingenuity.com), GeneGo (http://www.genego.com), Ariadne (http://www.ariadnegenomics.com/products/pathway-studio) and Transfac (http://www.biobase-international.com). Most of these tools are reviewed in [56,57].

Many of these tools can also be used to study networks of miRNAs. Due to the large number of discovered miRNAs in the human genome, and the hypothesis that many others have to be discovered, it is clear that the miRNA regulation of a specific target gene, protein, cellular behaviour, and its contribution to the development and progression of disease is very complex and thus the integration of both miRNA and target gene patterns of expression to then identify “network” deregulation is critical to our understanding of the role that miRNAs play as potential biomarkers and therapeutic targets. For example, recently, Volina and

colleagues examined the patterns of miRNA expression in over 4,000 human tissues: solid cancers, normal tissues and leukemia [58]. The most popular and reviewed tools are below.

5.3.1.1 MetaCore

MetaCore is a commercial bioinformatics product, developed and available by GeneGo (http://www.genego.com) that helps the research scientists in the following main topics:

- Identification of biomarkers for specific disease states
- Drug target selection and validation

Data from microarray gene expression studies, SNPs, metabolic profiles or High Content Screening (HCS) assays, can be imported for further analysis of the most relevant pathways, networks and cellular processes affected by the experimental condition. Access is commercial but a free trial is available.

5.3.1.2 Ingenuity

Ingenuity Pathways Analysis (IPA) (http://www.ingenuity.com) is a software for omics data analysis, pathway study and network interpretation derived from gene expression, microRNA, and SNP microarrays, metabolomics, proteomics and RNA-seq experiments. Also, the tool Path Designer is available to transform networks and pathways into publication quality representations. Several applications are available with a specific function and aim:

- Core analysis allows to extract signaling and metabolic pathways, performs molecular networks and define most significantly perturbed biological processes in a dataset.
- Metabolomics allows to gain biological insight into cell physiology and metabolism from metabolite data by understanding which biological processes and phenotypes user's data metabolites are involved in and what regulates their synthesis; it furthermore integrate mRNA, microRNA, SNP, proteomics, and metabolomics data for an integrated systems biology approach.
- Tox delivers, given a compound, toxicity7safety information, provides understanding of the pharmacological response and clarifies the drug mechanism of action/toxicity.
- Biomarker identifies the most promising and relevant biomarker candidates within experimental datasets.
- MicroRNA Target Filter is a tool that combines filtering methods and microRNA-mRNA content to provide information on the biological effects of microRNAs.

IPA has been broadly used and cited in many papers, although after the advent of MetaCore, the quality of IPA has been assessed as lower than MetaCore and others equivalent [57]. Access is commercial but a free trial is available.

5.3.1.3 Biobase

BIOBASE provides resources (databases, software and services) for the life sciences. They are manually curated by experts (http://www.biobase-international.com). Among them are

- BKL TransPath: Signalling pathway database.
- BKL TransFac: Knowledge base containing data on transcription factors, their experimentally proven binding sites, and regulated genes.

5.3.1.4 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG (http://www.genome.jp/kegg) is a suite of databases and associated software, integrating the function and utility of biological systems (PATHWAY and BRITE databases), genes and proteins (GENES database), and chemical compounds (COMPOUND database) and reactions (REACTION database). The PATHWAY database covers 137,977 pathways generated from 395 reference pathways, over 6,5 millions genes in their GENES database, over 17000 compounds in their COMPOUND Database and over 8000 reaction in the REACTION database. The main pathways covered include:

- Metabolism (Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Glycan, PK/NRP, Cofactor/vitamin, Secondary metabolite, Xenobiotics)
- Genetic Information Processing
- Environmental Information Processing
- Cellular Processes
- Human Diseases
- Drug Development

KEGG provides a browser which offers searching functionality together with a pictographic representation of the various pathways and several analysis tools:

- KEGG Mapper: KEGG PATHWAY and BRITE mapping tools
- KEGG Atlas: navigation tool to explore KEGG global maps
- KAAS: KEGG automatic annotation server
- BLAST/FASTA: sequence similarity search
- SIMCOMP: chemical structure similarity search
- PathPred: biodegradation/biosynthesis pathway prediction

KEGG plays an important role in oncology research. The PATHWAY database provides information relevant to the pathological processes involved in carcinoma initiation and development. Apart from the pathway-related information, KEGG also provides information on carcinoma-relevant genes and proteins with their mutant variants. KEGG is available for free usage within the terms of licensing.

5.3.1.5 WikiPathways

WikiPathways was established to facilitate the contribution and maintenance of pathway information by the biology community. WikiPathways is a collaborative platform to improve the curation of biological pathways. Building on the same user-friendly MediaWiki software that powers Wikipedia, the authors added a custom graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems. It is freely available and open at http://www.wikipathways.org.

5.3.1.6 GeneSpring

GeneSpring GX is the industry bioinformatics platform for gene expression analysis and pathways enrichment provided by Agilent (http://www.agilent.com/chem/genespring).

Several classical tools are available, like t-test, ANOVA, clustering and identification of differentially expressed genes and miRNAs as well as recent applications like the detection of alternate splicing and consideration of results at gene ontology level helping to understand and interpret the results within a biological context.

Transcriptomics, genomics, proteomics and metabolomics data can be easily analysed and visualized. The use is under commercial license.

5.3.1.7 Pathway Studio

Ariadne Pathway Studio (http://www.ariadnegenomics.com/products/pathway-studio) enables in-depth analysis of any interrelated biological data. This is particularly useful for the interpretation of gene expression or proteomics experiments.

Pathway Studio finds common regulators and associates pathway components with like-behaving biological entities and processes. All relationships are supported and validated by citations, linked to their references of origin. In summary, it assists scientists in:

- Interpreting high throughput data
- Building, expanding and analyzing pathways
- Finding relationships among genes, proteins, cell processes and diseases
- Draw and visualize pathway diagrams

It can be used under commercial license.

5.3.1.8 Cytoscape

This bioinformatics software platform to help scientists in:

- Visualizing molecular interaction networks and biological pathways
- Integrating the networks with annotations, gene expression profiles, other state data

The user can

- Input and construct molecular interaction networks from raw interaction files (SIF format), GML format or XGMML format
- Input mRNA expression profiles from tab- or space-delimited text files
- Load and save arbitrary attributes on nodes and edges
- Import gene functional annotations from the Gene Ontology (GO) and KEGG databases
- Directly import GO Terms and annotations from OBO and Gene Association files
- Load and save state of the Cytoscape session (.cys) file

The user can visualize

- Customized network data display using powerful visual styles
- Expression data mapped to node color, label, border thickness, or border colour, etc. according to user-configurable colors and visualization schemes
- Layout networks in two dimensions
- Use the network manager to easily organize multiple networks

It is also possible to

- Filter network to select subsets of nodes and/or interactions based on current data
- Find active subnetworks/pathway modules
- Find clusters (highly interconnected regions) in any loaded network

It can be freely downloaded and installed from http://www.cytoscape.org, although not all the plugins are freely available. For more information, please see [59].

5.4 References

[1] Azuaje F (2010) Bioinformatics and biomarker discovery: “omic” data analysis for personalized medicine. Wiley-Blackwell. ISBN: 978-0-470-74460-4

[2] Gavaghan D, Coveney P, Kohl P (2009) The virtual physiological human: tools and applications I Phil. Trans. R. Soc. A 367, pp. 1817-1821. doi: 10.1098/rsta.2009.0070

[3] Coveney P, Diaz V, Hunter P, Kohl P, Viceconti M (2011) The Virtual Physiological Human. Interface Focus 1:281-285

[4] Hood L, Friend S (2011) Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 8(3):184-7.

[5] Gorini A, Pravettoni G (2011) P5 medicine: a plus for a personalized approach to oncology. May 2011 Nat Rev Clin Oncol. PMID: 21629214

[6] http://www.vph-noe.eu/vphrepository/doc_download/175-vphnoeextendedslideset2011ppt

[7] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg L, Eilbeck K, et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25 (11): 1251–1255

[8] Rossi S, Sevignani C, Nnadi S, Siracusa L, Calin G. (2008) Cancer-associated genomic regions (CAGRs) and noncoding RNAs: bioinformatics and therapeutic implications. Mamm Genome. 2008 Aug;19(7-8):526-40. PMID: 18636290

[9] Rossi S, Tsirigos A, Amoroso A, Mascellani N, Rigoutsos I, Calin GA, Volinia S (2011) OMiR: identification of associations between OMIM diseases and microRNAs. Genomics. 97(2):71-6. PMID: 20974243

[10] Liang D, Meyer L, Chang DW, Lin J, Pu X, Ye Y, Gu J, Wu X, Lu K. (2010) Genetic variants in MicroRNA biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response. Cancer Res. 70(23):9765-76. PMID: 21118967

[11] Zhang B, Sun T, Zhang B, Zheng S, Lü N, Xu B, Wang X, Chen G, Yu D, Lin D (2011) Polymorphisms of GSTP1 is associated with differences of chemotherapy response and toxicity in breast cancer. Chin Med J (Engl). 124(2):199-204. PMID: 21362365

[12] Guillem V, Collado M, Terol M, Calasanz M, Esteve J, Gonzalez M, Sanzo C, Nomdedeu J, Bolufer P, Lluch A, Tormo M (2007) Role of MTHFR (677, 1298) haplotype in the risk of developing secondary leukemia after treatment of breast cancer and hematological malignancies. Leukemia. 21(7):1413-22. PMID: 17476281

[13] Calin G, Croce C (2006) MicroRNA signatures in human cancers. Nat Rev Cancer. 6:857–866

[14] Almeida M, Reis R, Calin G (2010) MYC-microRNA-9-metastasis connection in breast cancer. Cell Res. 20, 603–604

[15] Ballabio E, Mitchell T, van Kester M, Taylor S, Dunlop H, Chi J, Tosi I, Vermeer M, Tramonti D, Saunders N, et al. (2010) MicroRNA expression in Sézary syndrome: identification, function and diagnostic potential. Blood 116, 1105–1113

[16] Calin G, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik S, Iorio M, Visone R, Sever N, Fabbri M, et al. (2005) MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N. Engl. J. Med. 353, 1793–1801

[17] Rossi S, Shimizu M, Barbarotto E, Nicoloso M, Dimitri F, Sampath D, Fabbri M, Lerner S, Barron L, Rassenti L, et al. (2010) MicroRNA fingerprinting of CLL patients with chromosome 17p deletion identify a miR-21 score that stratifies early survival. Blood 116, 945–952

[18] Perkins D, Jeffries C, Jarskog L, Thomson J, Woods K, et al. (2007) microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. Genome Biol. 8:R27

[19] Wang Q, Wang Y, Minto A, Wang J, Shi Q, et al. (2008) MicroRNA-377 is up-regulated and can lead to increased fibronectin production in diabetic nephropathy. FASEB J. 22:4126–4135.

[20] Galasso M, Sana M, Volinia S (2010) Non-coding RNAs: a key to future personalized molecular therapy? Genome Med. 2010 Feb 18;2(2):12. PMID: 20236487

[21] Gupta R, Shah N, Wang K, Kim J, Horlings H, Wong D, Tsai M, Hung T, Argani P, Rinn J, Wang Y, Brzoska P, Kong B, Li R, West R, van de Vijver M, Sukumar S, Chang H (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 15;464(7291):1071-6. PMID: 20393566

[22] Wojcik S, Rossi S, Shimizu M, Nicoloso M, Cimmino A, Alder H, Herlea V, Rassenti L, Rai K, Kipps T, Keating M, Croce C, Calin G (2010) Non-coding RNA sequence variations in human chronic lymphocytic leukemia and colorectal cancer. Carcinogenesis. 31(2):208-15

[23] Rossi S, Kopetz S, Davuluri R, Hamilton S, Calin G (2010) MicroRNAs, ultraconserved genes and colorectal cancers. Int J Biochem Cell Biol. 42(8):1291-7

[24] Bertone P, Stolc V, Royce T, Rozowsky J, Urban A, Zhu X, Rinn J, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306(5705):2242-2246

[25] Birney E, Stamatoyannopoulos J, Dutta A, Guigo R, Gingeras T, Margulies E, Weng Z, Snyder M, Dermitzakis E, Thurman R, Kuehn M, Taylor C, Neph S, Koch C, Asthana S, Malhotra A, Adzhubei I, Greenbaum J, Andrews R, Flicek P, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799-816

[26] Kerr M, Churchill, G (2001) Experimental design for gene expression microarrays. Biostatistics, 2(2):183- 201

[27] Simon R, Dobbin K (2003) Experimental design of DNA microarray experiments. Biotechniques, Mar; Suppl: 16-21

[28] Yang Y, Speed T (2002) Design issues for cDNA microarray experiments. Nat Rev Genet. 3(8):579-588

[29] Wildsmith S, Archer G, Winkley A, Lane P, Bugelski P (2001) Maximization of signal derived from cDNA microarrays. Biotechniques 30, pp. 202–206, 208

[30] Schadt E, Li C, Su C, Wong W (2000) Analyzing high-density oligonucleotide gene expression array data. J. Cellul. Biochem. 80, pp. 192-202

[31] Bolstad B, Irizarry R, Astrand M, Speed T (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 19(2):185-193

[32] Kepler T, Crosby L, Morgan K (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biol 3, RESEARCH0037

[33] Tseng G, Oh M, Rohlin L, Liao J, Wong W (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. Nucleic Acids Res 29, pp. 2549–2557

[34] Yang Y, Dudoit S, Lu P, Lin D, Peng V, Ngai J, Speed T (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl. Acids Res. 30:4

[35] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96, pp. 2907–2912

[36] De Risi J, Penland L, Brown P, Bittner M, Meltzer P, Ray M, Chen Y, Su Y, Trent J Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14, pp. 457–460

[37] Eisen M, Spellman P, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95, pp. 14863–14868

[38] Iyer V, Eisen M, Ross D, Schuler G, Moore T, Lee J, Trent J, Staudt L, Hudson J, Boguski M, Lashkari D, Shalon D, Botstein D, Brown P (1999) The transcriptional program in the response of human fibroblasts to serum. Science 283, pp. 83–87

[39] Butte A, Kohane I. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput, pp. 418– 429

[40] Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43, pp. 59–69

[41] Tavazoie S, Hughes J, Campbell M, Cho R, Church G (1999) Systematic determination of genetic network architecture. Nat Genet 22, pp. 281–285

[42] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E, Golub T (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96, pp. 2907–2712

[43] Yeung K, Fraley C, Murua A, Raftery A, Ruzzo W (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17, pp. 977–987

[44] Shi H, Bevier M, Johansson R, Grzybowska E, Chen B, Eyfjörd J, Hamann U, Manjer J, Enquist K, Henriksson R, Carlson J, Brandt A, Lascorz J, Butkiewicz D, Pamula-Pilat J, Tecza K, Herms S, Hoffmann P, Hemminki K, Lenner P, Försti A (2011) Single nucleotide polymorphisms in the 20q13 amplicon genes in relation to breast cancer risk and clinical outcome. Breast Cancer Res Treat, PMID: 21630024

[45] Bartkova J, Horejsí Z, Koed K, Krämer A, Tort F, Zieger K, Guldberg P, Sehested M, Nesland J, Lukas C, Ørntoft T, Lukas J, Bartek J (2005) DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature. 434(7035):864-70. PMID: 15829956

[46] Spinola M, Leoni V, Galvan A, Korsching E, Conti B, Pastorino U, Ravagnani F, Columbano A, Skaug V, Haugen A, Dragani T. (2007) Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. Cancer Lett. 251(2):311-6. PMID: 17223258

[47] Milne A, Carneiro F, O'Morain C, Offerhaus G (2009) Nature meets nurture: molecular genetics of gastric cancer. Hum Genet. 126(5):615-28. PMID: 19657673

[48] Harries L, Perry J, McCullagh P, Crundwell M (2010) Alterations in LMTK2, MSMB and HNF1B gene expression are associated with the development of prostate cancer. BMC Cancer. 10:315. PMID: 20569440

[49] Nicoloso M, Sun H, Spizzo R, Kim H, Wickramasinghe P, Shimizu M, Wojcik S, Ferdin J, Kunej T, Xiao L, Manoukian S, Secreto G, Ravagnani F, Wang X, Radice P, Croce C, Davuluri R, Calin G (2010) Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. Cancer Res. 70(7):2789-98. PMID: 20332227

[50] SOAP: short oligonucleotide alignment program (2008) Bioinformatics, Vol. 24 no.5 2008, pages 713–714 doi:10.1093/bioinformatics/btn02

[51] Hawkins R, Hon G, Ren B. (2010) Next-generation genomics: an integrative approach. Nat Rev Genet. 11(7):476-86. Review. PMID: 20531367

[52] Lee W, Jiang Z, Liu J, Haverty P, Guan Y, Stinson J, Yue P, Zhang Y, Pant K, Bhatt D, Ha C, Johnson S, Kennemer M, Mohan S, Nazarenko I, Watanabe C, Sparks A, Shames D, Gentleman R, de Sauvage F, Stern H, Pandita A, Ballinger D, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 465(7297):473-7. PMID: 20505728

[53] Parsons D, Jones S, Zhang X, Lin J, Leary R, Angenendt P, Mankoo P, Carter H, Siu I, Gallia G, Olivi A, McLendon R, Rasheed B, Keir S, Nikolskaya T, Nikolsky Y, Busam D, Tekleab H, Diaz L, Hartigan J, Smith D, Strausberg R, Marie S, Shinjo S, Yan H, Riggins G, Bigner D, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu V, Kinzler K (2008) An integrated genomic analysis of human glioblastoma multiforme. Science. 321(5897):1807-12. PMID: 18772396

[54] Kim J, Dhanasekaran S, Prensner J, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M, Hu M, Sam L, Grasso C, Maher CA, Palanisamy N, Mehra R, Kominsky H, Siddiqui J, Yu J, Qin Z, Chinnaiyan A (2011) Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. 21(7):1028-41. PMID: 21724842

[55] Apica G, Ignjatovicb T, Boyerb S, Russell R (2005) Illuminating drug discovery with biological pathways. FEBS Letters 579:1872–1877

[56] Bonetta L (2004) Bioinformatics – from genes to pathways, Nature Methods 1(2):169

[57] Shmelkov E, Tang Z, Aifantis I, Statnikov A (2011) Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. Biol Direct. 6:15.

[58] Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, Marchesini J, Mascellani N, Sana M, Abu Jarour R, Desponts C, Teitell M, Baffa R, Aqeilan R, Iorio M, Taccioli C, Garzon R, Di Leva G, Fabbri M, Catozzi M, Previati M, Ambs S, Palumbo T, Garofalo M, Veronese A, Bottoni A, Gasparini P, Harris C, Visone R, Pekarsky Y, de la Chapelle A, Bloomston M, Dillhoff M, Rassenti L, Kipps T, Huebner K, Pichiorri F, Lenze D, Cairo S, Buendia M, Pineau P, Dejean A, Zanesi N, Rossi S, Calin G, Liu C, Palatini J, Negrini M, Vecchione A, Rosenberg A, Croce C (2010) Reprogramming of miRNA networks in cancer and leukemia. Genome Res. 20(5):589-99. PMID: 20439436

[59] Shannon P, Markiel A, Ozier O, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11): 2498–504. doi:10.1101/gr.1239303. PMC 403769. PMID 14597658.

6 High Performance and Cloud Computing

6.1 High-Performance and High-Throughput Computing Infrastructures in Europe

The European distributed computing landscape is shaped by the three biggest e-Infrastructures: EGI and PRACE as well as smaller initiatives driven by end-users communities, e.g. MAPPER. Each e-Infrastructure addresses the needs of specific research groups. PRACE forms the top layer of the European distributed computing ecosystem and operates state of the art peta- and exascale High Performance Computing resources, whereas EGI, established the bottom layer of the ecosystem, brings together national and regional providers to enable collaboration of research communities across Europe. Together the projects complement each other and provide a complete set of computing services to the European scientific community. MAPPER is focused on integrating services in the European e-Infrastructure to allow end-users an easier access via various application tools to different supercomputers and clusters, including many production sites.

6.1.1 Partnership for Advanced Computing in Europe (PRACE)

PRACE (http://www.prace-project.eu) is a unique persistent pan-European Research Infrastructure for High Performance Computing. PRACE provides Europe with world-class systems for world-class science and strengthens Europe's scientific and industrial competitiveness. PRACE maintains a pan-European HPC service consisting of up to six top of the line leadership systems (Tier-0) well-integrated into the European HPC ecosystem. Each system provides computing power of several Petaflop/s (one quadrillion operations per second) in midterm. On the longer term Exaflop/s (one quintillion) computing power will be targeted by PRACE. This infrastructure is managed as a single European entity. The partnership was established through the close collaboration of the European countries that prepared the legal, financial, and technical basis of the project. The First Implementation Phase of PRACE is in line with the objectives of the PRACE Research Infrastructure organisation: from coordinated system selection and design, coherent management of the distributed infrastructure, software deployment, porting, scaling, optimising applications and promoting and advancing application development and the skills.

A short description of the available Tier-0 systems is:

- IBM Blue Gene/P – JUGENE – hosted by GCS in Jülich, Germany. JUGENE has a peak performance of 1 Petaflop. It is composed of 294,912 processing cores with 4 cores forming a node with 2 GB of memory for a total of 147 TB. The total available capacity for JUGENE in this call is 360 million compute core hours. The allocation period for JUGENE ends on August 31st, 2012.
- Bull Bullx cluster – CURIE – funded by GENCI and installed at CEA, Bruyères-Le-Châtel, France. CURIE is composed by 2 different partitions:
    - A fat node partition open to PRACE calls since January 2011 and composed by 360 nodes with 32 cores per nodes, for a peak performance of 105 TFlops
    - A thin node partition, open to PRACE calls in Q1 2012 and composed by 5040 blades with 16 cores per node, for a peak performance of up to 1.5 PFlops

  The total available capacity in this call for CURIE is:
    - On the thin nodes partition: 125 million compute core hours, with an 8 months
    - On the fat nodes partition: 31 million compute core hours, with a one year allocation

- Cray XE6 – HERMIT – hosted by GCS in HLRS, Stuttgart, Germany. HERMIT has a peak performance of 1 Petaflop and is designed for sustained application performance and highly scalable applications. It is composed of 3552 dual socket nodes equipped with AMD Interlagos Processors leading to overall 113664 processing cores. Nodes are equipped with 32GB or 64GB main memory. The total available capacity in this call for HERMIT is 160 million compute core hours

There is also second level of resources provided by DEISA that consists of regional or national centers (Tier-1) which are available on the following architectures: Cray (XT and XE), IBM Blue Gene/P, IBM Power 6, Intel and PowerPC Clusters (various processor and memory configurations) and hybrid systems (clusters with GPGPU accelerators).

In total the number hours on the Tier-1 resources is more than 54 million core hours.

For instance PSNC is in the deployment phase of 224 node cluster (internal connections: Infiniband QDR) Each node has 2 AMD processors (12 or 24 cores) and 48 GB RAM . Additionally there are 336 NVIDIA graphic cards (GPU) associated with CPU nodes.

To get access to the PRACE e-Infrastructure, research groups have to submit project proposals describing resource requirements, methods and models used, research goals, and the scientific merit of their work. The proposals are evaluated by a committee composed of PRACE representatives and external experts based on a set of predefined criteria. After a number of proposals is selected a project is created and assigned to each of the respective research groups. PRACE takes over management of the projects including such activities as the allocation of resources, creation of user accounts and maintenance of authorization and accounting facilities. This allows the users to focus on the research but, at the same time, introduces some constrains since user requests can be processed only by PRACE representatives. To access PRACE resources and services each user has to authenticate him- or herself. The authentication process adopted by PRACE is based on the Public Key Infrastructure (PKI) scheme. As mentioned above the scheme is well-established across Europe. The majority of research facilities in France, Germany, or the Netherlands, for instance, are capable of issuing X.509 certificates that can be used to authenticate to PRACE resources. However, to get a certificate a person has to be physically present at the so-called Registration Authority. In cases when this is inconvenient or not possible, a number of alternative schemes are available. The certification process is not as seamless across all Europe and in some cases might require more time and effort. To circumvent this group certificates shared by several users could be used. Unfortunately, this is not possible since the PRACE policy demands that a certificate uniquely identifies a user.

6.1.2 European Grid Infrastructure (EGI)

Building a world-class pan-European High Performance Computing Service and infrastructure involves the scientific and industrial user communities with their leading edge applications. This needs to be done in a rapidly evolving context, where technologies change continuously and where the science focus changes as results are obtained and new directions are explored. The ultimate goal of EGI (http://www.egi.eu) is to provide European scientists and their international partners with a sustainable, reliable e-Infrastructure that can support their needs for large-scale data analysis. This is essential in order to solve the big questions facing science today, and in the decades to come.

Access to EGI resources is granted to users based on their virtual organization membership, where a Virtual Organization (VO) is a dynamic set of individuals and institutions active in a specific scientific area. From this point of view VOs are similar to virtual communities defined in DEISA. Yet a VO, generally speaking, has more responsibility and control over users, resources and services belonging to the VO. As such EGI defines generic policies for

authentication, authorization and accounting. Each VO is able to define and follows its internal rules as long as they are inline with EGI.

Similarly to PRACE, EGI relies on the PKI scheme for authentication. This means that users require an X.509 certificate to gain access to EGI resources. Yet, unlike the other European e-Infrastructures, EGI does not enforce the one-to-one mapping for users, user accounts and certificates. This means that several users are allowed to share a common certificate associated to a pool account and use it for authentication with EGI resources and services. This is a very useful and desired feature since it allows VOs to setup internal authentication mechanisms for simplifying the authentication process, for example in cases when not everyone who needs to access the e-Infrastructure is able to get a personal certificate. Regular user accounts requiring a personalized certificate are, of course, supported as well.

EGI does not define authorization and accounting policies and delegates this task to the individual VOs. As such, the VOs are able to internally control access to resources and services and specify mechanisms used for accounting. Policies defined by each VO should be inline with the local legislative regulations. For instance, user related data should be handled according to the enforced laws.

6.1.3 Multiscale Applications on European e-Infrastructures (MAPPER)

The MAPPER (http://www.mapper-project.eu) computing infrastructure, in close collaboration with EU-wide EGI and PRACE infrastructures, will extend existing capabilities offered by Uniform Interface to Computing Resources (UNICORE, http://www.unicore.eu) and gLite (http://glite.cern.ch) by deploying easy-to-setup QosCosGrid (http://www.qoscosgrid.org) middleware services to support new features for scientists, such as advance reservation of computing resources, co-allocation or multi-cluster parallel job executions. Various end-user communities, including clinical researchers that require on-demand access to high-performance computing facilities, will benefit immediately from the MAPPER infrastructure as some resource providers have already offered their allocations for MAPPER users. Technically speaking, the integration of QosCosGrid with HPC resources in PRACE and computing clusters in EGI is be done based on well-defined APIs and appropriate extensions to the following standards: OGF DRMAA 2.0, OGF HPC-Profile, OGF JSDL or OGF SAGA. In order to support interoperability in MAPPER we will take advantage of Vine Toolkit and SAGA libraries offering a uniform access for various end-user tools to a number of underlying grid middleware services used in Europe and worldwide. Additional support for programming and execution environments like QCG-OMPI, ProActive or workflows (e.g. Kepler, http://kepler-project.org) will be also provided.

As the QosCosGrid middleware has been successfully deployed and is currently supported in PL-Grid, computing resources from Poznan Supercomputing and Networking Center will be dedicated for demanding jobs in the p-medicine project. Moreover, PL-Grid has initiated functioning of the National Grid Initiative (NGI) in Poland as a part of EGI infrastructure. PL-Grid aims at significantly extending the amount of computing resources provided to scientific communities (by approximately 215 TFlops of computing power and 2500 TB of storage capacity), including p-medicine clinical researchers.

6.2 Cloud Computing

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. The cloud model of computing promotes availability.

6.2.1 Background Technologies

6.2.1.1 (Hardware) Virtualization

Virtuallization is the creation of a virtual machine (guest system) that acts like a real one in the context of environment for the applications (operation system). It can be run on some other machine (host system). Virtualization is the key technology for dynamic, on-demand environment creation in cloud systems. Three types of virtualization exist:

- Full virtualization: environment for virtual machine fully simulates underlying hardware, there is no need to modify guest operating system
- Partial virtualization: not all features of the hardware platform are simulated, it requires some modifications for the guest systems
- Paravirtualization: the VA does not necessarily simulate hardware but instead (or in addition) offers a special API that can only be used by modifying the guest system

6.2.1.2 Grid Computing

The idea to provide unified access to heterogeneous and geographically distributed hardware resources (computational nodes, storages). The key aspects of the grid implementation are to create middleware layer consists of services that manages the dynamically changing environment, with resources which are under many different administrative domains. The most important areas for grid technologies are:

- Resource management: assigning the computational task to the available machines
- Data management: integration of data with computation
- Security and user management (Virtual Organization): providing seamless access to resources in different administrative domains

Grid computing was finally replaced by cloud computing paradigm in many use cases because it failed to provide required QoS for the end-user (problem of availability of resources) and did not solve the problem of infrastructure heterogeneity (differences in hardware, operating systems configurations, installed library versions, etc.).

6.2.1.3 Service Oriented Architecture (SOA)

SOA defines software infrastructure as a set of loosely coupled software services that support the requirements of the business processes specified by the users. In a SOA environment, resources on a network are made available as independent services that can be accessed without knowledge of their underlying platform implementation. A service-oriented architecture is not tied to a specific technology. It may be implemented using a wide range of interoperability standards but the most important set of standard in the context of SOA is Web Services (WS). The key issue is that all services that building some environment are independent from each other, they publish their interfaces and that is enough for any other entity in the system to invoke their methods (meaning “use” the service). The idea of services as well-defined and self-contained that provides some business functionality fits ideally to the idea of software deployment in the cloud.

6.2.2 Service Models

- Software as a Service (SaaS) provides users application as a services available on-line via thin client interface (web browser)
- Platform as a Service (PaaS) provides not only application but also surrounding software environment (services) required for the application to run.

- Infrastructure as a Service (IaaS) provides whole IT environment in the terms of processing units, storage nodes, network connections etc., and make it available to deploy any services and applications.

6.2.3 Deployment Models

- Private cloud: Cloud infrastructure operated by some organization. The access to the resources is limited to some group of users based on rules defined by operator
- Community cloud: Infrastructure shared by group of people from different organizations sharing the same goal
- Public cloud: Infrastructure available for the general public, served by some organization based on commercial rules
- Hybrid cloud: Mixture or two or more cloud models

6.2.4 Commercial Cloud Providers

6.2.4.1 Amazon Web Service (AWS)

The Amazon Web Service (http://aws.amazon.com) is a cloud computing platform offered by Amazon as a set of services that could work together to provide cloud functionality. The most important are: Amazon EC2 (Elastic Compute Cloud) and Amazon S3 (Simple Storage Service). The first one is the central component of the cloud architecture responsible for providing virtual machines to run client's computations. Amazon S3 is the cloud storage service that provides web service interfaces (REST, SOAP) for storing the data that then can be used in the cloud environment (e.g. as an input for computation).

The most important ideas of AWS is to provide scalability, high availability and low latency.

6.2.4.2 Google App Engine (GAE)

GAE (http://www.google.com/enterprise/cloud/appengine) is the cloud computing platform (PaaS) provided by Google. It allows to run web application on Google resources in an easy way. GEA supports application written in several programming languages (e.g. Java, Python, Groovy, JRuby, Scala). The most important features of GEA are:

- Dynamic web serving, with full support for common web technologies
- Persistent storage with queries, sorting and transactions
- Automatic scaling and load balancing, APIs for authenticating users and sending e-mail using Google Accounts
- Fully featured local development environment that simulates GEA on your computer
- Task queues for performing work outside of the scope of a web request
- Scheduled tasks for triggering events at specified times and regular intervals

6.2.4.3 Microsoft Azure

Azure (http://www.microsoft.com/windowsazure) is the cloud platform designed and developed by Microsoft that can be used to build, host and scale web application through Microsoft datacenters. There are three product brands offered:

- Windows Azure: Operating system providing computation and storage facilities (consists of three components: Compute, Storage and Fabric)

- SQL Azure: Cloud-based version of SQL server
- Windows Azure AppFabric: Environment supporting application in the cloud (access control, tools and APIs to develop and host applications, service bus, caching)

6.2.4.4 Rackspace Cloud

Rackspace Cloud (http://www.rackspace.com/cloud) is a web application hosting and cloud platform provider. The solution is based on following services:

- Cloud Files: Cloud storage infrastructure very similar to Amazon's S3. It provides RESTful API and open-source client code
- Cloud Servers: Environment for computational nodes deployment based on Xen virtualization technology
- Cloud Sites: Web hosting platform built on scalable hardware infrastructure

6.2.5 Open Cloud Solutions

6.2.5.1 Eucalyptus

Eucalyptus (http://open.eucalyptus.com) is a software platform for the implementation of private cloud computing on computer clusters. The platform provides a single interface for accessing computing infrastructure (machines, network, and storage). It's modular design with extensible Web-services architecture allows to provide a variety of APIs towards users via client tools. Currently, Eucalyptus implements the industry-standard Amazon Web Services (AWS) API, which allows the interoperability of Eucalyptus with existing AWS services and tools. The most important features are:

- Compatibility with Amazon Web Services API
- Installation and deployment from source or DEB and RPM packages
- Secure communication between internal processes via SOAP and WS-Security
- Support for Linux and Windows virtual machines (VMs)
- Support for multiple clusters as a single cloud
- Elastic IPs and Security Groups
- Users and Groups Management
- Accounting reports
- Configurable scheduling policies and SLAs.

Eucalyptus consists of set of services:

- Cloud Controller (CLC) is responsible for exposing and managing the underlying virtualized resources
- Walrus is the cloud storage service (interface compatible with Amazon's S3)
- Cluster Controller (CC) controls the execution of VMs running on the nodes and manages the virtual networking between VMs and between VMs and external users
- Storage Controller (SC) provides block-level network storage that can be dynamically attached by VMs
- Node Controller (NC) controls VM activities, including the execution, inspection, and termination of VM instances

6.2.5.2 OpenNebula

OpenNebula (http://opennebula.org) is an open-source cloud computing toolkit for managing heterogeneous distributed data centre infrastructures. The OpenNebula toolkit manages a data centre's virtual infrastructure to build private, public and hybrid IaaS clouds. OpenNebula orchestrates storage, network, virtualization, monitoring, and security technologies to deploy multi-tier services (e.g. compute clusters) as virtual machines on distributed infrastructures, combining both data centre resources and remote cloud resources, according to allocation policies.

6.2.5.3 OpenStack

OpenStack (http://www.openstack.org) is an IaaS cloud computing project by Rackspace Cloud and NASA. Currently more than 100 companies have joined the project among which are Citrix Systems, Dell, AMD, Intel, Canonical, and Cisco. It is free open-source software released under the terms of the Apache License. There are two main system components:

- OpenStack Compute (Nova) is an open-source software designed to provision and manage large networks of virtual machines, creating a redundant and scalable cloud computing platform. It gives you the software, control panels, and APIs required to orchestrate a cloud, including running instances, managing networks, and controlling access through users and projects. OpenStack Compute strives to be both hardware and hypervisor-agnostic, currently supporting a variety of standard hardware configurations and seven major hypervisors.
- OpenStack Object Store (Swift) is an open-source software for creating redundant, scalable object storage using clusters of standardized servers to store petabytes of accessible data. It is not a file system or real-time data storage system, but rather a long-term storage system for a more permanent type of static data that can be retrieved, leveraged, and then updated if necessary. Primary examples of data that best fit this type of storage model are virtual machine images, photo storage, e-mail storage and backup archiving. Having no central “brain” or master point of control provides greater scalability, redundancy and permanence.

6.2.5.4 Nimbus

Nimbus (http://www.nimbusproject.org) is an open-source toolkit that allows you to turn your cluster into an IaaS cloud. Feature highlights include:

- Three sets of remote interfaces: Amazon EC2 WSDLs, Amazon EC2 Query API and grid community WSRF. Read more about interfaces
- Storage implementation compatible with S3 REST API
- Virtualization implementation is based on Xen and KVM virtualization technologies
- Can be configured to use familiar schedulers like PBS or SGE to schedule VMs
- Launches self-configuring virtual clusters with one click
- Defines an extensible architecture to customize the software to the project needs

6.2.6 Cloud Technology Standards

6.2.6.1 Open Cloud Computing Interface (OCCI)

The OCCI comprises a set of open community-lead specifications delivered through the Open Grid Forum. OCCI is a protocol and API for all kinds of Management tasks. OCCI was originally initiated to create a remote management API for IaaS model-based Services, allowing for the development of interoperable tools for common tasks including deployment, autonomic scaling and monitoring. It has since evolved into a flexible API with a strong focus on integration, portability, interoperability and innovation while still offering a high degree of extensibility. The current release of the Open Cloud Computing Interface is suitable to serve many other models in addition to IaaS, including e.g. PaaS and SaaS.

6.2.6.2 Cloud Data Management Interface (CDMI)

The Cloud Data Management Interface defines the functional interface that applications will use to create, retrieve, update and delete data elements from the Cloud. As part of this interface the client will be able to discover the capabilities of the cloud storage offering and use this interface to manage containers and the data that is placed in them. In addition, metadata can be set on containers and their contained data elements through this interface. This interface is also used by administrative and management applications to manage containers, accounts, security access and monitoring/billing information, even for storage that is accessible by other protocols. The capabilities of the underlying storage and data services are exposed so that clients can understand the offering.

7 Semantic Mediation and Data Integration

Today's biomedical research involves managing heterogeneous data from different provenance [1-5]. Clinicians must face the difficulties involved in dealing with incompatible schemas, formats and data codifications. The aim of biomedical semantic mediation (or data integration – systems is to alleviate this problem by hiding such complexities from end-users.

There exist different approaches for tackling this problem. Systems either adopt centralized approaches – e.g. data warehouses, where data is stored locally – or federated approaches – where data is left at the sources and accessed on demand. The selection of either approach depends on the type of solution to be deployed. Data warehouses might deal with data privacy issues and with outdated data. They however provide better efficiency and allow tighter control to data managers over what data will be available. Federated approaches always access updated data, but suffer from efficiency issues. During the last 15 years, numerous systems have been developed, often targeting specific problems or areas. Next subsection provides further detail on the existing approaches for semantic mediation.

In any of the approaches, a series of heterogeneities must be solved. These are divided in syntactic heterogeneities and semantic heterogeneities. The former refer to those due to differences in the access interface, querying language and database models. The latter are caused by different data representations – schemas or instances. Syntactic heterogeneities are often dealt with wrapper architectures, in which specific database wrapper modules provide a uniform interface to all databases to be integrated.

7.1 Data Integration Approaches

The two main approaches in semantic mediation are, as described previously, centralized and federated – also called data translation and query translation, respectively. Information linkage can be considered as a third approach in this area, although it does not deal with all the problems the other two deal with. All these approaches are described in detail below.

7.1.1 Information Linkage

The simplest and most often employed approach for data integration is information linkage (IL). This method is mostly used in websites or web-based databases. It consists simply on offering the user links to related items, like in the examples PubMed, Medline, Prosite, etc.

The main advantages are the simplicity of implementation, and the ease of use for the end-users, as they are used to work with hyperlinks on their daily basis. The downside is that it offers limited features, and no real integration is performed – the user is simply offered to navigate through relations in a unidirectional way. This approach cannot cope with serious data integration requirements and thus becomes insufficient for many research fields.

7.1.2 Data Translation

Data translation has often been adopted in business applications where a single organization has control over the databases to integrate. The central repository can be built taking into account the specific needs to meet by the applications, and the uses that the end-users will make of it. The data translation approach offers a tighter control over what is available to the end-users since the available data is selected at "push" time when data is loaded to the repository from the external databases. Having the data collected in a single store improves efficiency and allows easier implementation of high-level tools like visualization, KDD services, etc. In case of biomedical research dealing with cancer, clinical or genetic

information, the data warehouse can act as the central node storing and managing data for subsequent research [6-8]. In case of third party database integration, out of the control of data warehouse developers, care must be taken to obtain consent for the data copy process.

7.1.3 Query Translation

Query translation approaches rely on a virtual schema that represents the space of queries that the user can submit to the system. It is called virtual because no data is stored centrally. Instead, each query is dynamically translated into an array of subqueries for the databases to integrate, and their single results are merged into a global result which is presented to the end-user as answer to his initial query [9]. The translation process is supported by *mappings*, i.e. translations from the virtual schema to the underlying schemas. The benefits are better adaptability upon changes or new databases in include, and avoiding returning outdated data. Nevertheless, in environments in which we do not foresee frequent changes and in which we are allowed to copy data in a central node, the data warehouse offers a better solution, as it avoids the performance penalties associated to the query dynamic translation.

There are four different query translation systems, according to existing reviews on the area [10,11]. Namely, i) pure mediation, ii) global conceptual schema, iii) multiple conceptual schemas and iv) hybrid approaches. Other studies divide systems depending on the approach employed in the query translation process: i) global as view [12] and ii) local as view [13]. The former method aims to define the global schema in terms of the underlying schemas, while the latter opts for the opposite. In practical terms, these differ in the relatively low computational resources needed to translate queries with mappings defined in a global as view model, compared to the better adaptability upon changes in existing schemas (or inclusion of new ones) in models based on local as view.

7.2 Examples of Biomedical Database Integration Initiatives

Huge amounts of resources have been dedicated to solve the problem of data heterogeneity in biomedicine. Several past and present international projects (see Table 1) include this issue as a central problem to tackle.

Name	Domain	Founding source	Duration	# Partners	Nationality
Birn	Biomedical	US NIH	Since 2001	35	International
ACGT	Post-genomic clinical trials	EC	2006-2010	25	European
caBIG	Cancer	US NCI	Since 2003	Over 80	National (US)
HeC	Pediatrics	EC	2006-2009	14	European
Infogenmed	Genetic, medical	EC	2001-2004	5	European

Table 1: Past and current projects on database integration

7.2.1 Database Integration Systems

Apart from the mention projects, there have been numerous initiatives to develop data integration systems for the biomedical domain. Each approach opts for different design characteristics, depending on the specific requirements to meet. Usually, closed environments with fixed sets of databases select data warehouses, while more open systems expecting to grow with time opt for query translation approaches. Below, short descriptions of existing data integration systems are given. Table 2 gives details about these systems.

System	Integration approach	Query language	Transparency	Semantic model
JXP4BIGI (framework)	Warehouse	Extended SQL	-	Relational
GeneMapper	Warehouse	-	Yes	GAM (EAV evolution)
Atlas	Warehouse	SQL	No	Relational
iProClass	Warehouse	Form-based	No	Relational
DataFoundry	Hybrid	SQL	No	Relational
TINet	Hybrid	OPM Multidatabase query language	No	Object-Protocol Model
BioDataServer	Federated (with cache)	SQL	Yes	Relational
BioBench	Federated	-	Yes	Object oriented
Kleisli	Federated	CPL	Yes	CPL
KIND	Federated	XML	Yes	F-logic
TAMBIS	Federated	GRAIL (graphically constructed)	Yes	Ontologies
P/FDM	Federated	Daplex	Yes	FDM
SEMEDA	Federated	Form-based	Yes	Ontologies
DiscoveryLink	Federated	SQL	Yes	Relational
BioBroker	Federated	XQuery	Yes	XML
BioMediator	Federated	PQL	Yes	Ontologies
INDUS	Federated	-	Yes	Ontologies

GeXpert	Federated	-	Yes	-
OntoFusion	Federated	RDQL	Yes	Ontologies
BioFuice	Peer-to-peer	Graphically constructed	No	-
Bio2RDF	Federated	SPARQL	Yes	Ontologies
ACGT SM	Federated	SPARQL	Yes	Ontologies

Table 2: Most popular database integration systems, with main characteristics attached

7.2.1.1 JXP4BIGI

JXP4BIGI [14] was developed as an independent framework allowing heterogeneous data integration for constructing biological data warehouses and targeted at data integrators. Its functionality is distributed across the four components:

- XML bio-entity templates for representing custom bio-entities built by biologists, e.g. a gene or protein
- SQL-based query and extraction logics, capable of defining the elements and attributes to be retrieved from the integrated repositories
- Generalized wrappers providing uniform access to syntactically heterogeneous interfaces
- JXP processor in charge of organizing and executing the tasks involved in the data retrieval process

7.2.1.2 GeneMapper

GeneMapper [15] is a data warehouse system that allows integrating biological databases in a central repository. GeneMapper avoids the user of a global schema. Instead of that, a generic data model called GAM is adopted. GeneMapper has been tested with large-scale functional gene profiles and public biological data sources, such as LocusLink and Unigene.

7.2.1.3 Atlas

Atlas [16] is a biological data warehouse aimed at bioinformatics doing research in that area. SQL access is possible to each stored databases by creating a relational data model for each of them. Integration is done by cross-referencing protein sequence and gene identifiers.

7.2.1.4 iProClass

iProClass [17] is a data infrastructure offering protein data integration. It employs a data warehouse approach for fast access to integrated data, storing data from the UniProtKB and iProClass databases. It includes navigational capabilities linking to the data sources. On top of that, a user interface is provided to search, retrieve and analyse the integrated data.

7.3 DataFoundry

DataFoundry [18] is a hybrid integration system targeting homogeneous access to scientific data. It maintains a local data store that allows improving efficiency by caching frequently accessed. In addition, federated access to data sources is available. A global relational schema represents a view of the integrated sources that can be queried using SQL.

7.3.1.1 TINet

TINet [19] adopts a hybrid approach to implement biological data integration. A federated access model is supported for most databases but GenBank and SwissProt are maintained centrally. The result is a good balance between flexibility and performance. The system's focus is on overcoming syntactic rather than semantic heterogeneities. No global, uniform view of the integrated sources exists. Little more than links to distributed data are provided.

7.3.1.2 BioDataServer

BioDataServer [20] is a mediator-based system for homogeneous access to distributed life science databases, with focus on genomic databases. It provides transparent access to the integrated sources through a global relational data model (with SQL query possibility) and adopts a federated approach including caching to fasten retrieval of frequently queried data.

7.3.1.3 BioBench

BioBench [21] is a federated system for integrating semi-structured, heterogeneous bioinformatics databases. It is designed to enable access to unstructured data repositories, e.g. flat file repositories. It employs an object-oriented data model to support data integration.

7.3.1.4 Kleisli

Kleisli [22,23] is a system designed to perform integrated queries across distributed and heterogeneous databases. A federated approach is adopted for integrating biological databases and tools, like BLAST. Its type-inference system makes unnecessary the adoption of a global data model for describing the sources of data, enhancing system flexibility.

7.3.1.5 KIND

KIND [24] is a wrapper/mediator-based architecture for the biological domain. The wrapper level takes care of syntactic heterogeneities, while the mediator level provides the semantic interoperability through F-logic schemas of the sources.

7.3.1.6 TAMBIS

TAMBIS [25] is a database integration system for the molecular biology and bioinformatics area. Developers of TAMBIS focused on offering high transparency level while allowing users to perform complex queries on the integrated repositories. It adopts a federated approach, using a self-designed ontology, TaO [26], as a global schema.

7.3.1.7 P/FDM

P/FDM [27,28] is a federated data integration system for integrating heterogeneous biological data sources. It employs the functional data model (FDM) to describe the global schema and integrated repository schemas. Queries are formulated in the Daplex query language.

7.3.1.8 SEMEDA

SEMEDA [29] is a federated system for semantic integration of biological databases. Users are hidden from the internal structure of data sources. SEMEDA uses a custom ontology containing small top-level biological concepts. Other domain-related ontologies are employed as controlled vocabularies.

7.3.1.9 DiscoveryLink

DiscoveryLink [30] is a federated system for semantic integration of life science data sources. DiscoveryLink was born from the fusion of the Garlic and the DataJoiner systems. It adopts a wrapper-based architecture, with wrappers dealing with interfacing with the sources, and a middleware engine processing end-user SQL queries.

7.3.1.10 BioBroker

BioBroker [31] is an integration system created from a framework for constructing integration systems in the biological domain. The framework provides wrapper construction schemes for building wrappers for relational sources and XML documents. BioBroker was designed to integrate data from repositories like EMBL, SWISS-PROT, PDB, MICADO, DIP and BIND.

7.3.1.11 BioMediator

BioMediator [32,33] is a system for solving queries across an integrated set of heterogeneous databases. It uses a federated approach for providing transparent access to the sources. Global schema is represented through an ontology.

7.3.1.12 INDUS

INDUS [34] is an ontology-based system to integrate heterogeneous biological data sources. Users can define their own ontology to reflect their domain view on the underlying data.

7.3.1.13 GeXpert

GeXpert [25] is a framework for the integration of heterogeneous biological data sources. It uses open standards to enhance flexibility and adoption. The GeXpert data integration system was built with this framework with the goal of offering integrated access to bioinformatics resources focused on metabolic pathway reconstruction.

7.3.1.14 OntoFusion

Ontofusion [36] is a system designed to perform integration at the schema level. It is based on ontologies for representing the source schemas and the global schema. Integration is achieved in two steps: first, virtual schemas for each of the sources are constructed. Afterwards, automatic unification of the schemas takes place to build the global schema.

7.3.1.15 BioFuice

BioFuice [37] is a peer-to-peer-based biological data integration system. Unlike other approaches, it only stores bilateral mappings with integrated sources. The inclusion of a new repository only needs a mapping with one of the existing sources, facilitating the task of updating the global schema. BioFuice was developed as an extension of the iFuice system [38] in order to include new features and capabilities targeted at the biomedical domain.

7.3.1.16 Bio2RDF

Bio2RDF [39] aims to enable RDF-based accessed to a myriad of biological public resources, such as KEGG, PDB, MGI and HGNC among others over the web.

7.3.1.17 ACGT Semantic Mediator

The European project ACGT included in its task the development of a semantic mediation layer for accessing data from clinical trials [40]. It adopted a federated approach, with a wrapper-mediator architecture. An ontology built within the same project, the ACGT Master Ontology [41], served as global schema. Users could perform SPARQL queries in terms of this ontology to access the underlying sources.

7.4 Conclusions

Current biomedical research is mainly based on the analysis of distributed and heterogeneous data repositories. Different data integration approaches to offer homogeneous access to these data have been developed during the last years. Each method is better suited for different needs and requirements. The most important factors at the moment of selecting the appropriate approach are

- Level of control over the data sources
- Desired level of adaptability of the software
- Desired performance of the software.

Reviewing those systems developed during the last decade, we get the impression that even though systems adopt any of the existing approaches, their design and development end up being specific for each targeted problem. No generic solution exists due to the difficulty to adapt a single solution for all possible scenarios. Each situation must be analysed individually, and the design has to tackle the specific user needs of the involved scenarios.

7.5 References

[1] Bajic V, Brusic V, Li J, Ng S, Wong L (2003) From informatics to bioinformatics. Proc of the First Asia-Pacific bioinformatics conference on bioinformatics 3-12

[2] Grotkjaer T, Nielsen J (2004) Enhancing yeast transcription analysis through integration of heterogeneous data. Current Genomics. 5(8): 673-686

[3] Gurwitz D, Lunshof JE, Altman R (2006) A call for the creation of personalized medicine database. Nature Reviews, Drug Discovery. 5:23-26

[4] Philippi S, Köhler J (2006) Addressing the problems with life-science databases for traditional uses and systems biology, Nature Reviews Genetics. 7(6):482-488

[5] Galperin M (2008) The molecular biology database collection: 2008 update. Nucleic Acids Res. Vol 36, Database issue pp. 2-4

[6] Jarke M, Jeusfeld M, Quix C, Vassiliadis P (1998) Architecture and quality in data warehouses. In: Pernici B, Thanos C (eds.) Lecture Notes In Computer Science, vol. 1413. Springer-Verlag, 1998; 93-113

[7] Kimball R (1998) Bringing up supermarts. DBMS, 1998; 11, (1): 47-53

[8] Inmon W, Rudin K, Buss C, Sousa R (1999) Data Warehouse Performance. John Wiley & Sons, Inc.

[9] Lenzerini M (2002) Data integration: a theoretical perspective. Proc 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 233-246

[10] Sujansky W (2001) Heterogeneous database integration in biomedicine. J Biomedical Informatics. 34(4):285-298

[11] Wache H, Scholz T, Stieghahn H, König-Ries B (1999) An integration method for the specification of rule-oriented mediators. Proc of the International Symposium on Database Applications in Non-Traditional Environments. p. 109

[12] Ullman J (1997) Information integration using logical views. Proc International Conference on Database Theory. pp. 19–40

[13] Levy A, Rajaraman A, Ordille J (1996) Querying heterogeneous information sources using source descriptions. Proc 22nd International Conference on Very Large Data Bases. pp. 251–262

[14] Huang Y, Ni T, Zhou L, Su S (2003) JXP4BIGI: a generalized, Java XML-based approach for biological information gathering and integration. Bioinformatics. 19(18): 2351-2358.

[15] Do H, Rahm E (2004) Flexible integration of molecular-biological annotation data. Advances in database technology. 2992: 811-822

[16] Shah S, Huang Y, Xu T, Yuen M, Ling J, Ouellette B. Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics 2005; 6: 34

[17] Huang H, Hu Z, Arighi C, Wu C (2007) Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. Front Biosci. 12: 5071-88

[18] Critchlow T, Fidelis K, Ganesh M, Musick R, Slezak T (2000) DataFoundry: information management for scientific data. IEEE Trans Inf Technol Biomed. 4(1): 52-7

[19] Eckman B, Kosky A, Laroco L (2001). Extending traditional query-based integration approaches for functional characterization of post-genomic data. Bioinformatics. 17(7): 587-601

[20] Freier A, Hofestädt R, Lange M, Scholz U, Stephanik A (2002) BioDataServer: a SQL-based service for the online integration of life science data. In Silico Biol. 2(2): 37-57

[21] Höding M, Hofestädt R, Saake G, Scholz U (1998) Schema derivation for WWW information sources and their integration in databases in bioinformatics. Advances in Databases and Information Systems. 1475: 296-304

[22] Chung S, Wong L (1999) Kleisli: a new tool for data integration in biology. Trends in Biotechnology. 17(9):351-355

[23] Wong L (2001) Bioinformatics integration simplified: the Kleisli way. Frontiers in human genetics: diseases and technologies. pp. 79-90

[24] Gupta A, Ludascher B, Martone M (2000) Knowledge-based integration of neuroscience data sources. Scientific and Statistical Database Management. pp. 39-52

[25] Goble C, Stevens R, Ng G, Bechhofer S, Paton N, Baker P, et al. (2001) Transparent access to multiple bioinformatics information sources. IBM Syst J. 40(2): 532-551

[26] Baker P, Goble C, Bechhofer S, Paton N, Stevens R, Brass A (1999) An ontology for bioinformatics applications. Bioinformatics. 15(6): 510–520

[27] Kemp G, Angelopoulos N, Gray P (2000) A schema-based approach to building a bioinformatics database federation. Proc 1st IEEE International Symposium on Bioinformatics and Biomedical Engineering. pp. 13-20

[28] Kemp G, Angelopoulos N, Gray P (2002) Architecture of a mediator for a bioinformatics database federation. IEEE Trans Inf Technol Biomed. 6(2): 116-22

[29] Köhler J, Philippi S, Lange M (2003) SEMEDA: ontology based semantic integration of biological databases. Bioinformatics. 19(18): 2420-7

[30] Haas L, Schwarz P, Kodali P, Kotlar E, Rice J, Swope W (2003) DiscoveryLink: a system for integrated access to life sciences data sources. IBM Systems Journal. 40(2): 489-511

[31] Aldana J, Roldán M, Navas I, Pérez A, Trelles O (2004) Integrating Biological Data Sources and Data Analysis Tools through Mediators. Proc 2004 ACM symposium on applied computing. pp. 127-131

[32] Donelson L, Tarczy-Hornoch P, Mork P, Dolan C, Mitchell J, Barrier M, Mei H (2004) The BioMediator system as a data integration tool to answer diverse biologic queries. Stud Health Technol Inform. 107(Pt 2): 768-72

[33] Wang K, Tarczy-Hornoch P, Shaker R, Mork P, Brinkley J (2005) BioMediator data integration: beyond genomics to neuroscience data. AMIA Annu Symp Proc. pp. 779-83

[34] Caragea D, Pathak J, Bao J, Silvescu A, Andorf C, Dobbs D, Honavar V (2005) Information integration and knowledge acquisition from semantically heterogeneous biological data sources. Data Integration in the Life Sciences. 3615: 175-90

[35] Arredondo T, Seeger M, Dombrovskaia L, Avarias J, Calderón F, Candel D, Muñoz F, Latorre V, Agulló L, Cordova M, Gómez L (2006) Bioinformatics integration framework for metabolic pathway data-mining. Advances in Applied Artificial Intelligence. 4031: 917-26.

[36] Pérez-Rey D, Maojo V, García-Remesal M, Alonso-Calvo R, Billhardt H, Martín-Sánchez F, Sousa A (2006) ONTOFUSION: ontology-based integration of genomic and clinical databases. Computers in Biology and Medicine. 36: 712-30

[37] Kirsten T, Rahm E (2006) BioFuice, mapping-based data integration in bioinformatics. Data Integration in the Life Sciences. 4075: 124-35

[38] Rahm E, Thor A, Aumueller D, Do H, Golovin N, Kirsten T (2005) iFuice - information fusion utilizing instance correspondences and peer mappings. Proc 8th Int Workshop on the Web & Databases (WebDB)

[39] Belleau F, Nolin M, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008; 41:706–716

[40] Martin L, Anguita A, de la Calle G, Garcia-Remesal M, Crespo J, Tsiknakis M, et al. (2007) Semantic data integration in the Europ. ACGT project, AMIA Annu Symp Proc p. 1042

[41] Brochhausen M, Spear A, Cocos C, Weiler G, Martin L, Anguita A, et al. (2010) The ACGT Master Ontology and its applications - towards an ontology-driven cancer research and management system. J Biomedical Informatics 44(1): 8-25

8 System Design and Architecture

Today's large-scale software systems are among the most complex structures ever built by humans, containing millions of lines of code, thousands of database tables, and hundreds of components, all running on dozens of computers. This presents some formidable challenges to software development teams – and if these challenges aren't addressed early, systems are delivered late, over budget, or with an unacceptably poor level of quality.

In order to comprehend a complex computer system, we have to understand what each of its important parts actually do, how they work together, and how they interact with the world around them – in other words, its architecture.

8.1 Software Architecture Definitions

Software architecture is a description of a software system in terms of its major components, their relationships, and the information that passes among them. In essence, architecture is a plan for building systems that meet well-defined requirements and, by extension, systems that possess the characteristics needed to meet those requirements now and in the future.

A fundamental purpose of software architecture is to help manage the complexity of software systems and the modifications that systems inevitably undergo in response to external changes in the business, organizational, and technical environments. There is no single, industry-wide definition of software architecture. The Software Engineering Institute (SEI) web site includes a long list of definitions for the term "software architecture" (http://www.sei.cmu.edu/architecture/definitions.html) with some definitions providing details and context to the abstract definition given above, and expanding on the notions of a system description, requirement specification, and planning. Others are just as abstract but provide a different viewpoint for thinking about architecture.

Perry and Wolf [1] have perhaps the most classic definition, though it's a little sketchy:

Architecture = {elements, form, rationale}

where elements are Processing, Data, or Connecting elements:

- *A viewpoint is a collection of patterns, templates, and conventions for constructing one type of view. It defines the stakeholders whose concerns are reflected in the viewpoint and the guidelines, principles, template models for constructing its views.*
- *A stakeholder in a software architecture is a person, group, or entity with an interest in or concerns about the realization of the architecture.*
- *A concern about an architecture is a requirement, an objective, an intention, or an aspiration a stakeholder has for that architecture.*

A software architecture encompasses the significant decisions about the organization of a system, the structural elements that make up the system, the system composition from those elements, and how the systems are deployed to provide run-time capabilities.

> *A view is a representation of one or more structural aspects of an architecture that illustrates how the architecture addresses one or more concerns held by one or more of its stakeholders.*

There are many different aspects to building software [2], and once again, architectural perspectives provide a mechanism to divide the problem into individual concerns. The most common approach to software perspectives comes from the ''4+1 Views,'' described below.

8.1.1 4+1 Views Model

The “4+1” views model was originally developed in 1987 by Phillipe Kruchten of Rational Software [3,4]. According to this model, each view represents a different set of important, related concepts that can be understood separately and that often have their own sets of expertise. This means that each view can be modelled, i.e., each view can be represented by a distinct set of models, and these models can be assembled to create a complete system.

The logical view primarily supports behavioural requirements: the services to be provided to its end-users. Designers decompose the system into a set of key abstractions, taken mainly from the problem domain. Those are objects or object classes that exploit the principles of abstraction, encapsulation, and inheritance. In addition to aiding functional analysis, decomposition identifies mechanisms and design elements common across the system.

The process view addresses concurrency and distribution, system integrity, and fault tolerance. The process view also specifies which thread of control executes each operation of each class identified in the logical view. The process view can be seen as a set of independently executing logical networks of communicating programs – processes – that are distributed across a set of hardware resources which in turn are connected by a bus or a local area network or a wide area network.

The development view focuses on the organization of the software modules in the software development environment. The units of this view are small chunks of software – program libraries or subsystems – that can be developed by one or more developers. This view supports allocating requirements and work to teams and supports cost evaluation, planning, monitoring of project progress and reasoning about software reuse, portability and security.

The physical view considers system requirements like availability, reliability, performance and scalability. This view maps the various elements identified in the logical, process, and development views (networks, processes, tasks, and objects) onto the processing nodes.

The graphical depiction of an architectural view is called an architectural blueprint. For the various views described above, the blueprints are composed of the UML diagrams:

- Logical View: Class diagrams, sequence diagrams and collaboration diagrams
- Process View: Class diagrams and collaboration diagrams encompassing processes
- Development View: Component diagrams
- Physical View: Deployment diagrams
- Use Case View: Use case diagrams

8.1.2 Rozanski and Woods Viewpoint Set

Rozanski and Woods [5] describe a set of six viewpoints (in the ISO 42010 sense) – extending above’s 4+1 set, to use in documenting software architectures. They are:

8.1.2.1 Functional View

The functional view documents the system’s functional elements, their responsibilities, interfaces, and primary interactions. A functional view is the cornerstone of most architecture documents and is often the first part of the documentation that stakeholders try to read. It drives the shape of other system structures such as the information structure, concurrency structure, deployment structure, and so on. It also has a significant impact on the system’s quality properties, such as its ability to change or to be secured, and its runtime performance.

8.1.2.2 Information View

The information view documents the way that the architecture stores, manipulates, manages, and distributes information. The purpose of virtually any computer system is to manipulate information in some form, and this viewpoint develops a complete but broad view of static data structure and information flow. The objective of this analysis is to answer the important questions around content, structure, ownership, latency, references, and data migration.

8.1.2.3 Concurrency View

The concurrency view describes the concurrency structure of the system and maps functional elements to concurrency units to clearly identify the parts of the system that can execute concurrently and how this is coordinated and controlled. This entails the creation of models that show the process and thread structures that the system will use and the inter-process communication mechanisms used to coordinate their operation.

8.1.2.4 Development View

The development view describes the architecture that supports the software development process. Development views communicate the aspects of the architecture of interest to those stakeholders involved in building, testing, maintaining, and enhancing the system.

8.1.2.5 Deployment View

This view describes the environment into which the system will be deployed, including capturing the dependencies the system has on its runtime environment, i.e. the hardware environment the system needs, the technical environment requirements for each element, and the mapping of the software elements to the runtime environment that will execute them.

8.1.2.6 Operational View

The operational view describes how the system will be operated, administered, and supported when it is running in its production environment. For all but the simplest systems, installing, managing, and operating the system is a significant task that must be considered and planned at design time. The aim of the operational view is to identify system-wide strategies to address and solve the operational concerns of the system's stakeholders.

8.2 Architectural Styles

When many applications share the same structure and the relationships between the parts are very similar, we call it an "architecture style". It is basically a set of principles – a coarse grained pattern that provides an abstract framework for a family of systems. An architectural style improves partitioning and promotes design reuse by reducing the set of possible forms to choose from, and imposing a certain degree of uniformity to the architecture. Garlan and Shaw [6] define an architectural style as

> ..a family of systems in terms of a pattern of structural organization. More specifically, an architectural style determines the vocabulary of components and connectors that can be used in instances of that style, together with a set of constraints on how they can be combined. These can include topological constraints on architectural descriptions (e.g. no cycles). Other constraints – say, having to do with execution semantics – might also be part of the style definition.

In a shorter definition, [7] describe architectural styles as:

> *An architectural style is a specialization of element and relation types, together with a set of constraints on how they can be used.*

Therefore an architectural style defines:

> *A family of systems in terms of a pattern of structural organization, much like "design patterns" for the structure and interconnection within and between software systems.*

A vocabulary of components and connectors with constraints on how they can be combined

The introduction of architectural styles provides several benefits. The most important one is that, since they are a type of "pattern", they provide a common language in a technology-agnostic language. This facilitates a higher level of conversation that is inclusive of patterns and principles without getting into specifics. E.g. by using architecture styles, you can talk about client/server versus *n*-tier. Some architectural styles are shown in the table below:

Architectural style	**Description**
Client/Server	Separates the system into two applications, client program initiates contact with a separate server program (usually on a different machine) for a specific function or purpose. The client exists in the position of the requester for the service provided by the server.
Component-Based Architecture	Decomposes application design into reusable functional or logical components that expose well-defined communication interfaces.
Layered Architecture	Partitions the concerns of the application into layers (stacked groups) so that changes can be made in one layer without affecting the others.
Message Bus	An architecture style that prescribes use of a software system that can receive and send messages using one or more communication channels, so that applications can interact without needing to know specific details about each other.
N-Tier / 3-Tier	3-Tier is a client–server/layered architecture in which the presentation, the application processing, and the data management are logically separate processes with each process being located on a physically separate computer. N-Tier is a generalization where more "tiers" (layers) are introduced.
Object-Oriented	A design paradigm based on division of responsibilities for an application or system into individual reusable and self-sufficient objects, each containing the data and the behaviour relevant to the object.
Service-Oriented Architecture (SOA)	Refers to applications that expose and consume functionality as a service using contracts and messages.
Representational State Transfer (REST)	This is a kind of client-server architectural style where the clients initiate requests to the servers and the servers return appropriate responses. Requests and responses convey the representations of resources where a resource can be anything that may be addressed.

Architectural styles can be organized by their key focus area. The following table lists the major areas of focus and the corresponding architectural styles:

Category	Architectural styles
Communication	Service-Oriented Architecture (SOA), Message Bus
Deployment	Client/Server, N-Tier, 3-Tier
Structure	Component-Based, Object-Oriented, Layered Architecture

It is important to note that the architecture of a software system is almost never limited to a single architectural style but often combines architectural styles that make up the complete system. There are different ways to architectural styles. One way is through hierarchy where a component of a system organized in one architectural style may have an internal structure that is developed in a completely different style. Another way may be when a component uses a mixture of architectural "connectors" to interact with different parts of the system.

8.3 Standards

8.3.1 IEEE 1471

The IEEE 1471 standard "Recommended practice for Architecture Description of Software-Intensive Systems" (http://www.iso-architecture.org/ieee-1471) addresses the activities of the creation, analysis, and sustainment of architectures of software-intensive systems, and the recording of such architectures in terms of architectural descriptions. A conceptual framework for an architectural description is established and the content of an architectural description is defined. Annexes provide the rationale for key concepts and terminology, the relationships to other standards and examples of usage. This recommended practice has been also adopted since 2007 as an ISO standard, ISO/IEC 42010:2007. Figure 1 illustrates the conceptual model of the architectural description, as defined in IEEE 1471.

According to this conceptual model, a system has an architecture and this can be described in an architectural description. Note the distinction between the architecture of a system, which is conceptual, from the description of this architecture, which is concrete. Architectural description (AD) is defined as "a collection of products to document an architecture". The AD can be divided into one or several views. Each view covers one or more stakeholder concerns. View is defined as "a representation of a whole system from the perspective of a related set of concerns". A view is created according to rules and conventions defined in a viewpoint. Viewpoint is defined as "a specification of the conventions for constructing and using a view. A pattern or template from which to develop individual views by establishing the purposes and audience for a view and the techniques for its creation and analysis". An AD selects one or more viewpoints for use and this choice depends on the concerns of the stakeholders that need to be addressed by the architectural description. A view may consist of one or more models and a model may participate in one or more views. Each such model is defined according to the methods established in the corresponding viewpoint definition. The AD aggregates the models, organized into views.

IEEE 1471 defines a set of requirements for conforming ADs that can be summarized as:

- AD identification, version, and overview information
- Identification of the system stakeholders and their concerns
- Specification of each selected viewpoint and the rationale for those selections
- One or more architectural views

- A record of all known inconsistencies among the AD's required constituents
- A rationale for selection of the architecture

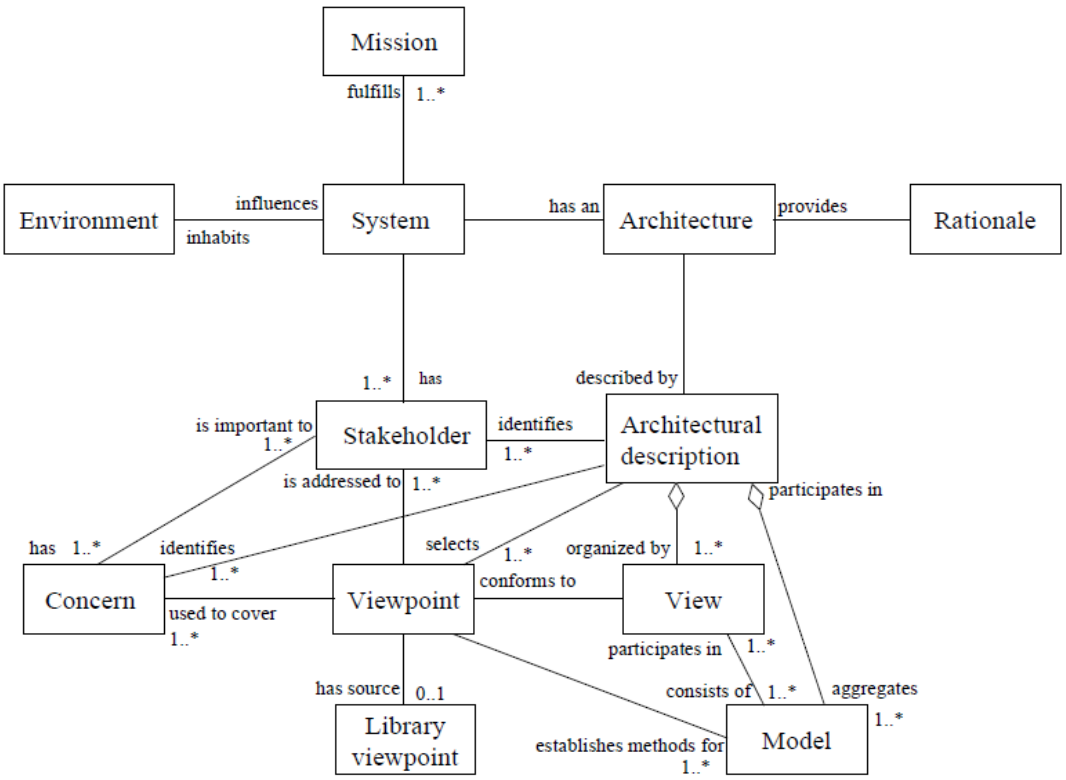


Figure 1: Conceptual model of architectural description from IEEE1471.

8.3.2 The Open Group Architecture Framework (TOGAF)

The Open Group Architecture Framework (TOGAF) is an industry standard architecture framework which provides an approach for designing, planning, implementing and governing of enterprise information architecture. TOGAF has been developed and evolved by the Architecture Forum of The Open Group, which is a consortium of industry companies and organizations. The Open Group is most famous as the certifying body for the UNIX trademark, and its publication of the "Single UNIX Specification" technical standard, which extends the POSIX standards and is the official definition of a UNIX system.

TOGAF, similarly to the IEEE 1471 standard, provides a general framework for developing a broad range of architectures. In essence, it describes how to describe an information system in terms of building blocks, show how the building blocks fit together, provide a set of common vocabulary and refer to a list of recommended standards and compliant products that can be used to implement the building blocks.

TOGAF is a high-level approach to design, typically modelled at four levels: business, application, data and technology and it tries to give an overall starting model to information architects upon which to built. It is based on four pillars, called the architecture domains:

- Business architecture defines the business strategy, governance, organization and key business processes of the organization
- Application architecture provides a blueprint for the individual application systems to be deployed, the interactions between the application systems, and their relationships

to the core business processes of the organization with the frameworks for services to be exposed as business functions for integration.

- Data architecture describes the structure of an organization's logical and physical data assets and the relevant data management resources.
- Technology architecture describes the hardware, software, network and other types of infrastructure needed to support the deployment of the applications.

The Architecture Development Method (ADM) (see Figure 2) is an iterative and cyclic process, described by TOGAF, which is applied to develop an enterprise architecture which will meet the requirements of an organization. In this process, each step checks the Requirements and involves some combination of the above mentioned architecture domains in order to provide a complete information architecture.

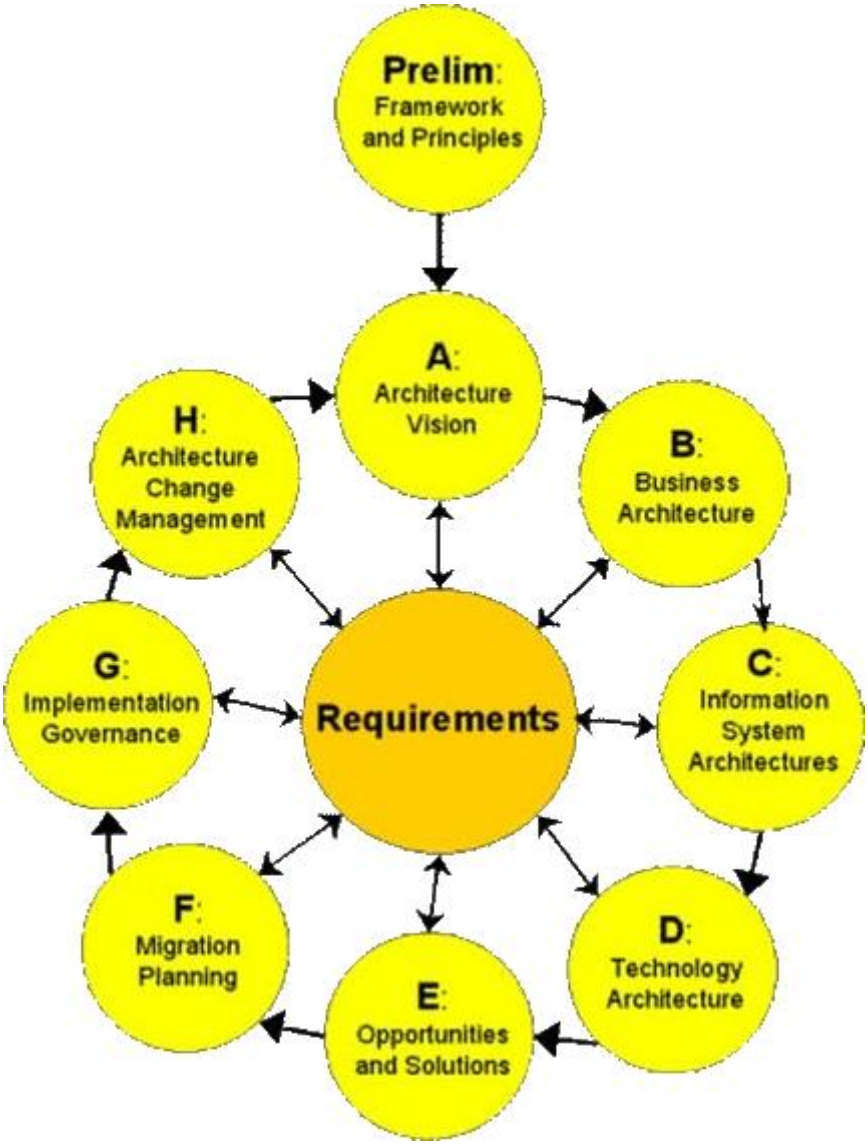


Figure 2: TOGAF Architecture Development Method

8.3.3 Model Driven Architecture (MDA)

Model-Driven Architecture (MDA) is a software design approach for the development of software systems. It provides a set of guidelines for the structuring of specifications, which are expressed as models. MDA was proposed and standardized by the Object Management

Group (OMG) which is a consortium of IT industry companies and organizations, most known for the creation of the CORBA (Common Object Request Broker Architecture) standard.

MDA exploits the emergence of a class of tools, which support model translation and allow meta-model manipulation. Meta-models are models of the formalism used to build models. They define the various kinds of contained model elements and the way they are arranged, related and constrained. The process of developing a model results in creating instances of the model elements defined in the meta-model then "populated" with instance data.

Model transformation is the process of converting a model expressed in one formalism to another model of the same system expressed in a different formalism. This can be achieved by building a meta-model of each of the source and target model representations and then defining a mapping between them. The meta-model of the source model is populated with instance data of the specific source model to be transformed. The mapping rules are applied as a set of operations invoked on the source meta-model, which results in a meta-model of the target model populated with instance data. This populated target meta-model is then used to generate the target model (or possibly the target text in the case of code generation).

In MDA, the definition of a system is done using a platform-independent model (PIM) by utilizing a domain-specific language (DSL). Then, given a platform definition model (PDM) that corresponds to a specific development environment, such as CORBA, .NET, Web services etc., the PIM can be translated to a platform specific model (PSM). This translation is usually performed by automated tools, which execute the corresponding mappings between the various models and the actual implementation mechanisms.

One of the main aims of the MDA is to separate the design from the architecture. As the concepts and the technologies change at different pace, decoupling them allows system developers to choose the best and most fitting in both domains. The design addresses the functional requirements while the architecture provides the infrastructure through which the non-functional requirements are realized. MDA envisages that the platform independent model (PIM)m which represents the conceptual design of a system, will survive changes in realization technologies.

The OMG also provides rough specifications for tools that can be used to develop, compare, verify, transform and otherwise use models and meta-models. The implementations of these specifications come from independent companies or open-source groups such as the Eclipse Foundation, which is developing some modelling tools according the specifications of OMG.

8.4 Modern Architectural Methodologies

In this section we present the two most popular architectural styles for the definition of the functional view of a system's software architecture, namely SOA and REST.

8.4.1 Service Oriented Architecture (SOA)

SOA is an architectural style for building enterprise solutions based on services. More specifically, SOA is concerned with the independent construction of business-aligned services that can be combined into meaningful, higher-level business processes and solutions within the context of the enterprise. Anybody can create a service: that is not the challenge of SOA. The real value of SOA comes when reusable services are combined to create agile, flexible, business processes. Achieving this might be easier to manage if a single organization is creating all of the services, but that is not the case at most large organizations. So, part of the architecture of SOA is responsible for creating the environment necessary to create and use composable services across the enterprise.

SOA is an architectural framework that breaks down software applications and business processes into component services that can be combined and reused with minimal effort. It doesn't simply define an implementation technology but is fundamentally an architectural solution for a specific problem in a given context. SOA uses XML web services as one possible implementation technology [8].

The central objective of a service-oriented approach is to reduce dependencies between "software islands," which basically comprise services and the clients accessing those services. These service-oriented software systems need to balance the following forces:

- Distribution: From a logical perspective the software environments under consideration consist of different software entities running on different network nodes that might need to cooperate via a communication protocol.
- Heterogeneity: The distributed software entities typically reside in heterogeneous environments, so client developers can't control remote services' implementation details. The developers have little or no control of the environments of the remote service consumers/providers.
- Dynamics: The software systems mostly comprise highly dynamic environments, so designers cannot statically predefine many decisions since the decisions must be dynamically configured at runtime. thus SOA environments are usually not static.
- Transparency: Remote-services providers and consumers should be oblivious to the underlying communication infrastructure's implementation details.
- Process-orientation: Services are often fine-grained. The clients compose the services into more coarse-grained building blocks and service composition is necessary for coordinated workflows.

SOA is agnostic in use of technology and in implementation. It can be implemented using by one of the following technologies: XML Web Services, CORBA, Java RMI,.NET Remoting, e-mail, Message-Oriented Middleware (MOM), and even "raw" TCP/IP. In particular we should emphasize that SOA is not Web Services (WS) in the sense that WS is a set of standards not only for defining the interfaces but also for accessing them.

8.4.2 REpresentational State Transfer (REST)

The Representational State Transfer (REST) style is an abstraction of architectural elements within a distributed hypermedia system. It ignores the details of component implementation and protocol syntax to focus on the roles of components, the constraints on their interaction with other components, and their interpretation of significant data elements. It encompasses the fundamental constraints upon components, connectors, and data that define the basis of the Web architecture, and thus the essence of its behavior as a network-based application.

REST-style architectures consist of clients and servers. Clients initiate requests to servers; servers process requests and return appropriate responses. Requests and responses are built around the transfer of representations of resources. A resource can be essentially any coherent and meaningful concept (an abstract or physical "thing") that may be addressed or talked about and be represented as a stream of bytes. A representation of a resource is typically a document that captures the current or intended state of a resource. Each resource has at least one identifier or URI that is used as its name. Examples of such URIs are mail addresses e.g. mailto:bob@example.com, Objects Identifiers (OID) like urn:oid:2.16.180, Life Science identifiers like urn:lsid:uniprot.org:enzymes:3.1.3.16, web addresses like https://www.mybank.com, etc. The client begins sending requests when it is ready to make the transition to a new state. While one or more requests are outstanding, the client is considered to be in transition. The representation of each application state contains links that may be used next time the client chooses to initiate a new state transition.

In HTTP(S)-based RESTful web services, the emphasis is on simple point-to-point communication over HTTP using XML. REST is a hybrid style derived from several of the network-based architectural styles and combined with additional constraints that define a uniform connector interface. REST architectures that use the HTTP application protocol can be summed up as using four verbs (GET, POST, PUT, and DELETE methods from HTTP 1.1) and the nouns, which are the resources available on the network (referenced in the URI). The verbs have the following operational equivalents:

HTTP	CRUD Equivalent	Safe	Idempotent
GET	Read	Yes	Yes
HEAD	Get metadata	Yes	Yes
POST	Create, Update, Delete	No	No
PUT	Create, Update	No	Yes
DELETE	Delete	No	Yes

The convention has been established that the GET and HEAD methods should not have the significance of taking an action other than retrieval, i.e. they don't have side effects. These methods ought to be considered “safe”. Methods can be “idempotent” in that (aside from error or expiration issues) the side-effects of two or more identical requests is the same as for a single request. The methods GET, HEAD, PUT and DELETE share this property.

8.4.2.1 Data Elements

REST data elements focus on a shared understanding of data types with metadata but limit the scope of what is revealed to a standardized interface. REST components communicate by transferring a representation of a resource in a format matching one of an evolving set of standard data types, selected dynamically based on the capabilities or desires of the recipient and the nature of the resource.

Data element	Modern Web examples
Resource	The intended conceptual target of a hypertext reference
Resource identifier	URL, URN
Representation	HTML document, JPEG image
Representation metadata	Media type, last-modified time
Resource metadata	Source link, alternates, vary
Control data	if-modified-since, cache-control

Table 1: REST Data Elements

8.4.2.2 Resources and Resource Identifiers

The key abstraction of information in REST is a resource. Any information that can be named can be one. A resource is a conceptual mapping to a set of entities and not the entity that corresponds to the mapping at any particular time. REST uses resource identifiers to identify the particular resources involved in an interaction between components. REST connectors provide a generic interface to access and manipulate the value set of a resource, regardless of how the membership function is defined or the type of software handling the request. The naming authority that assigned the resource identifier, making it possible to reference the resource, is responsible for maintaining the semantic validity of the mapping over time.

8.4.2.3 Representations

REST components perform actions on a resource by using a representation to capture the current or intended state of that resource and transferring that representation between components. A representation consists of data, metadata describing the data and, on occasion, metadata to describe the metadata (usually for the purpose of verifying message integrity). Metadata is in the form of name-value pairs where the name corresponds to a standard defining the value's structure and semantics. Response messages may include both representation metadata and resource metadata: information about the resource that is not specific to the supplied representation. Control data defines the purpose of a message between components, like the requested action or the meaning of a response. It is also used to parameterize requests and override the default behaviour of some connecting elements.

8.4.2.4 Connectors

REST uses various connector types to encapsulate the activities of accessing resources and transferring resource representations. The connectors present an abstract interface for component communication, enhancing simplicity by providing a clean separation of concerns and hiding the underlying implementation of resources and communication mechanisms.

All REST interactions are stateless. This restriction accomplishes four functions: 1) it removes any need for the connectors to retain application state between requests, thus reducing consumption of physical resources and improving scalability; 2) it allows interactions to be processed in parallel without requiring that the processing mechanism understand the interaction semantics; 3) it allows an intermediary to view and understand a request in isolation, which may be necessary when services are dynamically rearranged; and, 4) it forces all of the information that might factor into the reusability of a cached response to be present in each request.

Connector	Modern Web examples
Client	Libwww, libwww-perl, Apache HttpComponents (java)
Server	Libwww, Apache API, NSAPI, ISAPI, WSGI (Python)
Cache	Browser cache, Web Proxy servers, Akamai cache network
Resolver	Bind (DNS lookup library)
Tunnel	SOCKS, SSL after HTTP CONNECT

Table 2: REST Connectors

8.4.2.5 Components

REST components are typed by their roles in an overall application action. A user agent uses a client connector to initiate a request and becomes the final recipient of the response. An origin server uses a server connector to govern the namespace for a requested resource. It is the definitive source for representations of its resources and must be the final recipient of any request that intends to modify the value of its resources. Intermediary components act as both client and server in order to forward, with possible translation, requests and responses.

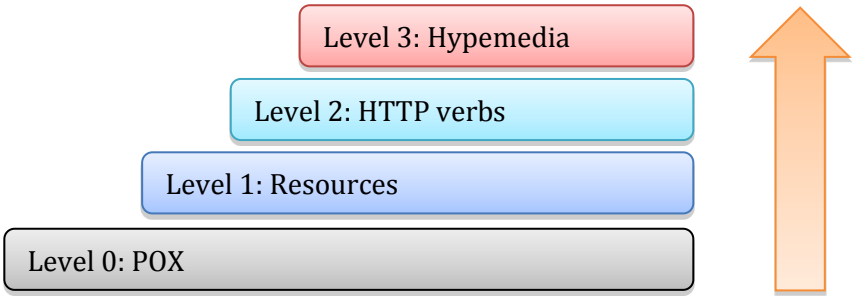
Component	Modern Web examples
Origin server	Apache httpd, Microsoft IIS
Gateway	Squid, CGI, Reverse Proxy
Proxy	CERN Proxy, Netscape Proxy, Gauntlet
User agent	Browsers (e.g. Firefox), “search bots”/“spiders” (e.g. GoogleBot)

Table 3: REST Components

8.4.3 Resource Oriented Architecture

Based on the tenets of REST the Resource Oriented Architecture (ROA) has been proposed by Richardson and Ruby [9] (see Figure 3). This is based on the following four concepts

- Resources are the fundamental abstract or concrete entities that the system needs to manage, interact with, etc.
- Names (URIs), of these resources
- Representations, one or more for each resource
- Links that interconnect the resources



- **Figure 3:** Richardson Maturity Model

They also define four properties that need to be supported:

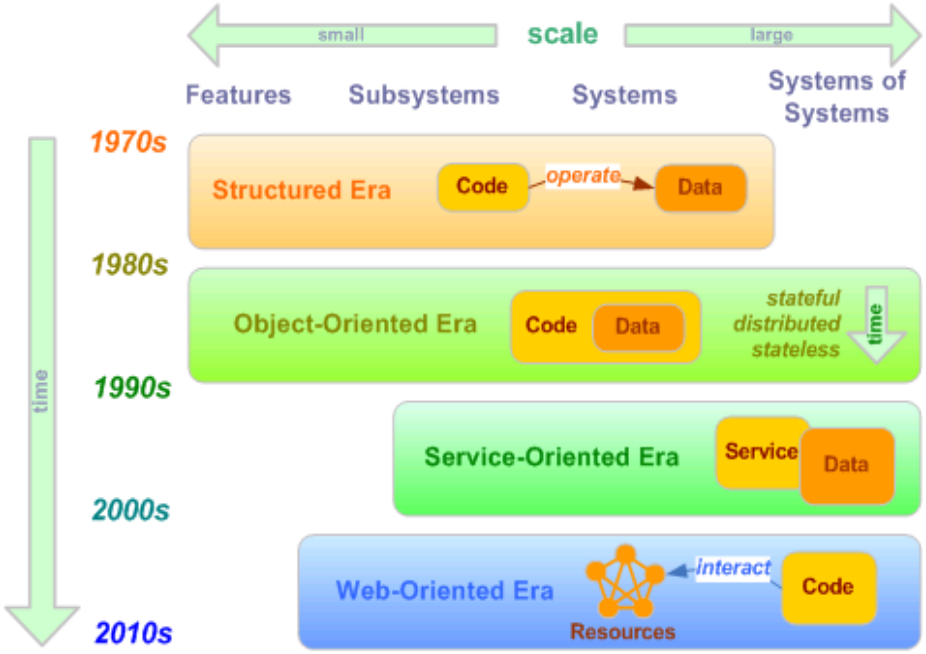
- Addressability, i.e. each resource can be addressed by its name (URI)
- Statelessness, i.e. all interactions are autonomous and stateless as defined by REST
- Connectedness, which means that the system supports the linking between resources and a certain level of navigation from one resource to its connected peers.

- Uniform interface, i.e. all resources supports the same limited set of operations (while the number of resources can be unlimited)

8.5 Technologies

For the successful implementation of the project a large number of technologies need to be seamlessly integrated. The main challenge of the integrated architecture is the interoperability of systems, tools and services made available to the users. A heterogeneous, scalable and flexible environment is needed and the following technologies, which have gained momentum in the recent years (see Figure 4), are being considered for adoption:

- Web Services technologies
- Semantic Web technologies
- Scientific Workflows



- **Figure 4:** The evolution of technologies

8.5.1 Web Services

A Web service is a method of communication between two electronic devices over a network. The W3C describes a Web service as “A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.” (http://www.w3.org/TR/ws-gloss)

Web services are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web, following the generic architecture depicted in Figure 5. This picture shows a middle service repository or registry that stores “offers” of functionality as these are published by service providers, and subsequently performs matching with the corresponding “requests”. After some matching has been performed the corresponding parties (services and their clients) are free to communicate and exchange data. Web services “providers” perform functions that can be anything from simple requests to

complicated business processes. Furthermore, services are meant to target machines i.e. to support machine-to-machine interoperable interactions.

As defined by the W3C Web Service Architecture Working Group "Web services provide a standard means of interoperating between different software applications, running on a variety of platforms and/or frameworks". The main requirements for interoperability are:

- To be vendor, platform, and language agnostic
- It must be simple for programmers to use the protocol and deploy applications, by easy access to client and server side implementations
- Open Internet standards should be used

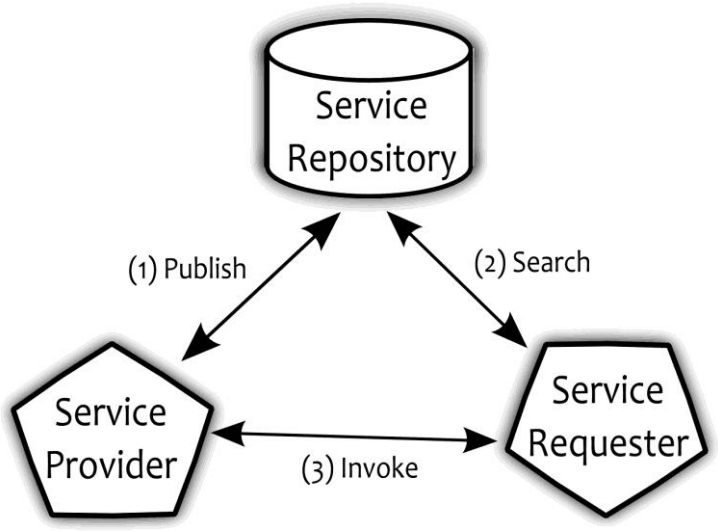


Figure 5: Web Service Architecture

On the technology side Web Services put more emphasis on the following:

- Transport over widely accepted web and internet protocols like HTTP/HTTPS, SMTP
- XML message payloads to provide the extensibility, introspection, and interoperability required in building complex multi party systems
- Platform and programming language independence

The Web itself is built around these very directions: open protocols, text-based (markup, e.g. HTML) message and document content, and abstraction over implementation details. In essence the underlying infrastructure is roughly based on the following technologies:

- SOAP messaging format, which is based on XML, to provide a wrapper format and protocol for data interchange between web services
- Web Service Description Language (WSDL) documents to describe the services' functionality and data exchange

Web services are a set of tools that can be used in a number of ways. The three most common styles of use are RPC, SOA and REST. Technologies implementing RPC includes PRC-XML, Poor-Old-XML (POX) messages, and SOAP-based services. RPC Traditional web services (WS-*) standards are more attached to the activity oriented service architecture, and are rather SOAP-based. The principal components of the so-called "Big" Web Services are SOAP (http://www.w3.org/TR/soap), WSDL (http://www.w3.org/TR/wsdl) and UDDI (http://uddi.xml.org).

W3C describes the set of interrelated technologies that can be utilized to construct and consume Web services, as illustrated in Figure 6.

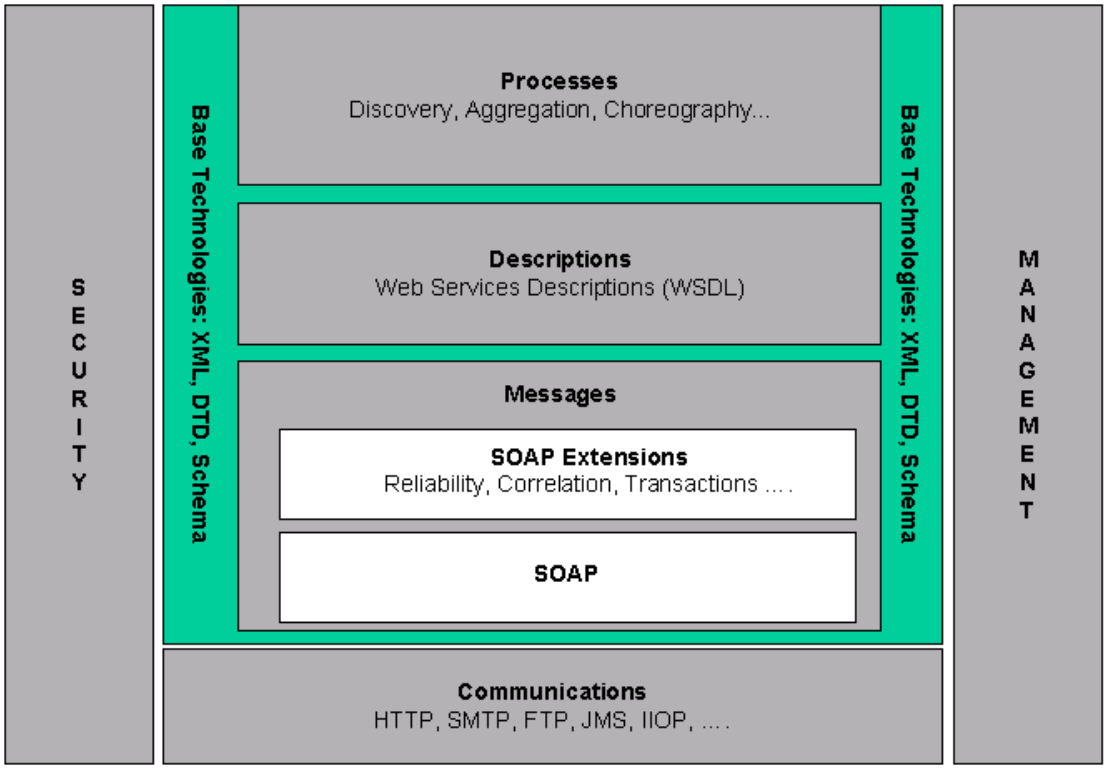


Figure 6: Web Services Architecture Stack (http://www.w3.org/TR/ws-arch)

- XML provides a standard, flexible, and extensible data format that reduces the burden of deploying the technologies utilized in Web services. Key concepts are XML core syntax, XML Infoset, XML Schema, and XML Namespaces.
- SOAP provides a standard, extensible, composable framework for packaging and exchanging XML messages. It also provides a mechanism for referencing capabilities (a named feature or piece of functionality that is declared or requested by an agent).
    - SOAP 1.2 Part 1 defines an XML-based messaging framework, a processing model, and an extensibility model.
    - SOAP 1.2 Part 2 defines encoding rules for expressing instances of application-defined data types, conventions for representing remote procedure calls and responses, and rules for using SOAP with HTTP.
- WSDL defines a language for describing Web services. WSDL describes the messages that are exchanged between requester and provider agents. The messages are described abstractly and bound to a concrete network protocol and message format. Web service definitions can be mapped to any implementation language, platform, object model, or messaging system.

8.5.1.1 Web Service Styles of Use

Web services are a set of tools that can be used in a number of ways. The three most common styles of use are RPC, SOA and REST (http://en.wikipedia.org/wiki/Web_service):

- Remote procedure calls (RPC): RPC Web services present a distributed function (or method) call interface that is familiar to many developers. Typically, the basic unit of

RPC Web services is the WSDL operation. Other approaches with nearly the same functionality as RPC are Object Management Group's (OMG) Common Object Request Broker Architecture (CORBA), Microsoft's Distributed Component Object Model (DCOM) or Sun Microsystems's Java/Remote Method Invocation (RMI).

- Service-oriented architecture (SOA): Web services can be implemented according to SOA concepts, where the communication basic unit is a message rather than an operation. Unlike RPC Web services, loose coupling is more likely since the focus is on the "contract" WSDL provides rather than the underlying implementation details.
- Representational state transfer (REST): It attempts to describe architectures using HTTP or similar protocols by constraining the interface to a set of well-known, standard operations (like GET, POST, PUT, DELETE for HTTP). The focus lies on interacting with stateful resources rather than messages or operations. A REST-based architecture can use WSDL to describe SOAP messaging over HTTP, can be implemented as an abstraction purely on top of SOAP or without using SOAP at all.

8.5.1.2 SOAP

SOAP, formerly defined as The Simple Object Access Protocol, is the successor of XML-RPC and a W3C (http://www.w3.org) recommendation. It provides a framework for a RPC middleware, widely used for activity oriented service architecture. It consists of several layers of specifications for message format, message exchange pattern, underlying protocol bindings, message processing models, and protocol extensibility. In general, it uses HTTP or Simple Mail Transfer Protocol (SMTP) as transport protocols and XML as message format. Documents can also being transferred using Attachments. A client invokes an activity by sending SOAP messages to a SOAP Router, or Dispatcher, that can be a Servlet or a Common Gateway Interface (CGI) script. The dispatcher interprets the message and sends a call to the service implementing the logic. The dispatcher further wraps the result and sends it back to the client in a standardized format. Both, client and server need a SOAP interpreter or engine, which must marshal and unmarshal the message context, mostly using XML. This mechanism causes the principal weakness of SOAP: performance inefficiency.

SOAP Version 1.2 is a lightweight protocol intended for exchanging structured information in a decentralized, distributed environment. "Part 1: Messaging Framework" defines, using XML technologies, an extensible messaging framework containing a message construct that can be exchanged over a variety of underlying protocols. SOAP is fundamentally a stateless, one-way message exchange paradigm, but applications can create more complex interaction patterns (e.g. request/response, request/multiple responses, etc.) by combining such one-way exchanges with features provided by an underlying protocol and/or application-specific information. SOAP is silent on the semantics of any application-specific data it conveys, as it is on issues such as the routing of SOAP messages, reliable data transfer, firewall traversal, etc. However, SOAP provides the framework by which application-specific information may be conveyed in an extensible manner.

8.5.1.2.1 SOAP Messages

A SOAP message is specified as an XML infoset whose comment, element, attribute, namespace and character information items are able to be serialized as XML 1.0.

- SOAP Envelope: This element information item has a local name (Envelope), a namespace name (http://www.w3.org/2003/05/soap-envelope), zero or more namespace-qualified attribute information items among its attributes property, one or two element information items in its children property in order of an optional Header element information item, a mandatory Body element information item (see Figure 7).
- SOAP Header: The SOAP Header element information item provides a mechanism for extending a SOAP message in a decentralized and modular way. This element is

optional and is an extension mechanism to provide a way to pass information in SOAP messages that is not application payload. Such “control” information includes, e.g. passing directives or contextual information related to the message processing. This allows a SOAP message to be extended in an application-specific manner.

- SOAP Body: A SOAP body provides a mechanism for transmitting information to an ultimate SOAP receiver. The SOAP body is the mandatory element within the SOAP env:Envelope, which implies that this is where the main end-to-end information conveyed in a SOAP message must be carried.
- SOAP Fault: A SOAP fault is used to carry error information within a SOAP message. To be recognized as carrying SOAP error information, a SOAP message MUST contain a single SOAP Fault element information item as the only child element information item of the SOAP Body.

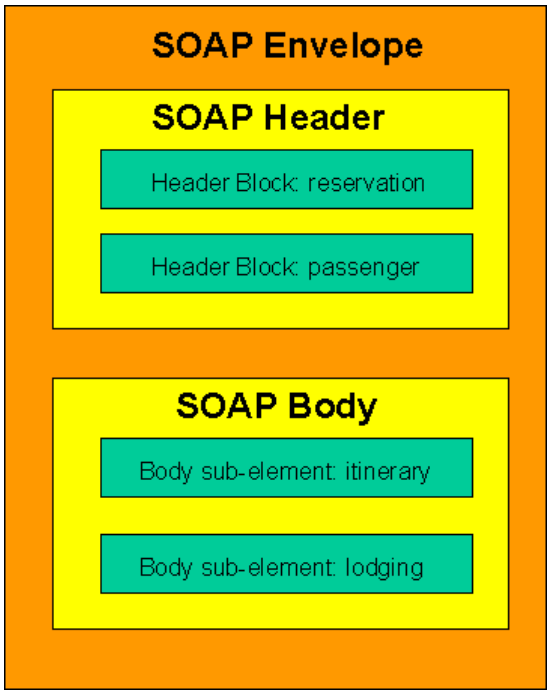


Figure 7: SOAP Message Structure from http://www.w3.org/TR/2007/REC-soap12-part0

8.5.1.3 WSDL

WSDL is an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint. Related concrete endpoints are combined into abstract endpoints (services). WSDL is extensible to allow description of endpoints and their messages regardless of what message formats or network protocols are used to communicate, however, the only bindings described in this document describe how to use WSDL in conjunction with SOAP 1.1, HTTP GET/POST, and MIME.

WSDL describes what the service can do, where it resides, and how to invoke it. A service providing a WSDL description can be “discovered” by a service broker. WSDL plays an important role in the overall Web Services architecture since it describes the complete contract for application communication. It provides a simple way for service providers to describe the basic request format to their systems regardless of the underlying protocol (such as SOAP or XML) or encoding (such as Multipurpose Internet Messaging Extensions).

WSDL is key part of the effort of the Universal Description, Discovery and Integration (UDDI) initiative to provide directories and descriptions of on-line services for electronic business.

8.5.2 Semantic Web

The Semantic Web is a "web of data" that facilitates machines to understand the semantics, or meaning, of information on the World Wide Web. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users.

The term “Semantic Web” is often used more specifically to refer to the formats and technologies that enable it. These technologies include the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL) (http://www.w3.org/TR/owl-features), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

8.5.2.1 RDF

The Resource Description Framework (http://www.w3.org/RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources using a variety of syntax formats. RDF is not only used to express individuals within semantics but also as a data exchange format.

RDF is based on XML and therefore has the same two levels of correctness. Whereas the main structuring element of XML is the hierarchy the main structuring element of RDF is relation: every information is decomposed into Subject-Predicate-Object triples. Subjects and objects are resources (or rather references to those, specified by URIs), which may be concepts, all over the Web.

RDF provides a framework for describing interchanging metadata. It is based on Web technologies (URI and XML). It is built on the following rules:

- Resource is anything that can have URIs (Web pages, XML document elements, etc.)
- PropertyType: Named resources that can be used as a property, e.g. Author or Title
- Property: Combination of a Resource, a PropertyType, and a value

RDF provides a model for metadata and syntax so that independent parties can exchange it and use it. The semantic web is based on RDF syntax. The simplicity and flexibility of the triple combined with using URIs for globally unique names, makes RDF unique and very powerful. It is a specification that fills a particular niche for decentralized, distributed knowledge providing a framework to enable reasoning in computer applications.

Like for XML there are different levels for the expression of Schemata for RDF: RDFS, OWL-Lite, OWL-DL and OWL-Full. They mainly differ in the power of expression of relations. The most common language is OWL-DL, giving a high power of expression and avoiding the technical complications of OWL-Full, which cannot be processed by reasoners. As those definitions of relations can not only be regarded as schemata for RDF but can be regarded as representations of knowledge and description rules, OWL documents are called ontologies. According to this naming the definitions are called concepts (written in OWL), and the realizations of those definitions are called individuals (written in RDF).

In addition to the checks if an RDF document is well-formed and valid according to an RDFs or OWL, OWL documents can be checked if they are logically consistent. The user also can define and look for derived (inferred) relations, based on the defined (asserted) ones.

8.5.2.2 SPARQL

SPARQL (http://www.w3.org/TR/rdf-sparql-query) Protocol and RDF Query Language is an RDF query language. It was standardized by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is considered a key semantic web technology. On 15 January 2008, SPARQL became an official W3C Recommendation. SPARQL is a general term for both a protocol and a query language.

SPARQL is a syntactically-SQL-like language for querying RDF graphs via pattern matching. The language's features include basic conjunctive patterns, value filters, optional patterns, and pattern disjunction. The SPARQL protocol is a method for remote invocation of SPARQL queries. It specifies a simple interface that can be supported via HTTP or SOAP that a client can use to issue SPARQL queries against some endpoint. Both the SPARQL query language and its protocol are products of the W3C's RDF Data Access Working Group.

8.5.3 Integration Technologies

Integration concerns the design of components that are easy to use as part of a larger suite of components. The goal of the integration process is to make disparate and heterogeneous applications work together so as to produce a unified set of functionality, possibly by complementing each other. In general this is a huge task due to the heterogeneity of the software and hardware platforms, diversity of architectural styles and paradigms, the security concerns, the geographic dispersion of the contributing software entities, etc. In some cases there are also non-technical impediments to the integration process, like crossing enterprise boundaries and rigid organizational policies. In spite of these problems, application/system integration is unavoidable in cases where the building of a single standalone application is difficult or even impossible because of the complexity of the application domain. Integration is also a viable solution for taking advantage of the available infrastructure, either existing (“legacy”) systems or deployed applications and computer resources, and increasing overall system capacity, performance, scalability, user functionality, and customer reach.

8.5.3.1 Criteria for an Optimal Integration Process

The following concerns (see [10]) should be when trying to build a good integration solution:

- Application coupling. The integrated applications should minimize their dependencies on each other so that each can evolve without causing problems to the others. In general tightly coupled applications make numerous assumptions about how the other applications work; when the applications change and break those assumptions, the integration between them breaks. Therefore, the interfaces for integrating applications should be specific enough to implement useful functionality but general enough to allow the implementation to change as needed.
- Intrusiveness. The integration process should not impose too many changes to the constituent applications. This non intrusiveness, to the degree that is possible, is necessary for reducing the integration costs and also ensuring that the integrated system maintains the virtues of the participating components. Nevertheless it can be the case that major changes are needed in order to achieve good integration.
- Technology selection. The choice of the technology platform and the relevant tools is also important. There are many offerings and usually the choice is made based on

the familiarity and the experience of the developers, the inherent costs (e.g. licenses), the applicability of open-source products, etc.

- Data format. Integrated applications must agree on the format of the data they exchange. Changing existing applications to use a unified data format may be difficult or impossible. Alternatively, an intermediate translator can unify applications that insist on different data formats. A related issue is data format evolution and extensibility, how the format can change over time and how this affects applications.
- Remote Communication. It is typical in an integrated environment to have different applications call each other. This communication although similar in principle with the local function call available in the majority of the programming languages is in fact quite different because of the intervention of the network. The common fallacies of distributed computing such as that the network latency is zero, the communication is secure and reliable, etc. (http://web.archive.org/web/20030208015752/http://java.sun.com/people/jag/Fallacies.html) have to be considered and avoided. Generally speaking the integrated applications should make as few assumptions about their environment as possible. The adoption of asynchronous communication and the preparation for and handling of the communication errors (e.g. unreachable network/computer/application) are good advices to follow.

These criteria are, of course, too general and in every integration scenario they must be further elaborated and analysed. Yet, they are indicative of the integration process' complexity and the effort necessary to build a system from different, heterogeneous parts.

8.5.3.2 Integration Styles

Over the last twenty years a number of integration approaches have been proposed and studied which can be roughly categorized as follows:

- Use of some shared area as the means for the communication and synchronization between the integrated applications. Examples of this setting are a common file system, a shared, possibly distributed, database, Tuple Spaces (http://c2.com/cgi/wiki?TupleSpace), etc.
- Have each application expose some kind of programmatic interface with procedures or methods that can be invoked remotely. This kind of integration is traditionally referred as Remote Procedure Call (RPC) or Distributed Objects Integration. Examples of this style are the OMG CORBA and the Web Services architectures.
- Use asynchronous message passing through some common messaging component (a "queue"). This way of interaction involves what is generally called Message Oriented Middleware (MOM) and examples of such middleware components are the Java Message Service (JMS, http://java.sun.com/products/jms) and the Advanced Message Queuing Protocol (AMQP, http://www.amqp.org). Additionally, there are many commercial products available, e.g. IBM WebSphere MQ.

There are advantages and disadvantages to all of the above approaches in terms of ease of integration, standardization, performance and scalability, etc. Also there are cases where different approaches are combined and the resulting integrated environment has features that do not allow its classification into a single integration paradigm.

Integrating applications has continued to become more common, and the growing availability of tools and standards (such as the Web services standards) and service-oriented architecture appear to hold a promise of easier integration. A particular technological framework promoting the composability of services has emerged in the recent years using workflows. Workflows, especially in the e-Science related domains, allow scientists to harness information technologies to accelerate scientific discovery.

8.5.4 Workflow Management Coalition (WfMC)

The Workflow Management Coalition (WfMC, http://www.wfmc.org) was funded to focus on workflow management and the interoperability of workflow management systems. WfMC developed the Workflow Reference Model where a workflow is defined as "the automation of a business process, in whole or part" and a Workflow Management System as a system that "defines, manages and executes workflows through the execution of software whose order of execution is driven by a computer representation of the workflow logic". The representation of the workflow logic (or process description) describes the tasks and activities to be executed and the order of their execution.

8.5.5 Workflows – Business Process Execution Language (BPEL)

A workflow is a pattern of activity, modelling real work for further assessment under a chosen aspect, often depicted as a sequence of operations. In computer programming the term workflow is used to capture and develop human to machine interaction, in order to provide end-users an easier way to orchestrate or describe complex processing of data visually.

A scientific workflow is the process of combining data and processes into a configurable, structured set of steps that implement semi-automated computational solutions of a scientific problem. Scientific workflow systems support in silico experiments, performing large-scale data analysis, integrating different software tools from diverse domains, and regularly provide visual programming interfaces for the modelling and uses grid technology for execution.

WS-BPEL (http://www.oasis-open.org/committees/wsbpel) is a language for specifying business process behaviour based on Web Services. It defines an interoperable integration model that should facilitate the expansion of automated process integration in both the intra-corporate and the business-to-business spaces. The processes described in WS-BPEL can be one of two kinds: Executable and Abstract processes. Executable business processes model actual behaviour of a participant in a business interaction. Abstract business processes are partially specified processes that are not intended to be executed and they may be used to hide some of the required concrete operational details.

8.5.6 Data

Data standards are a critical component in order to improve global public health. Inefficiencies in the collection, processing and analysis of patient and health-related information drive up the cost of drug development for life sciences companies and negatively affect the cost and quality of health care delivery for patients and consumers.

8.5.6.1 Clinical Data Interchange Standards Consortium (CDISC)

CDISC (http://www.cdisc.org) is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform-independent and freely available. Core principles of CDISC are:

- Lead the development of standards that improve efficiency while supporting the scientific nature of clinical research.
- Recognize the ultimate goal of creating regulatory submissions that allow for flexibility in scientific content and are easily interpreted, understood and navigated by regulatory reviewers.

- Acknowledge that the data content, structure and quality of the standard data models are of paramount importance, independent of implementation strategy and platform.
- Maintain a global, multidisciplinary, cross-functional composition for CDISC and its working groups.
- Work with other professional groups to encourage that there is maximum sharing of information and minimum duplication of efforts.
- Provide educational programs on CDISC standards, models, values and benefits.
- Accomplish the CDISC goals and mission without promoting any individual vendor or organization.

CDISC develops and supports controlled terminology in several areas such as:

- Study Data Tabulation Model (SDTM) is an international standard for clinical research data, and is approved by the FDA as a standard electronic submission format.
- Clinical Data Acquisition Standards Harmonization (CDASH) project, which develops clinical research study content standards in collaboration with sixteen partner organizations including NCI.
- Analysis Data Model (ADaM) project, which supports efficient generation, replication, review and submission of analysis results from clinical trial data.
- Standard for the Exchange of Nonclinical Data (SEND) project, which guides the organization, structure and format of standard nonclinical tabulation data sets for interchange between organizations such as sponsors and CROs and for submission to a regulatory authority such as the FDA. NCI EVS maintains and distributes SEND controlled terminology as part of NCI Thesaurus.

8.6 Related Projects and Initiatives

A large number of related projects have produced results that may prove useful in the context of p-Medicine. Of specific importance and relevance are specific data sharing and high performance computing related projects in the domain of Medical Informatics.

The description, objectives, architecture and services developed in the caBIG, ACGT, LifeWatch and myGrid projects, are presented below (by using published scientific papers of these projects and by using the available information on their respective websites).

8.6.1 Cancer Biomedical Informatics Grid (caBIG)

caBIG (https://cabig.nci.nih.gov) initiative was launched by the National Cancer Institute, aiming to create a virtual network of interconnected data, individuals and organizations that collaborate in order to redefine the way that cancer research is conducted. Several tools have been developed under this initiative that assist in collecting, analyzing, integrating and disseminating data information that is related with cancer care and research. Objective of these tools is to promote data sharing in a syntactically interoperable manner.

caGrid [11] is the underlying service-oriented infrastructure that supports caBIG. Driven primarily by scientific use cases from the cancer research community, it provides the core infrastructure necessary to compose the Grid of caBIG.

The core caGrid services include the security services (Dorian, Grid Trust Service (GTS), and Grid Grouper), metadata services (Index Service, Global Model Exchange (GME), Enterprise Vocabulary Services (EVS), and cancer Data Standards Repository (caDSR)), and high-level services such as the Federated Query Processing service (FQP) and the

Workflow services. In addition, data and analytical services (e.g. caArray and caBIO services in the figure), provided by research groups, institutions, individual researchers, can be discovered and securely accessed through the caGrid core services and protocols.

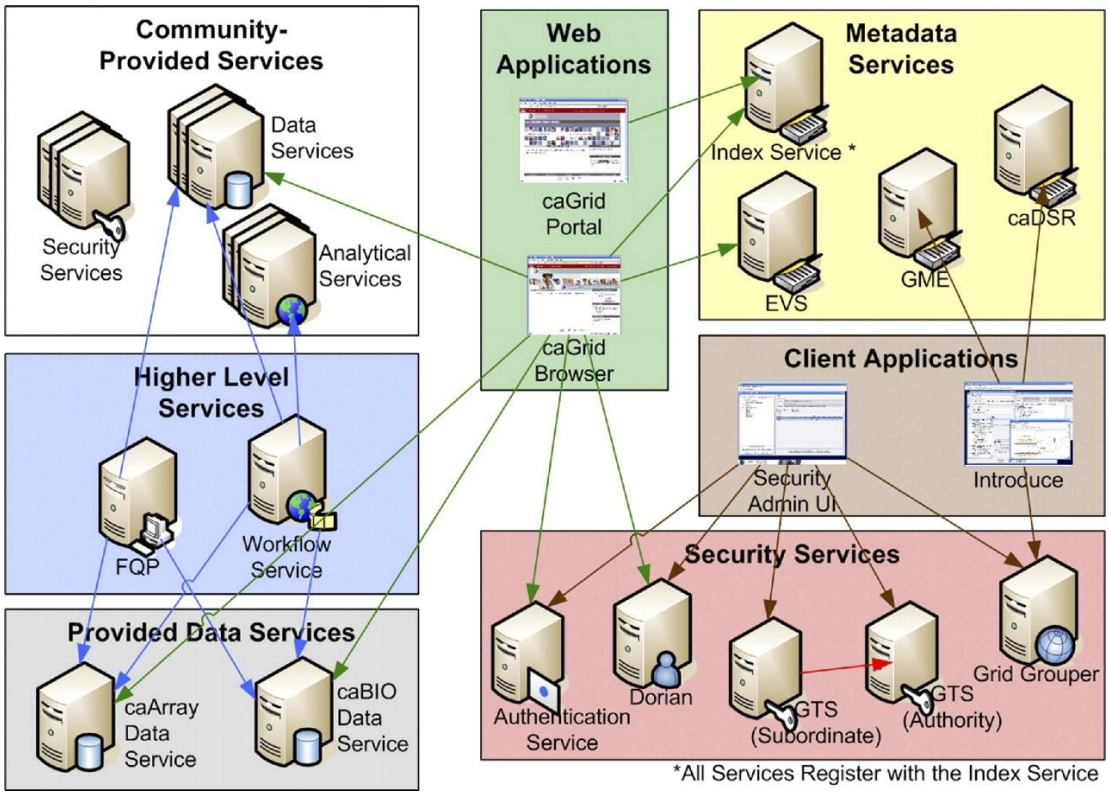


Figure 8: The caGrid infrastructure

8.6.1.1 Standards

A main principle of caBIG is to use open standards: caGrid is built as a service-oriented architecture based on standard Web Services specifications, particularly those standards defined by the Organization for the Advancement of Structured Information Standards (OASIS) (http://www.oasis-open.org) to represent service state (WS-Notification, Web Services Resource Framework (WSRF)) and security (WS-Security, Security Assertion Markup Language (SAML), Extensible Access Control Markup Language (XACML)). It aims to be programming language and toolkit agnostic. Each data and analytical resource in caGrid is implemented as a WSRF 1.2-compliant web service interacting with other resources/clients using standard protocols. The infrastructure implementation uses several grid systems, like the Globus Toolkit or Mobius, and NCI-tools, e.g. caCORE's infrastructure.

8.6.1.2 Data

The data services framework is a key component of the caGrid architecture. Each caGrid data service must conform to the standard query interface the framework requires and must expose standardized metadata describing its local domain model. The basic query language of caGrid, which must be supported by any caGrid data service, is called caGrid Query Language (CQL). It is based on the notion of objects and object hierarchies allowing the user to express queries searching for objects based on object classes and concept definitions and to specify criteria on the object properties and associations. The CQL schema includes:

- Target element to describe the data type queried and that the query should return
- Attribute element which defines a predicate or restriction for an attribute of an object
- Association element which describes restrictions on an object via criteria over another associated object (specifically, it defines a relationship down the object model tree)
- Group element which defines logical joins of two or more conditions that operate against the object to which they are attached

8.6.1.3 Services

caGrid is intended to be used by service developers with different levels of experience with grid computing, grid technologies, and implementation of grid services. The Introduce toolkit has been developed as an extensible grid service authoring tool with a graphical workbench for developers to easily develop and deploy strongly-typed services. Features:

- Reduces development time and knowledge required to implement/deploy services by abstracting WSRF specification details and integration with existing low-level tools
- Provides a graphical user interface, high-level usage-oriented functions and manages all service-specific files/directories required for correct compilation/deployment
- Generates appropriate, object-oriented client APIs that can be used by client applications to interact with the service
- Extensible via plug-ins, which allow for extensions to support common service types in an application domain and to implement mechanisms for customized discovery of common data types in creating strongly-typed services
- Integrates with the GAARDS infrastructure enabling secure services development and deployment

8.6.1.4 Metadata

Each caGrid service is required to describe itself using standard service metadata. When a service is deployed in caGrid, its service metadata is registered with an indexing registry service, called the Index Service. The common data elements are managed in the cancer Data Standards Repository (caDSR). The definitions of these data elements draw from vocabulary registered in the Enterprise Vocabulary Services (EVS). The concepts of data elements and the relationships among them are semantically annotated. The XML schemas corresponding to common data elements and object classes are registered in the Mobius Global Model Exchange (GME) service. In summary caGrid supports four metadata services:

- Cancer Data Standards Repository (caDSR) which
    - registers data models as Common Data Elements (CDEs) which are semantically harmonized and then centrally stored and managed the caDSR
    - provides model discovery/traversal, standard metadata generation capabilities
- Enterprise Vocabulary Services (EVS) provides
    - a set of services and resources addressing the need for controlled vocabulary
    - query access to the data semantics and EVS-managed controlled vocabula
- Global Model Exchange (GME)
    - is a DNS-like data definition registry and exchange service that is responsible for storing and linking structural data models in the form of XML schema

- provides access to the authoritative structural representation of data types on the grid

- Globus Information Services: Index Service provides
  - a generic framework for aggregation of service metadata, a registry of running grid services, and a dynamic data-generating and indexing node, suitable for use in a hierarchy or federation of services
  - Yellow and white pages for the grid

8.6.1.5 Security

The Grid Authentication and Authorization with Reliably Distributed Services (GAARDS) infrastructure of caGrid provides services and tools for the administration and enforcement of security policies in a multi-institutional environment. It consists of three main components: Dorian for provisioning and federation of grid user identities and credentials; Grid Trust Service (GTS) for maintaining a federated trust fabric of all the trusted credential providers in the Grid; and Grid Grouper for group-based authorization support. Its main features are:

- Grid User and Host Account Management
- Integration point between external security domains and the grid
- Allows accounts managed in external domains to be federated/managed in the grid
- Allows users to use their existing (grid-external) credentials to authenticate to the grid

Grid Trust Service (GTS) main features are:

- Creation and Management of a federated trust fabric
- Supports applications and services in deciding whether or not signers of digital credentials/user attributes can be trusted
- Supports the provisioning of trusted certificate authorities and corresponding CRLS

Grid Grouper's main features are:

- Group management service for the grid
- Provides a group-based authorization solution for the grid
- Enforce authorization policy based on membership to groups

8.6.2 Advancing Clinico-Genomic Trials on Cancer (ACGT)

ACGT (http://eu-acgt.org) focuses on the domain of Cancer research, and its ultimate objective is the design, development and validation of an integrated grid enabled technological platform in support of post-genomic, multi-centric Clinical Trials on Cancer. The driving motivation behind the project is the committed belief that the breadth and depth of information already available in the research community at large, present an enormous opportunity for improving the ability to reduce mortality from cancer, improve therapies and meet the demanding individualization of care needs.

Figure 9 below presents that general design view as a starting point for development of the ACGT infrastructure [12]. There are five horizontal layers presented on the picture. The lower ones are located closer to the physical resources, the top ones closer to the end-users. The mechanism used for providing distributed access to resources is grid technology. The layers on top are responsible to provide specific solutions for bioinformaticians and clinicians.

There are two vertical layers. The first one is logging infrastructure and is used by the services regardless their location within architecture. It is very important to have ability to track the activity that in many cases involves different services from different layers.

The other vertical layer is security that constitutes common infrastructure for all components in the infrastructure. It is very important to keep consistent security policies throughout the infrastructure and also to be able to dynamically manipulate the policies for the complete architecture in the context of virtual organisation management.

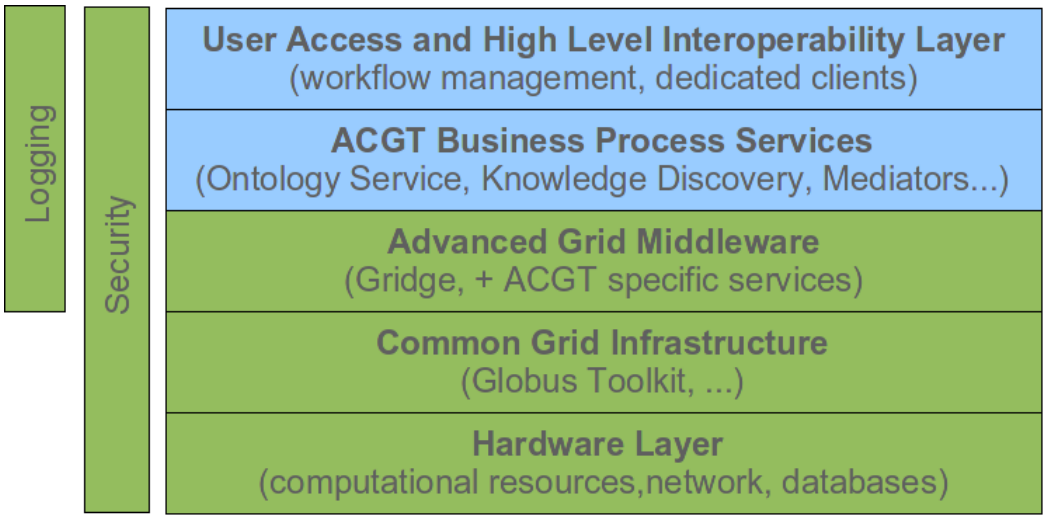


Figure 9: ACGT layered architecture.

8.6.2.1 Standards

The adoption of well-defined standards is the preferred way to guarantee the interoperability of the ACGT components. The basic technologies used to implement Services are:

- XML, as a common serialization format
- SOAP, as the XML-based messaging protocol
- Web Service Description Language (WSDL), as the service description language
- Security: WS-Security, WS-SecureConversation, WS-SecurityPolicy), WS-Trust
- Messaging and Addressing: WS-Addressing, WS-Notification, WS-Eventing
- Metadata: WS-Policy, WS-MetadataExchange, Semantic Annotations for WSDL

ACGT security relies on a commercial-grade PKI implementation (concerning the followed security practices) following the X.509 standards and is composed of several interdependent modules. The most important standards for the ACGT infrastructure implementation are:

- Web Service Description Language is an XML-based language to describe Web services interfaces. WSDL is the most important standard used for Web services implementation. In ACGT all services implemented are described using WSDL.
- Grid Security Infrastructure is a specification for secret, tamper-proof, delegable communication between software in a grid computing environment. Secure, authenticable communication is enabled using asymmetric encryption. In ACGT project GSI was introduced as a common security infrastructure not only for the services of grid layer, but for all services in ACGT environment.
- Job Submission Description Language is an extensible XML specification from the Global Grid Forum for the description of simple tasks to non-interactive computer execution systems. JSDL is used by resource management system for Grid (GRMS) that is used for submission of computational jobs to the grid.

- SPARQL is a query language. It is supported by the data access services that wrap the various data sources. The semantic mediator uses SPARQL as well, not only in constructing the queries that are sent to the data access service, but also for receiving queries (expressed in the ACGT Master Ontology).
- RDFS is a knowledge representation language whose final W3C recommendation was released in February 2004. It is designed to describe the schema of an RDF repository. RDFS is used to describe the schemas of the databases to be integrated in the Semantic Mediator. In some way, it is also utilized to describe the ACGT Master Ontology – in fact, OWL-DL is the language in which the MO is written, however this is only an extension of the RDFS language, as explained before.
- WS-BPEL is the technology used to support the high-level integration of the ACGT services and tools into complex scientific workflows for the implementation, testing and validation of user scenarios. The workflows defined in WS-BPEL are deployed as executable processes and they can be subsequently used as atomic services to construct even more complex and higher-level scenarios.

8.6.2.2 Data

Data storage, management and access in the Gridge environment are supported by the Gridge Data Management Suite (DMS). This suite, composed of several specialized components, allows building a distributed system of services capable of delivering mechanisms for seamless management of large amounts of data. It is based on the autonomic agents pattern using the accessible network infrastructure for communication. From the point of view of external applications, DMS is a virtual file system keeping the data organized in a tree-like structure. The main units of this structure are meta-directories, which allow creating a hierarchy over other objects and metafiles. Metafiles represent a logical view of data regardless of their physical storage location.

The Data Management System consists of three logical layers: the Data Broker, which serves as the access interface to the DMS system and implements the brokering of storage resources, the Metadata Repository that keeps information about the data managed by the system, and the Data Container, which is responsible for the physical storage of data. In addition, DMS contains modules which extend its functionality to fulfil common enterprise requirements. These include the fully functional web-based administrator interface and a Proxy to external scientific databases with a SOAP interface.

8.6.2.3 Services

In the ACGT environment there can be a lot of different services. The SOA paradigm defines architecture as a set of loosely coupled software services that support the requirements of business processes and software users. That kind of definition provides just a rough idea of the service without any notion what the functionality of the particular services should be. ACGT environment has been built in a SOA manner but it is required to add some semantic, some additional constraints and rules for building the ACGT services.

The goal of a layered architecture is to introduce different abstraction levels for services. Services from different layers operate on different world of terms. Some of them are located near the physical resources using hardware terms; the others are located near end-user and should be contacted using language of meta-descriptions. The other important fact is that upper layer services are specific to the scientific domain and lower level ones are more general (IT related) and could be used in a generic way by different clients:

- Service of Common Grid Layer: This layer's services access hardware resources
- Service of Advanced Grid Layer: Provides more advanced, collective functionality, using lower level services to realize client requests

- Service of Business Model Layer: These are specific services for ACGT environment. They are closer to the end-user, can operate on terms from the biomedical and cancer research domain (meta-descriptions, ontology)
- Role in ACGT: Obviously, services can play different roles in distributed environment

Most important services/tools implemented are:

- The ACGT Workflow Editor [13] is a graphical tool allowing users to combine different ACGT services into complex workflows. It is accessible through the ACGT Portal and thus has a web-based graphical user interface. It supports searching and browsing of available services and data sources and their composition through an intuitive, user-friendly interface. The workflows created can be stored in a user's specific area and later retrieved and edited so new versions of them can be produced. The publication/sharing of the workflows are also supported so the user community can exchange information and users benefit from other's research. Finally, the workflow editor supports executing the workflows and monitoring their enactment status.
- ObTiMA [14] is a complete suite for designing and running a clinical trial, it consists of a complex clinical trial management tool, workflow editor - the same application that is used in portal (can be also executed as a standalone tool) and visualization applications, various tools used for data visualization generated by other applications.
- GridR [15] is used as a tool to remotely execute R code in the grid. More specifically, the task of R code execution is submitted as a grid job to a remote grid machine. The current implementation of the server side GridR components that are related to the grid environment is based on several external software components, namely the GT4 grid middleware, an installation of the R environment on the grid machines which will execute the functions remotely and a GRMS-Server installation from the Gridge toolkit on a central machine in the grid environment that is responsible, for instance, for resource management. On the client side, GridR consists of a set of R functions and involves the Cogkit, which is responsible for proxy generation and data transfer, and a GRMS-Client. The client side part is structured around the components "RemoteExecution" (JobSubmission and JobDescription Generator) and "Locking".
- The Oncosimulator [16] is an advanced tool which is able to simulate the response of tumours and affected normal tissues to therapeutic schemes based on clinical, imaging, histopathologic and molecular data of a given cancer patient. It aims at optimizing cancer treatment on a patient-individualized basis by performing in silico (on the computer) experiments of candidate therapeutic schemes.
- The "Custodix Anonymisation Tool" (CAT) aims to simplify the process of de-identifying and exporting personal data, and is as such part of the ACGT Data Protection Framework. The tool was created to meet the demands for exporting pseudonymous data from the (internal) hospital data stores to their anonymous ACGT counterparts (i.e. the ACGT accessible data sources, also physically residing in the hospitals). The tool is innovative in a sense that it offers a generic solution regardless of the type of data to be treated or of de-identification requirements.

8.6.2.4 Metadata

The ACGT tool metadata repository handles metadata for the following main tasks:

- Publish (register) tools by service providers
- Find (discover) tools by service clients
- Bind (invoke) tools by service clients
- Modify existing tool metadata

- Retrieve all tool metadata (for metadata browsing tools)

The repository has been implemented in several layers:

- Modular API: This Application Programming Interface (API) integrates different tool repositories providing discovery/find functionality for tools, data types and functional categories using an access module to integrate with the ACGT repository databases.
- RepoServices API: This API is used internally in the Modular API access layer to the ACGT repository databases.
- The Semantic Mediator [17] is the core component of the ACGT Semantic Mediation layer. It is in charge of accepting queries in terms of the ACGT Master Ontology and translating them into terms of the physical databases included in the integration platform. The mediator can be accessed as an OGSA-DAI service making it available to any terminal connected to the internet. It is comprised by three different services:
  - SemanticMediator: This is the main service offered by the mediator (and it is called also as the resource). It offers a SPARQL interface to performe queries in terms of the Master Ontology. The received queries are handled by the mediator according to the existing database mappings, so a new query for each underlying data source is produced and their results are merged and sent back to the user
  - MappingList: This service allows retrieving the content of all mapping files included in the Semantic Mediator. This service is used by the Query Tool.
  - updateMappingList: Through this service, new mappings can be included in the integration platform. It is used by the Mapping Tool.

8.6.2.5 Security

One of the primary elements of the ACGT security infrastructure is the authorization service called GAS. The Gridge Authorization Service (GAS) is an authorization system which can act as the standard decision point for all components of a system. Security policies for all system components are stored in GAS. Using these policies GAS can return an authorization decision upon the client request. GAS has been designed in such a way that it is easy to perform integration with external components and it is easy to manage security policies for complex systems. The possibility to integrate with the Globus Toolkit and many operating system components makes GAS an attractive solution for grid applications.

GAS has been used for managing security policies: for many Virtual Organizations, for services (like Gridge Resource Management Service, Mobile Services and other) and for abstract objects like communicator conferences or computational centres. These and many other features give a possibility to integrate GAS with many existing solutions. Such integration can be very important, because it raises the security level of the existing solutions and makes it possible to use the newest security technologies.

ACGT security relies on PKI. The ACGT PKI is a commercial grade PKI implementation (with respect to the followed security practices) which follows the X.509 standards and is composed of several interdependent modules. The service is not specific for GRID infrastructures, but rather supportive to the Common Grid Infrastructure.

The Certificate Authority (CA) module is the central component that issues and signs certificates for end-users and services. It is based on the well-known EJBCA software (http://www.ejbca.org). It is not directly accessible by end-users but is used by their administration site (Registration Authority front-end, see above) and other PKI services.

Complex and long running tasks require that the end-users are able to delegate credentials to software agents (such as for example the workflow enactor), so that these agents can act

on behalf of the user while he is offline. Delegation means that a user "transfers" his access rights (typically for a restricted period of time) to another actor (service).

Delegation in ACGT is mainly provided through X.509 proxy credentials which are basically certificates signed by the end-user's certificate instead of a dedicated Certificate Authority. By issuing such a certificate the end-user delegates his rights to a specific service or person.

MyProxy is an online credential repository supporting this form of delegation. A dedicated ACGT MyProxy service was deployed and configured to allow only certificates generated by ACGT-approved CAs. Although delegation in ACGT is not restricted to MyProxy based delegation, the delegation is always bootstrapped by the MyProxy service at the portal level.

8.6.3 LifeWatch

LifeWatch (http://www.lifewatch.eu) will construct and bring into operation the facilities, hardware, software and governance structures for all aspects of biodiversity research. It will consist of: facilities for data generation and processing; a network of observatories; facilities for data integration and interoperability; virtual laboratories offering a range of analytical and modelling tools; and a Service Centre providing special services for scientific and policy users, including training and research opportunities for young scientists. The infrastructure has the support of all major European biodiversity research networks.

The LifeWatch infrastructure [18] for biodiversity research addresses the huge gaps faced in understanding of life on Earth. Its innovative design supports scientists to enter new research areas with large-scale data resources, advanced analytical and modelling capabilities with computational power. It not only serves the scientific community but is also an essential tool for local and global policy makers to understand and rationally manage our ecosystems.

Reusability, modularity, portability, interoperability, discoverability, and compliance with standards are common principles supported by LifeWatch. Its infrastructure shall:

- Rigorously use proven concepts/standards to avoid dependence on vendor specific solutions and maintain freedom to use all emerging solutions based on the standards
- Comply with the INSPIRE Directive and Implementation Guidelines for spatial data infrastructures in Europe
- Consist of loosely coupled components which can be interconnected using mediation
- Be independent of specific technologies to accommodate future technology changes
- Support an evolutionary style of development
- Be loosely coupled with external systems
- Be designed in a flexible, generic, and adaptable way for usage across different thematic areas and contexts
- Implement and deploy infrastructure using established techniques that guarantee rapid availability of components, whilst in parallel carrying out experimental research into cutting-edge technologies in selected areas to ensure adoption of new approaches, contributing to European Research Area informatics development

LifeWatch follows ORCHESTRA in promoting an incremental, iterative approach for the analysis and design phases. It distinguishes between an abstract service platform specified independently of any middleware technology and a concrete service platform that is implemented on a specific middleware – the Service Network (see Figure 10)

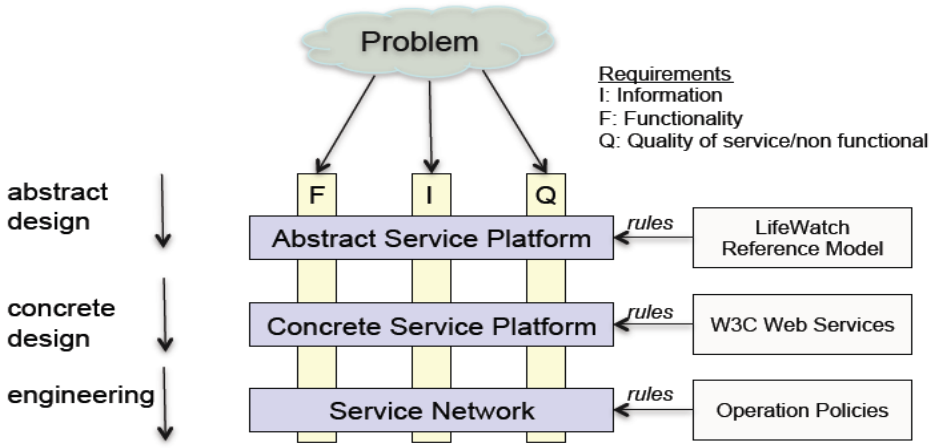


Figure 10: Abstract and concrete service platforms

The LifeWatch infrastructure is based on the assumptions that:

- Functionality is broken into component services based on the principles of SOA
- Workflows are used for the chaining of operations from multiple distributed services in order to perform specific user tasks
- Semantic Services provide uniform semantically defined interfaces enabling syntactical and semantic interoperability between and substitution of components
- Provenance information about documents, data and methods replaces traditional "laboratory notebooks" where provenance comprises all the information about the object's source/origin, history and pedigree, its derivation and passage through time

8.6.3.1 Standards

LifeWatch relies on and conforms to published standards whenever feasible. Several standards and best practices related to standards provide a guideline for the LifeWatch ICT conceptualisation and should be adopted whenever feasible and appropriate. Standardisation organisations of particular interest are:

- Biodiversity Information Standards (TDWG)
- Open Geospatial Consortium (OGC)
- Organization for the Advancement of Structured Information Standards (OASIS)
- Open Grid Forum (OGF)
- World Wide Web Consortium (W3C)

The LifeWatch Infrastructure should comply with the bases and principles of the W3C Recommendation "Architecture of the World Wide Web", which defines the WWW as an information space in which items of interests are called resources and people or software acting on this information space are called web agents. For the latter, LifeWatch will use the term participants following the usage of OASIS-SOARA.

LifeWatch Service Architecture should comply with the guidelines of ISO 19119 and the applicable W3C recommendations. From these, explicit policies shall be derived regarding the particular resources and participants involved. A service network may be mapped onto several service platforms. E.g., one platform may be based on SOAP and another on REST.

8.6.3.2 Data

LifeWatch is about biodiversity and biodiversity related data. Since biodiversity includes four levels of organisation (genetic, species, ecosystems and landscape) and related data can range from meteorological parameters to measurements of human impact, no single data model is sufficient to organise this complex information. Several, possibly overlapping, models are needed to cover multiple application areas.

The integration of biodiversity data resources faces similar problems as other fields. In addition, biodiversity data often has geospatial, temporal or taxonomic attributes, which can make discovery of appropriate data more difficult. Such data is exemplified by occurrence data, which record the observation of one or more organisms of a particular species at a particular location and time. The LifeWatch mechanism for integrating biodiversity data is the definition of information models and the mapping of data resource types to feature types, structured by an application schema. An information model specifies new data structures in terms of simpler data structures or basic data types.

Key data structures for the LifeWatch infrastructure are geospatial and temporal features and biodiversity data. These concepts will be drawn together through LifeWatch ontologies.

8.6.3.3 Services

The existing biodiversity networks and projects have already defined and implemented web services to provide access to biodiversity data and methods. The LifeWatch infrastructure will consider existing generic services when defining the LifeWatch interfaces. With the help of schema mapping services and other mediation mechanisms it should be easy to attach such existing services to the infrastructure, e.g. for taxonomic search or validation.

The Services Framework provides the basis for the specification of LifeWatch services, the service model, and a model for service chaining. The service model is based on the fundamentals of Service Oriented Architecture defined by the interaction between service consumers and service providers within a distributed system. The interaction between consumer and provider is performed by service requests, service responses and service exceptions. Services are determined by their interfaces that consist of operations, according to ISO/DIS 19119. A Service Type is defined by the specification of the externally visible behaviour of a service through its interfaces. The service model provides rules for the specification of service types with the target of providing the syntactic and semantic interoperability between services, source systems and applications. The service model considers therefore two levels of specification of service types:

- Platform neutral: Abstract description of the services and an abstract specification of their interfaces (in UML)
- Platform specific: Implementation of the services and interfaces in a platform specific paradigm, e.g. Web Services

Source system integration refers to the transformation of a non-Lifewatch source system into a LifeWatch Service Instance (within the LifeWatch service network conforming to the meta-models and interfaces). Due to the heterogeneity of External Source Systems, one cannot expect to define a service type with predefined interfaces covering all these systems. Hence the term Source System Integration Service is only used as a generic name for the class of services serving this purpose. Those integration Services must be implemented by the resource providers as part of the admission procedures to join the LifeWatch infrastructure.

The layered service groupings indicate broad dependencies between services, with services from the higher layers using services from the same layer or from the layers beneath. The

System Management Services are orthogonal to the other service groups, as they are used at all horizontal layers and use or are used by the other services.

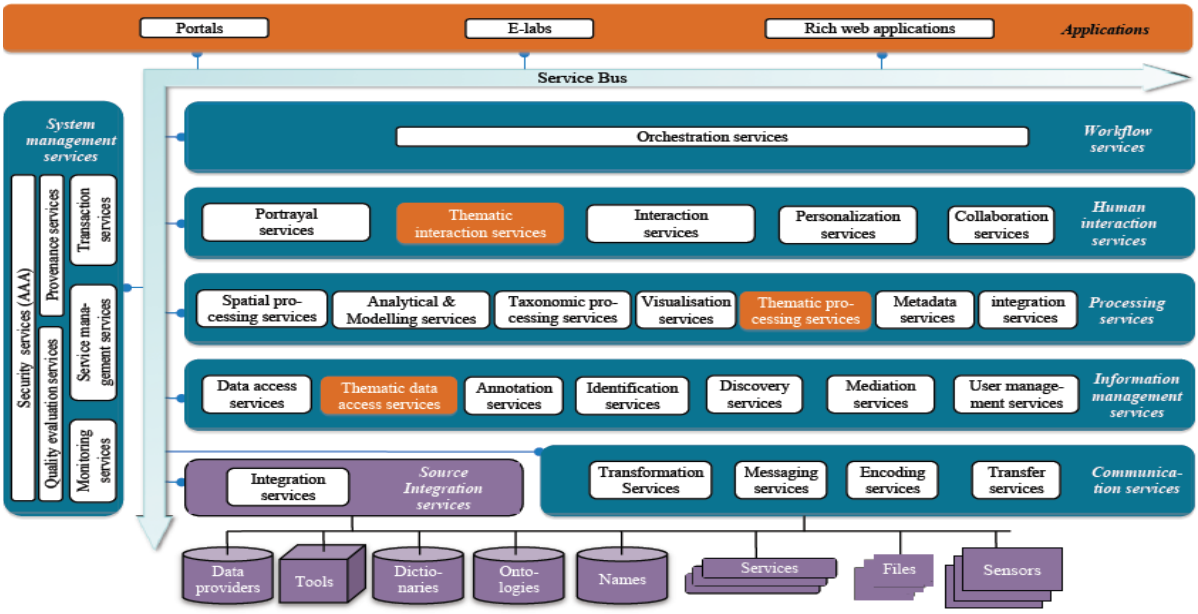


Figure 11: Service-oriented architecture

8.6.3.4 Metadata

Semantic mediation will be another cornerstone for interoperability. It will be a differentiating feature of the LifeWatch approach. The need for semantic mediation arises in several areas:

- Data and Service Discovery: Discover data and services based on, for example, specific domain ontologies
- Data Mediation: Processing data based on its semantics even if the data is provided by different data models
- Data Fusion: Combining data from different sources
- Data Interpretation: Multiple data models and heterogeneity at the data level itself, perhaps arising from differences of professional opinion, makes interpretation harder
- Service Integration: Chaining of services often needs transformation of data when passing data from one service to another
- Workflow Identification: Discovery of workflows that, e.g. may help to solve a particular modelling problem

Semantic mediation will be achieved by several means, presently under investigation. Mechanisms being considered include:

- Taxonomic checklists and associated tools for validating the integrity of checklists and crossmapping between checklists
- Use of ontologies, including distinct ontology classes for different application domains
- Use of semantic web innovations
- Use of rules-based reasoners: once feature types and ontologies are defined, it becomes possible to reason over the relations

ISO 19115 and INSPIRE propose models for metadata that help to describe resources and provide quality information about those resources. Following ORCHESTRA, LifeWatch will

define meta-information models, i.e. sets of metadata according to specific purpose, and provide rules how to specify these.

Meta-information models have particular relevance for LifeWatch to support description, discovery and use of resources by users. Yet, the extent to which the abstract infrastructure design should only provide generic meta-information models versus comprehensive meta-models for biodiversity data is not yet fully clear requiring further investigation.

8.6.3.5 Security

According to the OASIS-SOA-RA, security can be defined in terms of the "social structures that define the legitimate permissions, obligations, and roles of people in relation to the system, and mechanisms that must be put into place to realize a secure system". ISO/IEC 27002 characterizes the following key security concepts:

- Confidentiality: Protection of privacy of participants in their interactions: messages should not be readable to third parties, but the degree of visibility if messages are exchanged and of the participant's identity to third parties can be defined
- Integrity: Protection of altering exchanged information
- Availability: Concerns the reliability of a system, in other words, if the system offers the service for which it is designed, and the security concept needed to respond to active threats to the system
- Authentication: Concerns the means of identifying the participants in an interaction
- Authorization: Concerns the means of legitimacy of the interaction, the exchanged actions must be explicitly or implicitly approved
- Non-repudiation: Concerns the accountability of participants: participants should not, at a later time, successfully deny having participated in the interactions

8.6.3.5.1 LifeWatch Policies

- Security policies require mechanisms to support security description administration, storage, and distribution. These mechanisms shall be defined within LifeWatch.
- Security policies should be able to express trust relationships and domains, providing the ability to update trust relationships without changes in the hard- and software.
- Standard Protocols should be used to provide confidentiality, integrity, authentication, authorization, non-repudiation, and availability.
- Service Specifications and Service Descriptions should be able to reference one or more security policy artefacts.
- A Service Network should provide mechanisms for:
    - Protection of the confidentiality and integrity of messages exchange
    - Policy-based identification, authorization, and authentication
    - Ensuring service availability to the consumers
    - Ensuring security for a scalable network and between different platforms
- A Service Network should include instances of the following security services:
    - Authentication Service
    - Authorization Service
    - User Management Service

- Service Monitoring Service for Security including monitoring of intrusion detection and prevention, auditing and logging of interactions, security violations, service availability, and support for quality of services.

- LifeWatch Services should use an Encryption Interface to abstract encryption techniques, allowing the use of different techniques.
- There shall be an agreed-upon list of Security Policies, e.g. for the Network- and Transport-Layer to be supported by all LifeWatch Service Networks. These policies may also include generally valid aspects of the three security models (trust, treat model, and security response).
- A threat model shall be defined in form of a list of exceptions thrown by the Service Monitoring Service.

8.6.4 myGrid

myGrid (http://www.mygrid.org.uk) is an e-Science research project developing open-source high-level middleware to support in silico experiments in biology using databases and computational analysis tools rather than laboratory investigations to test hypothesis. Information repositories, service registries and change notification systems are all being developed to provide personalized views of resources. myGrid components make extensive use of metadata to support this need for personalisation and the project is pioneering the use of semantic web technologies to manage annotation, ontologies and semantic discovery. myGrid's ultimate goal is to supply this service collection as toolkit to build end applications.

8.6.4.1 Services

The myGrid middleware framework employs a service-based architecture, firstly prototyped with Web Services but with an anticipated migration path to the Open Grid Services Architecture (OGSA). The myGrid team develop and use tools to allow users access to the capabilities of e-Science. The tools can be categorized functionally as supporting:

- Workflow creation, editing and execution, such as Taverna
- Social networking by scientists and the sharing of workflows, like myExperiment.
- Interfaces that are specialized for the needs of users, such as UTOPIA
- Service creation and registry tools, like SoapLab and the BioCatalogue
- Ontology creation and checking tools, such as Protege
- Metadata encapsulation - Scientific Research Objects
- Use of portal creation software

The tools can also be categorized according to their origin:

- Tools developed by the consortium specifically for the e-Laboratory, like Taverna and myExperiment
- Tools developed by consortium members used by the e-Laboratory, such as COHSE
- External tools such as Protégé

8.6.5 Taverna

Taverna [19] is an open-source, domain-independent Workflow Management System, a suite of tools used to design and execute scientific workflows and aid in silico experimentation.

The Taverna suite is written in Java and includes the Taverna Engine (used for enacting workflows) that powers both the Taverna Workbench (the desktop client application) and the Taverna Server (which allows remote execution of workflows). Taverna is also available as a Command Line Tool for a quick execution of workflows from a terminal. Its main features are:

- Fully featured, extensible and scalable scientific Workflow Management System
- Available as a desktop Workbench, from a command line or remotely as a Server
- Access to local and remote resources and analysis tools, Web and grid services; 3500+ services available on startup
- Support for calling tools/scripts on local or remote machines as part of a workflow
- Not restricted to predetermined services, rapid services incorporation without coding
- Extensible service plug-in architecture for adding new service types
- Up-to-date R support (version 2.11.1)
- Excel and CSV spreadsheet support
- Secure access to resources on the Web
- Service packaging for customizing and sharing, e.g. “Taverna for chemists”
- Standards-compliant provenance collection
- Graphical workflow designer – drag and drop workflow components
- Workflow validation during design time and intermediate values during workflow runs for debugging workflows
- Cross platform (written in Java), open-source, LGPL-licensed

8.6.6 myExperiment

myExperiment [20] is a collaborative environment where scientists can safely publish their workflows and experiment plans, share them with groups and find those of others. Workflows, other digital objects and bundles (called Packs) can now be swapped, sorted and searched like photos and videos on the Web. Unlike Facebook or MySpace, myExperiment fully understands the needs of the researcher and makes it really easy for the next generation of scientists to contribute to a pool of scientific methods, build communities and form relationships, reducing time-to-experiment, sharing expertise and avoiding reinvention. It is now the largest public repository of scientific workflows and is Linked Data compliant.

8.6.6.1 Metadata

MyGrid’s Taverna Workbench services can be annotated with semantic descriptions based on ontologies and later discovered based on these descriptions. The myGRID ontology is expressed in DAML+OIL (a predecessor of OWL). In order to annotate a given service semantically, the Taverna Workbench integrates with the PeDRo tool.

The PeDRo tool [21] provides a graphical interface through which a user is guided to fill the missing semantic information and build XML descriptions of its services using the MyGRID ontology suite. These XML descriptions can then be published to a WebDAV server and advertised to a UDDI registry. On the other hand, the Taverna Workbench offers also the ability for the semantic discovery of services through the Feta component. Feta is composed of two sub-components, the Feta Client GUI that is the user interface for the formulation of semantic queries, and the Feta Engine which is a web service responsible for searching

service descriptions that match user's search criteria. The Feta client side GUI is currently able to formulate a number of canned queries such as:

- Find an operation accepting input of semantic type X or something more general, or find an operation that produces output of semantic type Y or something more specific
- Find an operation that performs task X or something more specific
- Find an operation that uses method X or something more specific
- Find an operation that is function of application/toolkit X or something more specific

8.6.7 Feta

Feta [22] is a semantic discovery tool that can be used to search available services and find those that best match the requirements of the user. Feta can assist the user in workflow design, shortening time taken to discover services and incorporate them into workflows. Feta can also provide extra information on the format required or produced for inputs or outputs.

Another important function of Feta is that it can provide information on alternative services. Several organisations can provide different implementations of the same application, or provide very similar applications that essentially perform the same function. The semantic discovery capabilities can be used to find alternative services if a particular web service is unavailable, or past experience has shown that a particular web service is unreliable or slow.

8.6.7.1 Security

In myGrid security is limited in comparison to the other integration environments. Taverna is the basic frontend of myGrid and can be used to access all sorts of services secured with HTTPS, HTTP Basic authentication and WS-Security for securing Web services.

For WS-Security Taverna supports the portion of the WS-Security standard that refers to username and password authentication. Depending on a service's settings, Taverna will add plaintext or digest password as part of SOAP messages sent to that service.

The Credential Manager is a Taverna tool to manage credentials and certificates of services you wish to invoke. It stores username/password pairs and private key certificates securely and remembers which credentials you want to use for which services. The user does not have to enter them each time when invoking a secure service from a workflow. In this respect, it is similar to the password manager in Firefox or Internet Explorer, or Keychain.

8.7 References

[1] Perry D, Wolf A (1992) Foundations for the study of software architecture. SIGSOFT Softw Eng Notes. 17(4): 40-52. DOI 10.1145/141874.141884

[2] Soni D, Nord R, Hsu L, Drongowski P (1993) Many faces of software architecture. In: Lamb D, Cracker S (eds.), Proc. Worksh. on studies in software design, Baltimore, Springer

[3] Kruchten P (1995) The 4+1 view model of architecture. IEEE Software, 12 (6): 42-50

[4] Kruchten P, Obbink H, Stafford J (2006) The past, present, and future for software architecture. Software, IEEE, 23(2): 22-30, DOI 10.1109/MS.2006.59

[5] Rozanski N, Woods E (2005) Software systems architecture: working with stakeholders using viewpoints and perspectives. Addison-Wesley Professional. ISBN 978-0321112293

[6] Garlan D, Shaw M (1994) An introduction to software architecture. CMU-CS-94-166

[7] Clements P, Bachmann F, Bass L, Garlan D, Ivers J, Little R, Merson P, Nord R, Stafford J (2010) Documenting software architectures: views and beyond. Addison-Wesley Professional (2nd edition) ISBN 978-0321552686

[8] Stal M (2006) Using architectural patterns and blueprints SOA

[9] Richardson L, Ruby S. RESTful web services (2007) O'Reilly, ISBN 978-0-596-52926-0

[10] Hohpe G, Woolf B (2003) Enterprise integration patterns: designing, building, and deploying messaging solutions. Addison-Wesley Professional, ISBN 978-0321200686

[11] Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, Kurc T, Siebenlist F, Covitz P, Shanbhag K, Foster I, Saltz J (2008) caGrid 1.0: an enterprise grid infrastructure for biomedical research. J Am Med Inform Assoc. 15(2): 138–149

[12] Tsiknakis M, Brochhausen M, Nabrzyski J, Pucacki J, Sfakianakis SG, Potamias G, Desmedt C, Kafetzopoulos D (2008) A semantic grid infrastructure enabling integrated access and analysis of multilevel biomedical data in support of postgenomic clinical trials on cancer. IEEE Trans Inf Technol Biomed. 12(2):205-17

[13] Sfakianakis S, Koumakis L, Zacharioudakis G, Tsiknakis M (2009) Web-based authoring and secure enactment of bioinformatics workflows. Intern Workshop Workflow Management

[14] Brochhausen M, Weiler G, Schera F, Rauch J, Graf N, Kiefer S (2009) Ontology-based trial management system (ObTiMA). nature preceedings. doi:10.1038

[15] Wegener D, Sengstag T, Sfakianakis S, Rüping S, Assi A (2007) GridR: an R-based grid-enabled tool for data analysis in ACGT clinico-genomics trials. Proc IEEE International Conference on e-Science and Grid Computing. pp. 228-235

[16] Stamatakos G, Dionysiou D, Graf N, Sofra N, Desmedt C, Hoppe A, Uzunoglu N, Tsiknakis M (2007) The Oncosimulator: a multilevel, clinically oriented simulation system of tumour growth and organism response to therapeutic schemes. Towards the clinical evaluation of in silico oncology. Proc 29th Annual Intern Conf IEEE EMBS. pp. 6628-6631

[17] Martín L, Anguita A, de la Calle G, García-Remesal M, Crespo J, Tsiknakis M, Maojo V (2007) Semantic data integration in the European ACGT project. AMIA Annu Symp. p. 1042

[18] Ernst V, Poigné A, Giddy J, Hardisty A, Voss A, Voss H (2009) Towards a reference model for the LifeWatch ICT Infrastructure. GI Jahrestagung. pp. 654-668

[19] Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, Oinn T (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Research. 34: 729-732

[20] De Roure D, Goble C, Stevens R (2008) The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems. 25: 561-567

[21] Garwood K, Lord P, Parkinson H, Paton N, Goble C (2005) Pedro ontology services: a framework for rapid ontology markup. Lecture Notes in Computer Science. 3532/2005: 41-55, DOI: 10.1007/11431053_39

[22] Lord P, Alper P, Wroe C, Goble C (2005). Feta: A light-weight architecture for user oriented semantic service discovery. Proc European Semantic Web Conference. pp. 17-31

9 Usability Process

9.1 Introduction

In the project p-medicine a service oriented clinical research infrastructure will be developed to improve prognosis of patients with cancer by individualizing treatment and going to personalized medicine. The use of computers, software applications and IT in daily medical life and in research is rapidly increasing; see its precursor ACGT (http://eu-acgt.org).

The main task to assure usability of the developed systems is the accomplishment of the users' needs. Generally, software is developed without the evaluation during the development process. To avoid this well-known risk, it is of utmost importance to involve the end-user from the design phase of new software, during the development process and to secure an iterative evaluation of the software by end-users.

Without taking the end-user into account, the software will fail usability at the product stage and end-users would not use the software and platform – a serious loss of time, money and resources for the project. To guarantee the necessity of usability, a series of international standards are developed with basic recommendations for products and process design.

In p-medicine the developed software is evaluated considering the following approaches:

- The first one described in D2.2 is the elicitation of context scenarios from interviews with the prospective user groups. These scenarios describe the whole context of use of the user's daily work with the software. From the resulting user needs mock ups can be generated which reflect the user's task.
- The second one is to assure that the software is working as expected from an end-user perspective (i.e. to ensure that it meets the demands of the target groups) by evaluating its usability as a whole, in a feedback loop between developers and users. This should guarantee the usability, and ultimately the actual use, of the software by the biomedical community.
- The third one is to involve the cancer patient in the development process. This means to take the patients' task into account and consider their user needs. These needs depend on the acquisition of information about their kind of disease, to become healthy again and other important issues.
- These evaluation criteria and processes are user driven from the beginning of the project to assure that all requirements of end-users are covered during the process of software development.
- To assure the delivery of a high-class research environment, p-medicine has to guarantee that only high quality software and tools are implemented in the platform. This document can hence be viewed as a set of guidelines for software developers, rendering explicit the criteria that have to be fulfilled by a "candidate" software (developed either inside or outside p-medicine) to meet the standards of p-medicine. By considering them in the development process, these evaluation criteria become an integral part of the quality-assurance mechanism in p-medicine.
- With the help of end-users an assessment of the usability of the p-medicine environment will be possible, even after publication. End-users will be asked to evaluate the software and tools they use according to the usability criteria provided here to give a direct feedback to the developers, thereby ensuring the continuity of the optimization process.

This document describes the usability process from the beginning of the project to the evaluation and testing of the developed software tools.

Different user groups are identified as major target groups in the present document: clinicians and healthcare professionals, biomedical researchers, data managers and patients, each having specific usability criteria in relation to the nature of their activities and associated tools, see deliverable D2.2. Specific sections of this document elicit those needs.

Data collection within clinic genomic trials and interdisciplinary analysis by clinicians, biostatisticians/-informaticians, data managers, patients and developers/researchers is mandatory to further improve the outcome of cancer patients. A major problem of sharing all the available clinical data to the different users for analysing and storing must be possible by all integrated VPH models, clinical practice and omics data into a comprehensive and usable platform for all involved user groups.

There are many problems related to the different data sources, data of the clinical pathologies may be stored in different formats in different hospitals. Sharing data to larger researcher communities is the first step that must be conducted assumed that there are standardize data and shared tools in the p-medicine environment. An important idea of patient empowerment is to enable patients themselves to be participants in their own health care. Patients can inform about her/his disease and the progress of treatment.

For all these various user groups and user interfaces built in the whole environment of p-medicine it is important to start with the usability process in the beginning of the project. The various user groups must be interviewed to get the respective user requirements and their needs to conduct their task in an efficient and satisfied way. This should be done before the implementation phase of software will be started.

9.2 General End-User Evaluation Aspects for Usability of Developed Software in p-medicine

9.2.1 State of the Art

Standards related to usability can be categorised as primarily concerned with:

- Use of the product (effectiveness, efficiency, satisfaction in a particular use context)
- User interface and interaction
- Process used to develop the product
- Capability of an organisation to apply user centred design

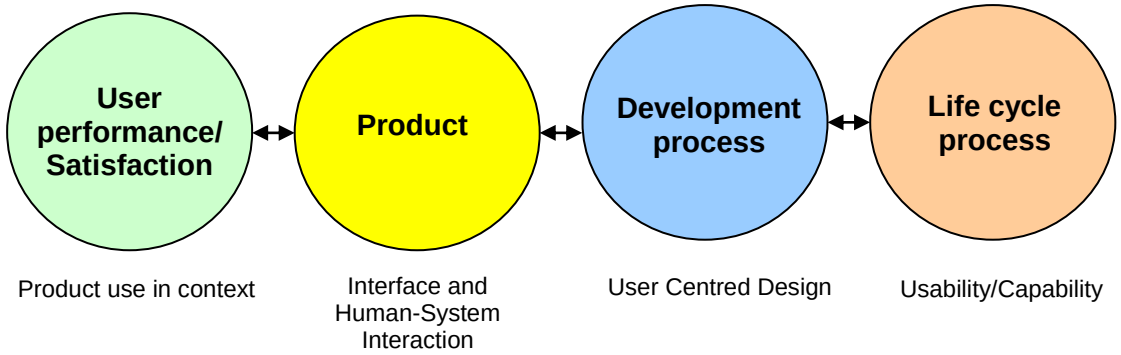


Figure 1: Coherences in software development (http://www.usabilitynet.org)

9.2.1.1 ISO Principles and Recommendations for Usability and Software Engineering

	Principles and recommendations	Specifications
Use in context	ISO/IEC 9126-1: Software Engineering – Product quality – Part 1: Quality model	ISO 20282-1:2006 Ease of operation of everyday products – Part 1: Design requirements for context of use and user characteristics
	ISO/IEC 25000:2005: Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE; refers only on product quality and not on process quality	
	ISO/IEC 25040:2011: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Evaluation process	
	ISO 9241-11: Ergonomics of human-system interaction. Part 11: Guidance on usability	
	ISO 9241-151:2008 Ergonomics of human-system interaction. Part 151: Guidance on software accessibility	
Interface and interaction	ISO/IEC TR 9126-2: Software Engineering – Product quality – Part 2: External metrics	ISO 9241:2006 Ergonomics of human-system interaction
	ISO/IEC TR 9126-3: Software Engineering - Product quality – Part 3 Internal metrics	ISO/IEC 10741-1: Information technology – User system interfaces – Dialogue interaction – Part 1: Cursor control for text editing
	ISO 9241-110:2008 Ergonomics of human-system interaction – Part 110: Dialogue principles	ISO/IEC 11581: Information technology – User System interfaces – Icon symbols and functions
	ISO 9241-12: Ergonomics of human-system interaction – Part 12: Presentation of information	ISO/IEC 14598: Information technology – Software product evaluation
	ISO 9241-151:2008 Ergonomics of human-system interaction – Part 151: Guidance on world wide web user interfaces	ISO 12119: Information technology – Software packages – Quality requirements and testing
Documen-tation	ISO/IEC 18019: Guidelines for the design and preparation of software user documentation	ISO/IEC 15910: 1999 Information technology - Software user documentation process

Develop-ment process	ISO 9241 - 210:2009 Ergonomics of human-system interaction – Part 210: Human - centred design for interactive systems; replaces the ISO 13407	ISO/IEC 14598 - 3: Software engineering – Product evaluation, Part 3: Process for developers
	ISO TR 16982: Usability methods supporting human centred design	

Table 1: ISO norm standards related to usability and software engineering

Many standards are related to or affect the usability of computer software and applications. These standards have to be taken into account during the developmental process. For the evaluation of usability regarding developed software and tools the ISO 9241-11 [1] guidance of usability is the most relevant standard today. It describes an objective, structured process to identify the users' requirements for the software and the mechanism to modify software applications and procedures with regard to the functionality and usability of the software. The process described in the ISO 9241-11 will be used as a guideline for the evaluation of the usability. To respect the complexity of the p-medicine project, the standard ISO/IEC 25000:2005 [2] is used to comprise the needs of software developers.

p-medicine will use these standards as basic quality criteria to test all prototypes. The usability process regarding the dialogue principles of ISO 9241 – part 110 [3], part 11 and 12 will be the basic standard regarding end-user needs, while the needs for software specification can be specified using the criteria of ISO 25000 norm.

9.2.2 Black Box Model

White or glass box testing requires the knowledge about the program internals, while black box testing is based on the requirements of the end-users from the perspective of end-users. Because of the complex structure and architecture of p-medicine both testing methods are useful and needed. In respect to the high complexity of the p-medicine software environment, it cannot be expected from end-users to understand it in detail. A white box approach is thus considered impractical. To secure the success of p-medicine clinicians, bio-researchers, data managers and cancer patients as end-users will use the p-medicine platform only as a black box. In respect to the user's requirements the black box testing and the white box testing have to be done by defining requirements in collaboration with all acknowledged end-users.

9.2.3 White Box Model

Structural testing is an approach in which the internal control structure of a program is used to guide the selection of test data. It is an attempt to take the internal functional properties of a program into account during test data generation and to avoid the limitations of black box functional testing. Functional testing takes into account both functional requirements of a system and important functional properties that are part of its design or implementation and which are not described in the requirements. In functional testing, a program is considered to be a function and is thought of in terms of input values and corresponding output values [2].

In the following the usability approach will be described in detail described according to the various user groups of p-medicine.

9.3 Approach of the Usability Process in p-medicine

The usability method that is used for the whole usability process in p-medicine is described in http://www.dakks.de/sites/default/files/71-SD-2-007_Leitfaden%20Usability%201.3.pdf. This document (only available in German) describes the requirements for usability and offers guidance for practitioners to conduct usability tests of interactive software systems in particular their conformance with part 110 and part 11 of ISO 9241. It has been developed by the DAkkS [4] as the successor of DATech, acquired the same content. The permanent working group DAkkS "Usability Engineering" has been working for more than 10 years in the development of test methods published in this guide. The DAkkS is the national accreditation body of the Federal Republic of Germany. It is legally mandated to carry out accreditations of conformity assessment bodies.

The DAkkS accredits certificate authorities in respect to the directives of the international organisations of standards (http://www.iso.org). The DIN EN ISO 13407 was the standard in Germany for „User-centred design of interactive systems" describing a prototyping software developmental process which was replaced by the new standard DIN EN ISO 9241- 210 [5] in March 2010. The process consists of four main topics:

1. Use-Context: documented description of relevant users, daily work and work station
2. Specify requirements: the documented description of the use context is to align the needs of users to the software demands and the relevant software specifications
3. Describe solutions: this can be done in the form of prototyping and mock ups or other iterative processes
4. Evaluate the solutions: the prototypes are evaluated by expert-reviews or usability tests, online-evaluation or a mixture of them. Modifications for the next developmental step are based on the evaluation of the discovered variances and the lack of usability

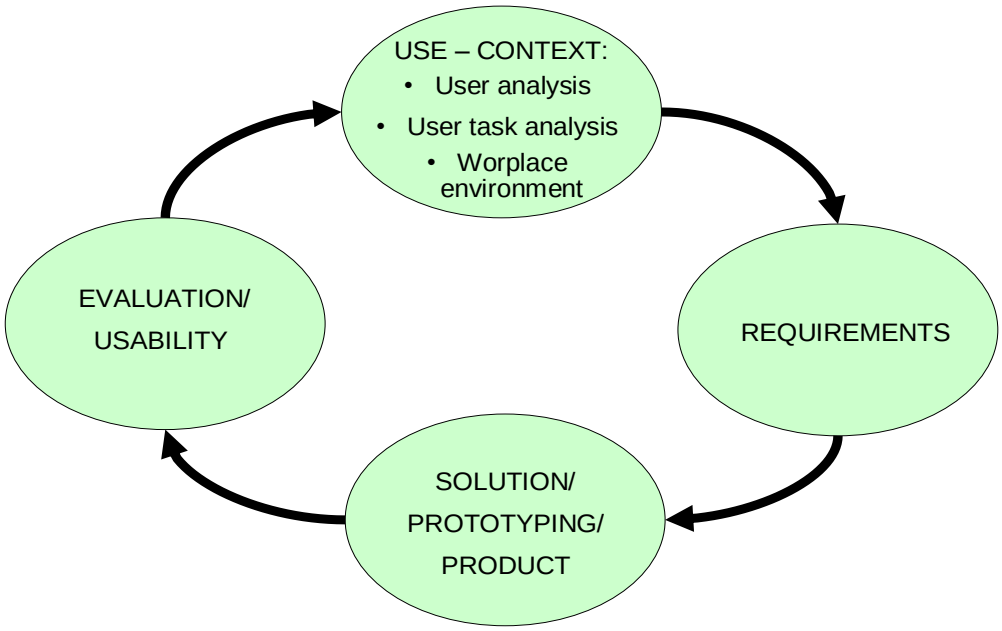


Figure 2: Schematic Structural process of ISO 9241-210

For the usability process based on ISO 9241 – part 11, the guidance on usability and – part 110 the dialogue principles (see below) play an important role in the evaluation process of developed tools in p-medicine. As a complex research project in the medical research area, it has to specify and elevate the important user requirements of all prospective user groups. This is generally necessary to use the developed tools in an efficient and satisfied way by all

user groups. Usability as part of the tool design and development involves the systematic identification of requirements for usability during the whole development loop (Figure 3).

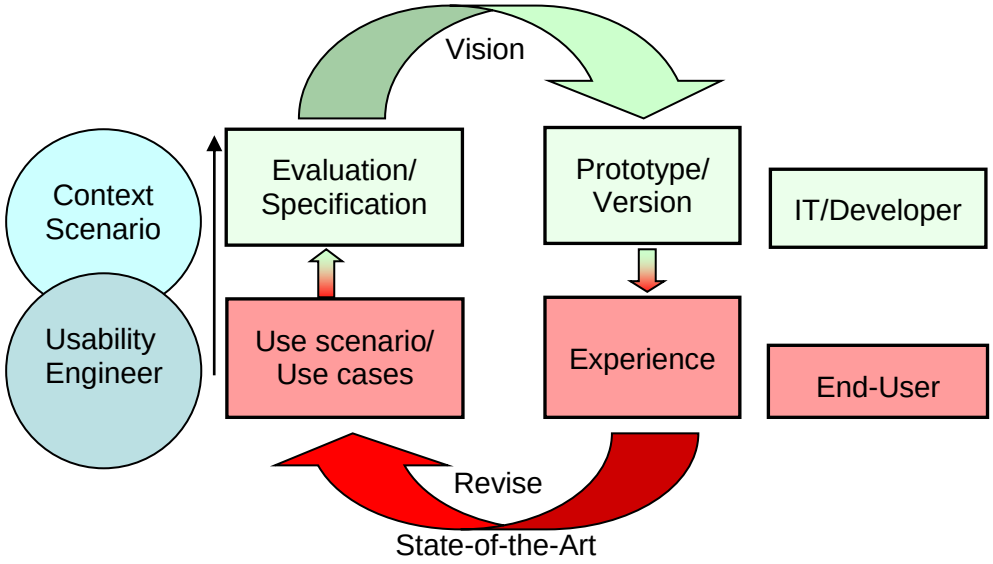


Figure 3: Development loop

9.3.1 Development Loop

Usability is concerned with the extent to which the users of products are able to work effectively, efficiently and with satisfaction using the product according to the implied needs of the context of use to achieve a specific goal [1]. The software has to fit to the users' needs and not vice versa as practiced in the past.

Based on the DAkkS usability method [4], the effectiveness represents the basic prerequisite for good quality software according to ISO 12119. This norm is in the standard norm 9241 – part 11 defined as follows, "the accuracy and completeness which enables users to achieve a specific goal". Efficiency means that the functionality can be used as part of the special context of use, i.e. it is the extent to which a product can be used. With the freedom from discomfort and positive attitudes towards the use of the tool the satisfaction will be described.

9.3.2 Usability Engineer Process (UEP)

Usability Engineering describes a pragmatic approach to interface design emphasizing empirical methods and operational definitions of user requirements for tools concerning software ergonomics. To define these requirements the usability engineer defines the users' needs in relation to his working place and the software concepts or prototypes developed. Extending as far as International Standards Organization-approved definitions (see e.g. ISO 9241 part 11) usability is considered a context-dependent agreement of the effectiveness, efficiency and satisfaction with which specific users should be able to perform tasks [1]. In p-medicine this process will be performed during the prototyping period to assure that the users' needs are satisfied. To objectify this approach, the usability engineer has conducted interviews with end-users which result in context scenarios D2.2. After confirmation regarding content and correctness by the end-users the report is sent to the software developer.

To develop very early a first prototype that must not have the complete functionality of the specification is an essential role in the usability engineering cycle. It should give a first impression of the interface for the end-user to start a first usability test.

9.3.3 Usability Engineer (UE)

To formalize and objectify the usability criteria and advance and administer the usability performance as well as interactive processes in the project, the usability engineer (UE) will define the concrete usability concepts written in this document and has the functionality of an independent agent between the end-users and the software developers. The engineer analysis the process during the whole development period from the first design defined in the user requirements D2.2 For the success of p-medicine it is of high importance that the software is self-explanatory and easy to use, because the main user groups have none or basic knowledge of computer systems or applications, especially the cancer patients. The user interfaces, as a gateway between project and end-users, is of fundamental impact. The first step in the usability process (the user interviews) started at the first meeting in Homburg.

9.3.4 Mechanism and Evaluation Strategy

This section states the mechanism used to interview and evaluate the p-medicine software from an end-users perspective regarding usability criteria with the help of an UE. The first step in the UEP is the identification of the prospective user groups. In p-medicine there are several user groups:

- Clinicians
- Researchers
- Bioinformaticians
- Biostatisticians
- Data managers
- Cancer patients

These user groups have to be enabled to conduct their daily work with the software tools and achieve their goal in an efficient and satisfied way. The tools developers belong not to these user groups we consider in the following procedure.

With the representatives of each user group interviews had to be taken early in the project. A general list of key questions (see Appendix B) exists to get a common understanding of the user's daily task in the whole context of use. These key questions depend on the task and qualification of the user's working place. Therefore we adapted the key questions to the task of the bioinformaticians-statisticians and data managers who are presented in D2.2.

The key questions recognize the task characteristics, users' prior knowledge and qualification, his working environment, and his specific way of working, organisational conditions and other elements of the context of use. This procedure is an essential process for the usability engineer. The aim is to understand the task and the whole context of use of the user to describe the user's requirements in the sense of context scenarios. The different context scenarios from a bioinformatician, a biostatistician, a clinician in the role of a physician, of an administrator and a biologist in a clinic are all given in D2.2

Ecancer elevated a corresponding survey for the cancer patients presented on the internet to give cancer patients the chance to respond. The answers and evaluation are shown in D2.2.

9.3.5 Schematic Procedure of Usability Testing

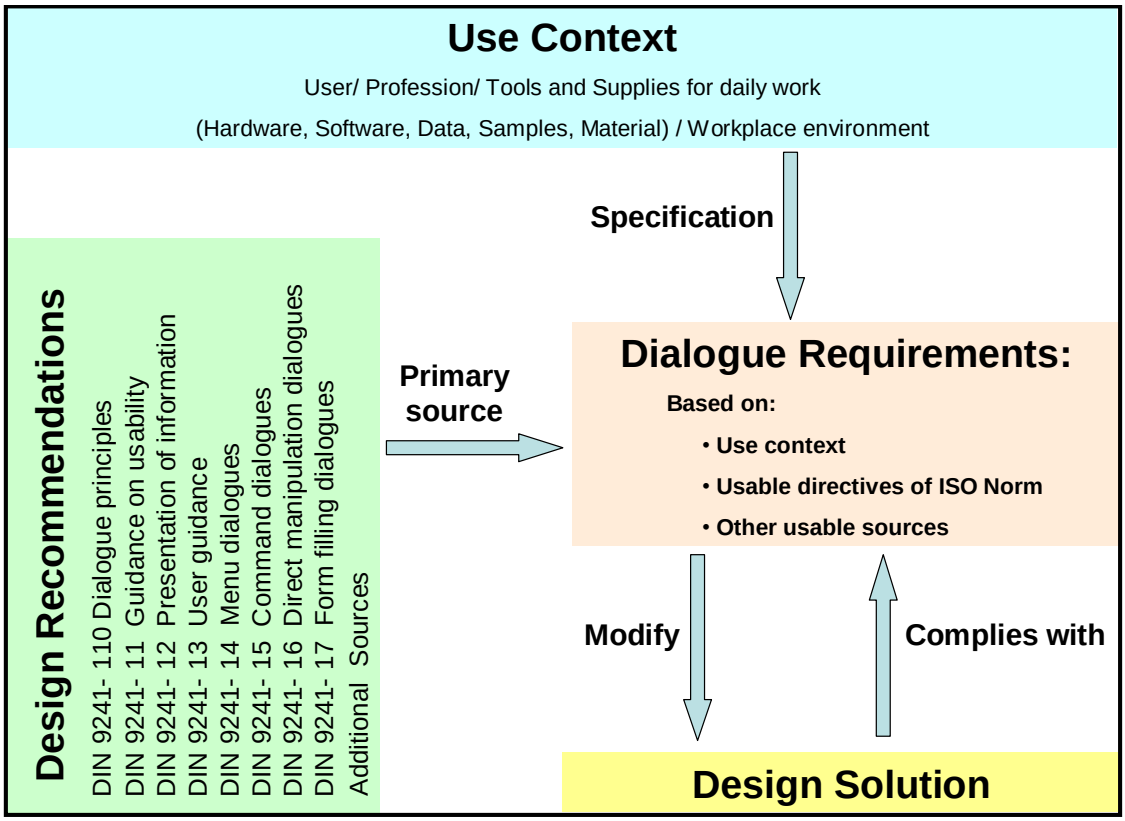


Figure 4: Framework execution regarding DIN EN ISO 9241-110

The framework execution (see Figure 4) and interviews are executed by a usability engineer in close collaboration with the users and developers, responsible for the evaluation and validation activities conducted in the project. The major steps in usability evaluation will be:

- Context scenario: Understand users' work and their work environment organization
- Usability prototyping
- Use scenario / Use cases
    - Use scenarios describe the action of the user with the system related to a predefined task and identify the user problems step by step
    - A use case describes only the action of the user and the system reaction
- Evaluation and elimination of initial usage problems
- Validation and Evaluation of product: Validation of defined use cases for usability

9.3.6 Context Scenario

The context scenario describes the task and the whole context of use in the real life situation of the end-users taking into account:

9.3.6.1 Design Recommendations

- DIN 9241-110 Dialogue principles

- DIN 9241-11 Guidance on usability
- DIN 9241-12 Presentation of information
- DIN 9241-13 User guidance
- DIN 9241-14 Menu dialogues
- DIN 9241-15 Command dialogues
- DIN 9241-16 Direct manipulation dialogues
- DIN 9241-17 Form filling dialogues

The UE documents the interviews in an objective report. With the support of the dialogue principles the UE identifies first the user needs and derives the system requirements. This report is sent to the end-user for validation and after that to the software developer to achieve a common understanding of the whole task and working steps. The relevant key questions for the interview are shown in the templates in the Appendix A.

With usability prototyping is meant to demonstrate the user the well-understood requirements and get a first feedback if it is what the user expected. This prototype must not have the full functionality of the tool. User requirements can be further augmented during the software development process and are not fixed. Software design means to have a common understanding about the functionality and the possibility of the developed software tools.

9.3.7 Dialogue Principles

Part 110 is related to the user-interfaces of interactive dialogue systems. It assigns seven criteria to be fundamental in the dialogue for successful interactive dialogue/interface design.

9.3.7.1 Suitability for the Task

A software program or a tool is suitable when it supports the actions of the user and executes his task in an efficient and effective way that helps the user to ease his daily life and not complicate it. While most of the procedures done manually today p-medicine will deliver an infrastructure to facilitate the translation from current practice to medicine by integrating VPH models, clinical practice, imaging and omics data into a single comprehensible biomedical platform. Nevertheless this aim can only be reached when the developed tools, software and workflows process the data in a suitable processing speed and without complicated editing and sufficient support.

9.3.7.2 Self-Descriptiveness

Self-descriptiveness is the degree to which a system or component contains enough information to explain its objectives and properties. The user should understand the tool by itself and the entries he should perform are obvious without external descriptions. In any step it must be obvious for the user to know in which dialogue and at which in the dialogue he will be and what will be the next action to execute.

9.3.7.3 Controllability

Controllability plays a crucial role in many computational processes. In the broadest sense it is the ability to use the entire configuration of a system without severe errors or failures of the system. The user should never get the feeling, that the system fails or the computer crashes or a workflow is not performed adequately. The user knows in every situation what is expected from him as next step or how to return. He can control the velocity and direction.

9.3.7.4 Conformity with User Expectations

Per definition conformity with user expectations means acting according to certain accepted standards and the users' knowledge and experience of software systems in means of consistency. In p-medicine conformity is important for the usability on several levels:

- Workplace conformity: The user must have the feeling that the software of p-medicine fits and supports him at his normal working place. For these needs the working places of the end-users will be described in an objective way by the usability engineer in a report and aligned with the software developed for the special needs of the individual user
- Design conformity: e.g. the software interfaces of tools developed for p-medicine should be designed in a similar way and recognised by the user as part of p-medicine
- Format conformity: The formats used in the daily research and used for the p-medicine platform are already defined. Nevertheless the need of implementation has to be iteratively evaluated.

9.3.7.5 Error Tolerance

Input data errors often occur during the test phase and when a user learns about a new tool. While during the test phase the discovery of failures in the system is the aim, the usability of the software regarding the user will be strictly bound to the way he can interact with the system. Missing usability can be expected if:

- Input data errors led to severe complication or handling problems
- Input data errors can not be cancelled
- The user is not warned before an input data error causes a problem
- No mechanism to avoid input data errors is available.

The user will judge the systems usability in the way data input errors are avoided. A loss of time or even worse of data will damage the reliability of the software. The user should have always the opportunity to solve problems in an easy and efficient way and get helpful instructions by error messaging described in the user's language.

9.3.7.6 Suitability for Individualization

Suitability for individualization of the software does not only mean to design the interface regarding colour and minor functions. The user should have the opportunity to set up settings that allow him to meet his individual needs. In case of the workflow editor for example, the user needs to store his workflows to use them again. Especially bio-molecular researcher often uses the same analysis redundant in different experimental tests. The more the needs for the specific user group are suited the more active users will use the p-medicine platform.

9.3.7.7 Suitability for Learning

Suitability for learning describes a process where the user has the chance to learn about the software by simple trial and error. The user should have the possibility to use the tools without damaging data, functions or the platform itself. Several equivalent ways might exist to perform an operation and at all times the possibility to cancel the progress should be offered without closing the whole program, e.g. simple step-by-step back functions in the program.

9.3.8 Use Scenario

In general a use scenario describes the user-system interaction with the aim to identify problems related to the interaction, to denote norm conformity and to discover critical incidences and weaknesses of the system.

The use scenario is based on the evaluation of the context scenario in which the minimal functions and requirements of the system were derived from the users implied needs.

During the use scenario the usability engineer is involved as a participatory observatory. The use scenario template is divided into four columns.

The first column shows the task to be performed by the user with the system. This can be subdivided into sub-tasks to describe it in a more detailed way. In the second column the

- Action of the user
- "Thinking Aloud" (Users are asked to say whatever they are looking at, thinking, doing, and feeling, as they go about their task to get also a subjective behaviour and the expectation towards the system)
- Identify critical incidence (behaviour of the user that results in an unsuspected reaction of the system or system failure)

are described.

The third column reports on the reaction of the system in detail (errors, failure messages,..) and the fourth column analysis the single task process in respect to norm conformity or norm violation. An exemplary template for a reporting sheet is given in Table 2.

Task	**User action**	**System action**	**Problems/** **Non Norm Conformity**
Part 1	Action	Reaction	e.g. Non Norm conformity
Part 2			
Part 2.1	Action Critical Incident	Reaction	e.g. Norm violation
Part 2.2	Action "Thinking aloud"	Reaction	No icon for planned task (e.g. no printing function) No task suitability

Table 2: Example of use scenario reporting template

It must be stressed that only negative or failure behaviour in the execution of the task is described in the reporting forms. To avoid general problems the user should have a basic understanding in the use of computers, but he should not be familiar with the software. By doing so initial usage problems of the software can be detected.

Ideally the use scenario is performed in a professional usability laboratory using auditing software. The usability engineer records only the direct interaction of the user with the system, excluding the general behaviour except the "thinking aloud". In p-medicine this effort is hard to realise in respect to the several user groups. The solution will be a usability

satellite session beside the regular p-medicine meetings, where the prototypes can be tested by the end-users of p-medicine and the use scenarios are recorded by the usability engineer.

The user questionnaire (see Appendix B) is protected by copyright by the DAkkS. To not breach the copyright p-medicine has consent to use the questionnaire of the DAkkS. This questionnaire should be distributed to users who have some experience with the software and do not work with the tools for the first time. Users who are not familiar with the software will have initial usage problems which they forget after some time and then the real usage problems occur that are outlast and recurrent. The usability engineer will get a first impression about these usage problems.

9.3.9 Use Case

A use case is a use scenario without participatory observation and detailed documentation of the human-system interaction. Because of the high efforts of man power and costs of the use scenarios, it is obvious that clear defined use cases are powerful elements to evaluate the software, tools and the platform of p-medicine. A use scenario should be an essential precondition for the design of use cases, the specification of functional requirements.

The use case is defined based on the experience of the received design recommendations. A use case defines a goal-oriented set of interactions between external actors and the system under consideration. Actors are parties outside the system that interact with the system [4]. A use case is initiated by a user with a particular goal in mind, and completes successfully when that goal is satisfied. It describes the sequence of interactions between actors and the system necessary to deliver the service that satisfies the goal. It also includes possible variants of this sequence, e.g., alternative sequences that may also satisfy the goal, as well as sequences that may lead to failure to complete the service because of exceptional behaviour, error handling, etc. The system is treated as a “black box”, and the interactions with system, including system responses, are as perceived from outside the system [2].

9.4 Appendix A: Key Questions for Describing and Structuring User Performance in Context

Introduction	1. Describe your daily work in one or two sentences. 2. Which tasks compose your work with the computer (list typical key tasks, which are time-consuming or frequently occurring or very important)? Which of the key tasks should be supported by the software? 3. How work is organised (e.g. in various tasks, as a sequence of tasks, as repetitive single task)?
Assumptions (or pre-condition)	4. What kind of qualification is needed for performing the tasks (for task completion / for using software)? What kind of skills are missing? 5. Who or which event decides what to do? (Who selects your jobs? jobs are performed autonomously, work is divided, data is needed from colleagues or external sources) 6. Which media or devices are necessary (for task completion / for software use)? Which of them are missing, which are desired additionally?

Routine activities (or usual performance)	7. Which working steps are executed? 8. Which working steps are performed repeatedly? (Automated execution desired / necessary)? 9. Which working steps are executed by the software? Can the user control the autonomous process / is control allowed / desired / required? 10. Are several users working in parallel with/ on the same object (e.g. transaction, file, document, data record)? 11. Is there a defined sequence of working steps? If so, how is it composed? (More flexibility needed / desired?) 12. Which are the results / partial results and how are they used / continued? 13. Which kind of feedback does the interviewee get concerning his working results and effects?
Special features during work performance	14. Which kind of interruptions appear? Why do they appear? When do they appear? (Organisational / Social / Technical)? 15. How are mistakes reported back and solved (Organisational / Social / Technical)? 16. Which important special cases have to be considered (Respectively comes up in the user's mind spontaneously; e.g. division of work / collaboration)?
Organisational conditions	17. Which organisational aims are defined for the working tasks? 18. Are there mechanisms to control the efficiency of work? (If yes, which ones? Are they necessary?) 19. Which overview has the user with respect to the overall workflow? 20. Which changes are expected or desired by the user considering the performance of work? Are there any suggestions from the interviewee? Visions! 21. Which results / working steps affect third parties (e.g. customers) directly? And which are the consequences? 22. Which are the stress factors and how are they handled?
Other comments to critical incidents which already occurred	Put examples in here, when the interviewee tells something about critical incidents concerning the software during the interview. Usually such problems analysed within use scenarios.

To derive requirements and test cases. Only this derived material is provided to the project members.

9.5 Appendix B: User Questionnaire

ErgoNorm

User Questionnaire

"Work & Software"

Dear user

With this questionnaire we want to learn more about your personal opinion on the work with your computer used in your daily work. It is only you who can estimate how well the computer supports you in your work situation. It is a matter of finding out which activities are difficult to perform with the software in question, which are the steps that annoy you or leave you puzzled.

Maybe you are no longer aware of the deficiencies of the software during execution of your work because you have become accustomed to them, or maybe you think, "That's just the way it is." The questionnaire helps you to identify and name those weaknesses in the software. Your answers to the questions help to capture deficiencies in quality. The aim is to improve the computer to suit your needs, and therefore ease your work at your workstation.

All data will be collected anonymously, so that none of your statements can be traced back to you personally.

Handling of the questionnaire

Probably you use the computer to execute different and self-contained tasks. Please be sure to fill in the questionnaire with respect to the execution of the following task:

_____

_____

_____

Before you start to fill in the questionnaire please read through all the questions first. You will notice that all questions point to very useful features the computer should have. When completing the questionnaire, it is important for you to think about the task initially described. You should only answer those questions that are important to this task. If you think that a question is not concerned with that task, mark the answer "question does not apply". When filling in you can also indicate deficiencies. If you think them to be very disturbing, please mark the corresponding item.

Please don't start filling in the questionnaire until you have carefully read all the questions. It has proved valuable to complete the questionnaire continuously over a period of time. If you come across problems during your work, insert them in the appropriate position.

Please send the completed questionnaire (on paper or online) to the following address:

Fraunhofer IAIS
Marie-Luise Christ-Neumann
Schloss Birlinghoven
D - 53754 Sankt Augustin
Marie-Luise.Christ-Neumann@iais.fraunhofer.de

Your comments and suggestions are also welcome, even if you have already returned the questionnaire.

Description of task

(Please remember your special task when filling in the questionnaire.)

_____

_____

_____

_____

_____

_____

_____

Suitability for the task

A program is suitable for a task when it is usable for the completion of your special kind of activity. "Usable" means, that all activities you have to perform are supported by the system in an effective and efficient way. The program should be a helpful tool not disturbing you by making your work harder or more complicated in some situations

1.)

Has the program all the features required for your task?	O Yes O No O Question does not apply

If no:

Please indicate the dialogue step which makes you wish, that the program "can do more" than is possible now.

_____
_____
_____
_____

O I feel this is very disturbing

2.)

Do you have to do redundant input actions or dialogue steps?	O Yes O No O Question does not apply

if "yes":

Please indicate redundant input actions and dialogue steps

_____
_____
_____
_____

O I feel this is very disturbing

3.)

Is it possible to facilitate repeated entering of data or text?	O Yes O No O Question does not apply

if "no":

In which situation do you wish that you do not have to enter the same thing again and again?

_____
_____
_____
_____

O I feel this is very disturbing

4.)

Do you think that the effort required to achieve the results of your work is appropriate?	O Yes O No O Question does not apply

if "no":

In which situation have you ever thought "This could be achieved with less effort"?

_____
_____
_____
_____

O I feel this is very disturbing

5.)

Do you think that you have to do task steps which should be done by the program?	O Yes O No O Question does not apply

if "yes":

Please specify these tasks.

_____
_____
_____
_____

O I feel this is very disturbing

6.)

Do you have to enter values and text that the computer could really know?	O Yes O No O Question does not apply

if "yes":

Please describe the situations which make you think e.g.: "The computer should really know by now. Why must I write this once again?"

_____
_____
_____
_____

O I feel this is very disturbing

7.)

Do you have to go some other way or use tricks to achieve your working results as intended?	O Yes O No O Question does not apply

if "yes":

Please describe the situations where you play tricks on the system in order to achieve the intended result.

_____
_____
_____
_____

O I feel this is very disturbing

8.)

Do you get help by the program which actually helps you?	O Yes O No O Question does not apply

if "no":

Specify the situations where the help information has not helped you.

_____
_____
_____
_____

O I feel this is very disturbing

9.)

Does the program fit to your forms and current formats?	O Yes O No O Question does not apply

if "no":

Describe the activity where the program does not fit to your paper forms or formats.

_____
_____
_____
_____

O I feel this is very disturbing

Self-descriptiveness

A program is self-descriptive when you are always informed of what the computer is doing and what is expected as your next input or reaction. This means, among other things, that all of the feedback is comprehensible and that you always know where to enter the next input and you always understand what consequences follows from your input.

10.)

Is the information you need to perform your task structured clearly on the display?	O Yes O No O Question does not apply

if "no":

Please specify the information you need but which is not available on the display "at a glance".

._____
_____
_____
_____

O I feel this is very disturbing

11.)

During your work with the program can you recognize which input is expected from you next?	O Yes O No O Question does not apply

if "no":

Please briefly describe the situation where you are not sure about the next step to do with the program.

_____
_____
_____
_____

O I feel this is very disturbing

12.)

Are system messages always comprehensible to you?	O Yes O No O Question does not apply

if "no":

Identify the situations where you noticed messages you do not understand.

_____
_____
_____
_____

O I feel this is very disturbing

13.)

Are you warned before you perform actions that cannot be undone by the software?	O Yes O No O Question does not apply

if "no":

Please specify the situations where you were not warned by the system.

_____
_____
_____
_____

O I feel this is very disturbing

14.)

Does the help function really help you when a dialogue step or menu item is not entirely clear to you?	O Yes O No O Question does not apply

if "no":

Describe the situations where you do not understand the help text.

_____
_____
_____
_____

O I feel this is very disturbing

15.)

Do you often have to ask colleagues or look up in the manual to continue with your work?	O Yes O No O Question does not apply

if "yes":

Please describe situations where you could not continue without the help of colleagues or a manual.

_____
_____
_____
_____

O I feel this is very disturbing

Controllability

A program is controllable if you can freely determine the sequence of your work steps. If it is required in your work situation, you can interrupt your work with the computer and resume work again without loss of previously attained results.

16.)

Can you execute your work steps in the order which makes most sense to you	O Yes O No O Question does not apply

if "no":

Please describe the work steps where a different order seems to make more sense.

_____
_____
_____
_____

O I feel this is very disturbing

17.)

Is there sometimes a (re)action of the program you do not want at that moment?	O Yes O No O Question does not apply

if "yes":

Please specify the behaviour of the program which occurs unintentionally.

_____
_____
_____
_____

O I feel this is very disturbing

18.)

Can you interrupt a task on demand and resume later, without having to re-enter everything?	O Yes O No O Question does not apply

if "no":

Please explain, in which situation you lost data already entered by a break?

_____
_____
_____
_____

O I feel this is very disturbing

19.)

Can you undo a working step when appropriate for your task performance?	O Yes O No O Question does not apply

if "no":

Please identify the situations where undoing a dialogue step would be advisable.

_____
_____
_____
_____

O I feel this is very disturbing

20.)

Do you feel slowed down by the program in the pace of interaction, e.g. due to long response time?	O Yes O No O Question does not apply

if "yes":

Please describe the situations where you would like to work faster.

_____
_____
_____
_____

O I feel this is very disturbing

Conformity with User Expectations

A program conforms with user expectations when you are not "surprised" by the program. Such a surprise can be, for example, a function being in a totally different position in the menu as you would have expected it, or tasks which cannot be performed as usual.

21.)

Do you find menu items or functions where you think they should be?	O Yes O No O Question does not apply

if "no":

Please specify the specific location in the menu or in another matrix, where the arrangement of information does **not** meet your expectation.

_____
_____
_____
_____

O I feel this is very disturbing

22.)

Are you still sure during waiting periods that the program continues to work?	O Yes O No O Question does not apply

if "no":

Please specify the situations where you are not sure if the program is still working, for example, when the program needs a very long time to store data.

_____
_____
_____
_____

O I feel this is very disturbing

23.)

Are you sometimes surprised at how the program reacts to your input?	O Yes O No O Question does not apply

if "yes":

Describe the situations where you are surprised about the reaction of the system.

_____
_____
_____
_____

O I feel this is very disturbing

Fault tolerance

A program is error-tolerant when the intended result can be achieved despite evident errors in input with either no or minimal corrections by the user. This means, it has to be allowed to mistype or to make a wrong working step without causing a system crash or having to correct the mistake with great effort. In addition, you should be noticed by the program when an error occurs and hints for possible correction should be given to you.

24.)

Do you get correction hints on an incorrect input?	O Yes O No O Question does not apply

if "no":

Please specify situations where you might wish that the program proposes a correct input?

_____
_____
_____
_____

O I feel this is very disturbing

25.)

Can you recover from an incorrect input with minimal effort?	O Yes O No O Question does not apply

if "no":

Please describe briefly the situations where the effort for recovery from of an erroneous input appears to be too high.

_____
_____
_____
_____

O I feel this is very disturbing

26.)

Does the program always work robustly and reliably during the execution of your task?	O Yes O No O Question does not apply

if "yes":

Please describe briefly the situation in which the effort for the correction of incorrect input is not affordable?

_____
_____
_____
_____

O I feel this is very disturbing

Suitability for Individualization

A program is capable of individualization when you are able to adapt the interface software for your individual needs.

27.)

Can you customize the computer so that you can read and work more comfortably?	O Yes O No O Question does not apply

if "no":

Indicate the places where working with the program is difficult for you.

_____
_____
_____
_____

O I feel this is very disturbing

Suitability for Learning

A program is suitable for learning when it allows you to explore the program without having to be afraid of spoiling something. Additionally, you should get relevant information by the system which you need in your opinion to understand the program better.

28.)

Does the program allow you to learn by "trial and error"?	O Yes O No O Question does not apply

if "no":

Please describe the "punishment" which you already got when exploring the program.

_____
_____
_____
_____

O I feel this is very disturbing

The last part of the questionnaire is reserved for your personal observations. There is room for further criticism of the program or for the problems you could not write down in special parts of the questionnaire.

Thank you for your effort and help to improve the p-medicine platform!

9.6 References

[1] ISO 9241–11:1998 Ergonomics of human-system interaction – Part 11 Guidance on usability

[2] ISO/IEC 25000:2005: Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE

[3] ISO 9241–110:2006 (former Part 10) Ergonomics of human-system interaction – Part 110 Dialogue principles

[4] Leitfaden Usability; available on DAkkS website (German's National Accreditation Body (former DATech) only in German language) http://www.dakks.de/sites/default/files/71-SD-2-007_Leitfaden%20Usability%201.3.pdf

[5] ISO 9241–210:2010 (former DIN EN ISO 13407:1999;) Ergonomics of human-system interaction – Part 210 Human-centred design for interactive systems

10 Clinical Decision Support Systems

10.1 Introduction

With medical knowledge expanding at an unprecedented rate – medical literature may be doubling every 22 months and more than two million articles being published annually by over 20,000 biomedical journals, clinicians have a significant amount of information and small amount of time to learn, evaluate and put this information into practical use.

In addition to keeping informed of the new research and therapeutic procedures, clinical guidelines and pathways to follow, oncologist face huge amount of patient data, administration tasks and payers conditions to manage. Moreover, it is reported that in cancer research there is always a delay in implementation of knowledge. Also, cancer patients have the risk of receiving less than optimal care with treatments that are not aligned with evidence-based guidelines. Meanwhile, providers are forced to pay greater attention to quality and efficiency, with an emphasis on using the latest technologies as support. These issues denote that the evidence-practice gap will only get worse without some sort of intervention.

A clinical decision support system (CDSS) is a computerized system that uses case-based reasoning to assist clinicians in assessing disease status, in making a diagnosis, in selecting appropriate therapy or in making other clinical decisions. DCSS can be used as a solution to improve the healthcare quality by bringing the right information to the right person by right interventions at the right time, that is at the point of care and help medical staff make a patient specific informed decision. The main goals of using CDSSs are improved care quality, avoidance of adverse events, and reduced costs.

Decision support systems have been used for different purposes ranging from diagnosis, preventive care and therapy to monitoring and follow-up. The types of CDSSs will be discussed in section 2, and their impact on the quality of care is explained in section 3.

10.2 Types of CDSS

There is no standard taxonomy for CDSS yet. Nevertheless, it is possible to categorize them from different perspectives.

Abbasi et al. [1] state that CDSS can be classified methodologically to knowledge-based and non-knowledge-based systems. The knowledge-based CDSS is further divided into rule-based or evidence-based systems while the non-knowledge-based ones include machine learning and artificial intelligence techniques. There are also hybrid CDSS that combine rule-based, evidence-based and machine learning approaches. Hybrid systems extract the best of all methodologies and provide an optimal solution for clinical decision support systems [2].

Other than the way of reasoning CDSSs have been categorized according to their intended function (prevention, treatment, diagnosis), interaction methods with EHR, CPOE, form of intervention (alerts, reminders, information/reference provision) and the factors related to human–computer interaction [5-13].

Wright et al. [3] have described the evolution of CDS systems based on their architecture. By architectures, they mean the way in which decision support systems interact (or choose not to interact) with other related systems, such as computerized physician order entry (CPOE) and electronic health record (EHR) systems. They have formulated a model with four distinct architectural phases for decision support:

10.2.1 Standalone decision support systems (beginning in 1959)

The early CDSS were not proactive and needed a lot of data entry, even for data already available in digital format. Systems like MYCIN suggesting a best antibiotic therapy, DXplain diagnosis decision support or QMR are among the systems operating stand-alone without any interaction with the patient clinical data available somewhere else in the clinical setting. Since they did not interact with other hospital systems no standardization was needed.

10.2.2 Integrated systems (beginning in 1967)

Integration of CDS systems into other clinical systems like CPOE and EHR started with HELP in 70s. HELP is a system which provides decision support for many clinical areas including laboratory, radiology, pharmacy, ICU and more. Regenstrief Medical record System (RMRS) is another prominent integrated CDS system that uses a large set of rules to make suggestions about care. WizOrder and BICS are other examples of CDS systems capable of integration with the clinical system in use. The integrated systems can be proactive and can e.g. alert clinicians without them seeking assistance.

The major down side of integrated systems, however, is that there is no easy way to share them or reuse their content. Also, integrating decision support into clinical system can create knowledge-management problems. Yet, successful integrated CDSSs are reported to be the ones that were integrated with clinical systems developed by the same system provider.

10.2.3 Standards-based systems (beginning in 1989)

Wright et al. explain that in order to overcome the inability to share decision support content standardization efforts were undertaken. By encoding and representation of knowledge using standards sharing knowledge became feasible. Despite the advantages of these standards, lack of standards terminology prevents them from being widely adopted.

10.2.4 Service models (beginning in 2005)

Service Models have been another effort introduced recently in order to overcome the vocabulary problem. They consider a standard application programming interface (API) that combines the two separate parts of CDSS and clinical information system.

Sharable Active Guideline Environment project (SAGE) [15,16] and SEBASTIAN [17] are the most prominent examples of DSS systems created based on this model. Both of these systems have significant advantages over the standards representation, and both are promising. Each system constrains itself to fully standardizing only one of the two interfaces at the junction between a CDSS and a clinical system, limiting their potential for success. SAGE's API is placed in front of the clinical system while SEBASTIAN places its interface in front of clinical decision support modules. Also, both systems principally look at only one clinical system and one decision support system at a time, although, in the real world, knowledge about a patient (that is stored in a clinical system) and knowledge of medicine (that is stored in a decision support system) can be fragmented across many sites [3].

10.3 Impact of CDSS

A large body of knowledge approves that CDSS can help to improve quality of care [18-24]. The embedding of a CDSS into patient care workflow offers opportunities to reduce medical errors as well as to improve patient safety, to enhance drug selection and dosing, and to improve preventive care [25]. It is less certain whether a CDSS can enhance diagnostic

accuracy, but it is known that CDSS can assist clinicians in reducing some errors and costs and that CDSSs can assist in preventing adverse drug reactions, reducing inappropriate drug dosing, and reinforcing the use of effective prophylactic measures [26-38, 22]. According to a study, hospitals with properly integrated CDS systems have lower costs, fewer complications and lower mortality rates [67]. However, CDSS is not widely in use due to challenges facing the effective implementation of these systems and which are further discussed in section 4.

10.4 CDS Challenges

Even the successful implementations of CDSS have not been widely repeated due to the major challenges that exist in design, development and implementation of CDSS. Some of these challenges have their root in the inherent complexity of the task of decision making while others originate from the integration to the clinical workflow, the technical aspects needed for CDS implementation, the knowledge maintenance, and so much more.

Sittig et al. [4] have categorized the challenges of building CDS using an iterative, consensus-building process. These challenges are aligned with with the strategic objectives recently outlined by an expert panel in a roadmap for national action on CDS [39].

According to them the 10 challenges are related to 3 main categories: improving the effectiveness of CDS interventions, creating new CDS interventions and disseminating existing CDS knowledge and interventions. Sittig et al. mention the following big challenges:

1. Improve the human-computer interface: CDS effectively remind clinicians of things they have truly overlooked and support corrections, or better yet, put key pieces of data and knowledge seamlessly into the context of the workflow or clinical decision-making process, so the right decisions are made in the first place [9,40].
2. Summarize patient-level information: Intelligently summarize a patient's electronically available clinical data, both free-text and coded, and to create one or more brief medical histories, current condition(s), physiologic parameters or current treatment(s) is another CDS challenge. Ultimately, vast amounts of data may be reduced to a summary set of indicators allowing 'at a glance' assessment of patient status [4].
3. Prioritize and filter recommendations to the user: Automatically prioritizing recommendations according to a multi-attribute utility model means combining patient- and provider-specific data to take into account expected mortality or morbidity reduction, patient preferences, cost to the individual or organization, effectiveness of the test or therapy, how the patient might tolerate the recommended intervention, location in the clinician's workflow, insurance coverage, genetic and genomic considerations, clinician's past performance, and other factors [4].
4. Combine recommendations for patients with co-morbidities: Current clinical care guidelines, for the most part, ignore the fact that the majority of elderly patients have multiple co-morbidities that must be addressed by their patient care team [41]. One of several reasons why clinical guidelines are underutilized in practice is because they do not adequately address these co-morbidity issues. Addressing this challenge may require new combinatorial, logical, or semantic approaches to combining and cross-checking recommendations from two or more guidelines [4].
5. Use free-text information to drive clinical decision support: At least 50% of the clinical information describing a patient's current condition and stage of therapy resides in the free-text portions of the HER [42]. Automatically extracting information from free-text documents and structuring it appropriately for the use of CDS is a challenging task that should be addressed in order to fully benefit from the CDS.
6. Prioritize CDS content development and implementation: Deciding which content to develop or implement first (e.g. interventions to improve patient safety, chronic

disease management, or preventive health interventions) according to pertinent factors like value to patients, cost to the health care system, availability of reliable data, difficulty of implementation, acceptability to clinicians and patients, and national interests and overall health care value is another challenge of CDS development [4].

7. Mine large clinical databases to create new CDS: There is always a large amount of new guidelines, CDS interventions and knowledge that are produced but not yet compiled and made ready for the use of CDSS. Some methods for mining large clinical knowledge repositories are needed to be developed so that clinicians can have access to latest knowledge out there. However, this development has another aspect that needs to be attended and that is the legal issues for accessing the databases as well as privacy issues [4,43].
8. Disseminate best practices in CDS design, development, and implementation: Common success factors can be derived from the best practices of CDSS. This kind of knowledge is frequently not readily available to other organizations seeking to develop CDS programs [44,45]. To accomplish this, a consensus on a standard taxonomy of clinical decision support interventions and outcomes that would allow us accurately describe the best practices is needed as well as comparison of outcomes between implementations of different systems and across organizations [4,30].
9. Create an architecture for sharing executable CDS modules and services: The goal is to create a set of standards-based interfaces to externally maintained clinical decision support services that any EHR could “subscribe to”, in such a way that healthcare organizations and practices can implement new state of the art clinical decision support interventions with little or no extra effort on their part [46,39]. These knowledge modules can be loaded into a clinical information system, or to execute as a remote service, with the local clinical system invoking them over a network according to a standardized interface [4,47].
10. Create internet-accessible clinical decision support repositories: The challenge is to build one or more internet-accessible repositories of high quality, evidence-based, tested, clinical decision support interventions and services that could be easily downloaded, maintained, locally modified, installed, and used on any Certification Commission for Healthcare Information Technology (CCHIT)-certified HER product [48], for instance using the architecture described in Challenge 9 [4]. The challenges mentioned above are not all of the challenges residing in the process of building CDSS but certainly are among the central ones.

10.5 CDS Standards

There have been some efforts to form standards for CDS functionalities, information exchange, knowledge representation and quality of data [49-52]. However, still approximately half of the costs to develop the CDS involves clinician time in selection and design of content [53] since there are multiple standards to choose from and no consensus-based single standard, for example, for the specific evidence-based guidelines to be used in the CDS.

Moreover, even with the use of standards some relevant extent of adaptation work is needed. Miller et al. believe that any CDS implementation requires some degree of customization, ranging from configuring the CDS for local needs and sometimes paying for added features that are needed at their site [9].

By encoding, modelling and representation of knowledge using standards, sharing knowledge became feasible. Despite the advantages of these standards, lack of standards terminology prevents them from being widely adopted. Due to these terminology variances and the different standards for data expression like normal laboratory values, medication formularies, or norms for processes of care at different sites and within different CDS [54],

still some extra synchronization effort is needed even when commercially available knowledge sources and modules are used.

Arden Syntax is one of the knowledge modelling standards being developed based on HELP [59] and RMRS [60]. It is a grammar for representing and processing medical recommendations as Medical Logic Modules (MLMs) that can only encode event driven and patient-specific rules. Thus, it cannot be used for the point-of-care reference or information retrieval support. Moreover, it does not define a standard vocabulary and therefore, the rules cannot be used in other systems without modification [61].

Arden Syntax (2.6) is accepted as a standard by the American National Standards Institute (ANSI) and Health Level 7 (HL7) [55]. Other standards such as Guideline Interchange Format (GLIF) [58] and GELLO have also been developed. GLIF created by InterMed Consortium [56], is more complex than Arden Syntax which mostly is useful for alerts and reminders. It focuses on multi-part guidelines, including complex clinical pathways and defines an ontology for representing guidelines and one for medical data and concepts [61].

OpenClinical.org [57] provides an overview over the guideline and knowledge modelling standards including GELLO, GLIF and Arden Syntax.

10.6 Characteristics of a Successful CDSS

According to Friedlin et al. [24,62] several practical factors contribute to the success of a CDSS. Kawamoto et al. [24] have also performed a systematic review on the literature and came up with the some design characteristics of a successful CDS. The following factors are the summary of the points mentioned by these studies:

1. The potential impact of CDS on clinical workflow should be considered. Automatic CDS interventions which also merge into the workflow (for example, embedded within the CPOE and EMR) are most used by the clinicians.
2. Creating an intuitive and configurable user interface can be a success factor.
3. Delivering decision support in real time at the point of care has the highest impact. There, CDS interventions should provide information and decision making.
4. Providing actionable alerts/reminders/recommendations that are succinct and relevant to patient care.
5. CDS systems that generate suggestion for action are more effective than the ones that perform diagnosis and assessment.

Osheroff et al. [46] also identified three key elements for fully realizing a CDSS' potential:

1. The best available clinical knowledge is well organized, accessible to clinicians, and encapsulated in a format facilitating effective support for the decision making process.
2. A useful CDSS is extensively adopted, and generates significant clinical value that contributes financial and operational benefits to its stakeholders.
3. Both clinical interventions and knowledge undergo constant improvement via user feedback, experience and data analysis that are easy to aggregate, assess, apply.

As mentioned, knowledge maintenance plays an important part in the effectiveness and success of CDSS. With the rate that medical knowledge is growing, the introduction of new medication and treatments, and the expansion of evidence-based guidelines, keeping the knowledge sources updated is a major task. In addition to the maintenance of external knowledge, accuracy of data provided for CDS usage by the clinical systems like the EMR is also of utmost importance. Studies found out that outdated clinical records of patients are the reason of generating most alerts that are overridden by the clinicians [63,64].

In order to fulfil this task, some approaches have been in favour of building a home-grown knowledge management process which needs significant resources to maintain, while others like the Clinical Decision Support Consortium (CDSC) have opted for a web-based knowledge repository using SOA and including various types of content from guidelines to "plug and play" CDS. [47] Another way is to utilize commercially available knowledge sources. They have to be purchased considering their sources of knowledge and the frequency of their updates [65,66]. All in all, knowledge management is an expensive task and it must be noted that even when a commercial knowledge source or a central external one is used, still some local effort is needed to integrate and adapt it for the local site.

Workflow integration is a key success factor for the CDSS. "Do CDS with users not to them." Osheroff proposes. The workflow integration of CDS is a main issue that must be addressed with care, both by designers of the CDS and people who implement it. It is best to involve the clinicians in the entire process of CDS design and implementation. Workflow integration should be attended in the early phases of CDS requirement analysis to clarify how the CDS should fit into the workflow and what possible modifications are needed before implementing the CDS into the current workflow. However, the workflow changes should only be performed in case of the need for process improvement not only for the sake of specific CDS, since there is a possibility that the CDS itself was not optimally designed [68].

Considering the cost of treatments, payers and insurance policies while suggesting the treatments is becoming a necessary part of CDSS. Clinicians want to make sure they get reimbursed while payers want to ensure that the treatment and procedures were evidence-based and the tests and medications were necessary.

Patient empowerment has recently gained much attention as a patient-centric approach to improve care. The concept of patient empowerment is described as a "social process of recognizing, promoting, and enhancing people's abilities to meet their own needs, solve their own problems, and mobilize necessary resources to take control of their own lives" [69]. In other words, patient empowerment is a process of helping people to assert control over factors that affect their health.

Patient empowerment and education can be an effective and important aspect of a sophisticated CDS. CDSS can provide a basis to transfer the important information to patients so that they know their condition and how they can best manage their situation. However, more research is needed to better evaluate the real effect of its inclusion in CDSS.

10.7 CDS Systems and Tools for Oncology

In this section the CDS systems related to oncology domain are introduced. Some of these systems are commercially available, while others are free to use. Also, some CDS systems related to oncology discussed in literature are not available for general use are discussed.

10.7.1 Evidence-based Treatment Intelligence (eviti)

eviti (evidence based treatment intelligence) (http://www.driveideas.com/ita/html/about-eviti), launched in October 2010, is a web-based oncology decision support platform that equips providers with real-time access to evidence-based intelligence at the point of prescribing a treatment. It provides comprehensive, verified treatment options for the physicians just when they need to determine the most suitable patient-specific cancer treatment.

The treatment options offered to oncologists by eviti originate from the knowledge sources that are used by eviti. It benefits from a comprehensive digital library of evidence-based oncology treatment regimens available, compiled from FDA, NCI, ASCO, ASH, NCCN, and others, taking into account costs, outcomes, efficacy and toxicities. These sources are maintained by a team of expert oncologists, medical advisory board and medical informatics

professionals. eviti's library is also enriched by the reports and journal publications, including JCO, JNCI, Lancet, NEJM and JAMA.

In addition to offering nearly 1,000 nationally accepted treatment regimens for more than 120 cancer types across all modalities, it will soon include the leading cancer trial search engine TrialCheck and its thousands of clinical trials. This makes eviti a CDS system that brings clinical trials and evidence based medicine together in one CDS tool.

Moreover, eviti aligns providers with the payers and facilitates timely reimbursement from insurance companies. Using a unique eviti code containing all the information like CPT codes, ICD9 codes, J code and treatment plan itself, payers are aware that the right treatment was selected for a specific patient and oncologists are offered an automatic pre-certification by the payers. This streamlines the reimbursement process and ensures that the correct treatment is being prescribed from the beginning using an up-to-date, comprehensive evidence-based library and that the appropriate procedure will be paid for by the insurance.

10.7.2 Proventys CDS Oncology

Proventys CDS Oncology (http://proventys.com/CDSTrial) is a web-based solution to support the decision-making needs of oncologists by providing real-time access to protocols and guidelines for patient specific treatment decisions. It allows clinicians to navigate the NCCN Clinical Practice Guidelines in Oncology (http://www.nccn.org/professionals/physician_gls/f_guidelines.asp) efficiently within the clinical workflow, and create personalized care plans based on patient-specific data.

The first release of CDS Oncology includes NCCN guidelines, covering four major cancer types including breast, colon, non-small cell lung and hodgkin lymphoma. Future releases of CDS Oncology would also include clinical trial searching functionality.

The core edition of CDS Oncology provides all key decision support features of the product to meet the needs of any oncology practice, even those with an existing EMR system. The Expanded Edition of the includes all of the elements of the Core Edition, but also adds CPOE capabilities for practices that do not have, or are unsatisfied with their current solutions.

CDS Oncology serves as the point-of-care platform for the AlignQI pathways program. As a tool for documenting clinical decisions, AlignQI highlights any clinical pathways or payer coverage policies that apply to a particular patient in health plan. AlignQI can be customized with payer- or practice-specific pathways to provide clinicians with a full set of evidence-based care choices derived from the NCCN Guidelines.

10.7.3 Adjuvant Online

The purpose of Adjuvant (http://www.adjuvantonline.com) is to help health professionals and patients with early cancer to discuss risks and benefits of getting additional therapy (adjuvant therapy: usually chemotherapy, hormone therapy or both) after surgery and make estimates of negative outcome risk (cancer related mortality or relapse) without systemic adjuvant therapy, estimates of the reduction of the risks afforded by therapy and risks of side effects.

Adjuvant integrates patient related information (age, sex, and comorbidity) to make estimates of non-breast cancer related mortality, tumour related information (nodal status, tumour size, histologic grade, oestrogen receptor status, and histologic subtype) to project breast cancer related mortality and relapse, and tumour and patient related information (patient age, oestrogen receptor status, and treatment type) to make estimates of treatment efficacy.

10.7.4 MATE

Hospital-based cancer care in the UK is typically managed via multidisciplinary team meetings (MTM). MATE supports evidence-based decision making in MDT meetings for breast cancer by evaluating a patient's clinical facts. It provides patient-specific decision support including patient-specific clinical assessments (diagnosis, prognosis, etc.) and management recommendations (e.g. investigations, treatment adjuvant therapies).

The MATE knowledge base is based on high quality evidence-based sources from the breast cancer literature. These include clinical practice guidelines, systematic reviews and meta-analyses, randomised controlled trials and other published evidence. The knowledge base takes account of 16 different guidelines and 220 individual recommendations. It includes recommendations for 10 different types of decisions covering 32 types of interventions.

Various published breast cancer prognostication algorithms such as the Nottingham Prognostic Index, Adjuvant online!, the Van Nuys Prognostic Index and the MSKCC nomogram have also been included in the knowledge base to provide risk calculations.

The scope of the knowledge base includes diagnosis, staging, treatment and surveillance decisions covering benign breast conditions as well as operable breast cancers. It is currently being extended to cover recurrent and metastatic cancers and genetic risk assessment, but in order to limit its scale for practical reasons in the first instance, guidelines for metastatic, locally advanced and recurrent breast cancers have been excluded.

PROforma provides the knowledge representation and decision support functionality of MATE, originally developed by Cancer Research UK. PROforma is a logic-based formalism that provides a task/class hierarchy for representing the system's clinical knowledge, and an argumentation-based decision framework to output the system's recommendations.

MATE has its own standalone clinical database for secure storage of patient records. The system provides standard data capture and other database services such as form-based data recording, automatic creation of a clinical summary for each patient, and prompts and reminders. It supports dynamic and concurrent audit. The system evaluates the quality of decision-making and compliance with national guidelines. MATE automatically identifies and flags patients found to be suitable for entry into ongoing national and local clinical trials.

10.7.5 Dukes B Adjuvant Chemotherapy Risk Prognostication Tool

More than a third of all bowel cancer patients are offered postoperative chemotherapy to reduce the risk of recurrence and prolong survival. At present, however, many patients are exposed to chemotherapy with the risk of side effects but negligible benefit. This is particularly true for patients at low risk of recurrence (Dukes B).

New genomic tests to predict prognosis have recently been developed such as the microarray gene expression profiling for Dukes B patients by chip provider ALMAC Diagnostics. This has been shown to provide good overall prognostic accuracy in a multivariate analysis. But since it is currently unclear how microarray gene expression profiling can be optimally combined with conventional evidence to give accurate risk estimates a CDSS combining standard clinicopathological information with gene expression profile data has been developed (http://credo.cossac.org/applications2/dukes.html).

The system's knowledge base uses the most recent guidelines and textbooks and expert interpretation of the current knowledge regarding the value of adjuvant chemotherapy in Dukes B stage patients. It also uses literature from the fields of risk representation and communication, informed decision making, risk/benefit calculation models, colorectal cancer datasets. The data model has been developed to be fully compliant with the National Bowel Cancer Audit Programme (NBOCAP). This enables data collection at the national level.

SNOMED has also been used when the NBOCAP dataset was found insufficient. The PROforma language has been used to model the workflow as a series of tasks.

The decision process has been modelled in three steps that is believed to help the patient make a more informed decision:

1. Calculation and rationale for the risk of recurrence
2. Estimation of the potential benefit from adjuvant chemotherapy
3. Choice of regime taking into account the patient's preferences.

An important aspect of the architecture of the system's software framework is its generic design. The Dukes B colorectal service, designed to deliver decision support in complex problems such as the choice of adjuvant treatment, is a specialisation of the platform. The Dukes B framework has seen some of its technology used in the breast MDT tool, MATE.

10.7.6 MedSolutions' Oncology Management Program

MedSolutions' Oncology Management program (http://www.medsolutions.com/services/intelligent_util/oncology/solution.html) is designed to deliver savings by managing oncology drugs, diagnostic imaging and radiation therapy. The program uses evidence-based preauthorization, predictive intelligence technology, and nationally recognized guidelines to reduce delays in care, detect inappropriate studies and therapies, and reward clinically-accurate providers with expedited approvals. The program aims to achieve cost savings and quality improvement for patients, payers and providers.

MedSolutions' oncology management program uses evidence-based guidelines from national physician-led organizations such as the American Society for Therapeutic Radiology and Oncology, the NCCN, the American College of Radiology, and the U.S. FDA, as well as highly trained oncology physicians, nurses and clinicians to review cancer therapies and treatment pathways. Additionally, MedSolutions' program offers access to trained oncology case management professionals. These case managers work closely with patients and their treatment teams to monitor adherence to and effectiveness of ongoing therapies and proactively manage chemotherapy side effects.

10.7.7 Arezzo Optimal Pathways

The Arezzo Optimal Pathways Application (http://www.infermed.com/index.php/arezzo) is a web-based solution that takes the best clinical evidence available, matches it with an individual patient's condition and provides a recommended care pathway specifically for that patient. Arezzo OPA generates decision options, takes into account predictive risk factors and enables tailored prescribing and complete care pathway planning.

Arezzo is a workflow and inference rules engine used for the design, creation and execution of clinical pathways and guidelines, order sets and patient care protocols. During consultation Arezzo instantaneously customizes guidelines, protocols, and care pathways using the individual patient's data together with the rules of the guideline, so that only options relevant for that patient's condition (and all such options) are returned to the clinician, together with pros and cons for each. It provides active decision support as part of normal workflow, monitoring clinical situations and providing recommendations from relevant clinical pathways. It supplies patient-specific recommendations, with arguments for and against, in real-time, while the patient is being seen by the clinician. There may be sound clinical reasons for not following guideline recommendations. In such cases Arezzo can request documentation of these reasons, enabling detailed clinical audit.

Arezzo is being used in an integrated solution with electronic medical record (EMR) systems. The Arezzo-based application integrates with the EMR system to allow GP access to best practice guidelines, displaying the results and recommendations within their own system.

10.7.8 CREDO Applications for Breast and Colon Cancer

CREDO (http://credo.cossac.org) aims to develop and trial a comprehensive suite of computer services to support care in various clinical domains, particularly in cancer. The initial focus of CREDO is on developing and evaluating applications for the complete breast cancer pathway from first presentation and diagnosis through treatment and follow-up and facilitating communication and coordination across the multidisciplinary team. CREDO will also cover other types of cancer (an application to support for colon cancer is under development) and develop applications in other clinical domains.

Cancer care services to be offered by CREDO applications include:

- GP referrals
- Genetic risk assessment
- Therapy planning
- Support for recruitment into trials
- Patient-tailored information
- Support for multidisciplinary treatment.

Inadequate co-ordination between primary care (where initial detection, primary risk assessment and sometimes follow-up of a cancer occurs) and secondary care (where most of the treatment takes place) can increase patients' feelings of uncertainty. One of the reasons for this is that general practitioners and breast specialists have different roles and view the process from different perspectives. However, from the patient's point of view, the entire process, from her first visit to the GP through to follow-up, is a single journey.

CREDO aims to enhance shared care by acknowledging the different goals and roles of the different stakeholders (primary and secondary care clinicians, nurses, patients …) involved in a complex pathway. Credo systems will be designed to allow all stakeholders to access the same underlying evidence-based computerized care process, but provide different representations of it to meet their individual needs.

The first clinical-strength CREDO application is MATE described in the previous section, which is in routine clinical use at the Royal Free Hospital NHS Trust in London.

10.7.9 Management of Pediatric Asthma Exacerbation (MET3-AE)

MET3-AE is a CDSS that provides a comprehensive decision support for data collection, diagnosis formulation, treatment planning and evidence retrieval. It is aiming to help emergency physicians to collect data about the patient, diagnose the severity of exacerbation, and plan a treatment based on evidence. The system is developed using ontology-driven [70] and multi-agent methodologies [71,72].

MET3-AE supports the HL7 standard to interact and exchange data with HIS. Mirth Connect has been used to enable communication between the MET3-AE and HIS. MET3-AE is implemented using open-source software JADE (Java Agent DEvelopment Framework). The repositories with abstract models (the model repository) and with patient data (the data repository) were created with Protégé [71].

MET3-AE belongs to the latest generation of CDSSs [61] that depending on desired functionality are designed according to service-oriented principles and implemented as web services for example the SEBASTIAN system [71].

10.7.10 Knowledge ON ONcology through Ontology (KON^3)

KON^3 is a joint effort among universities, companies and regional government agencies to build a CDSS in “breast cancer” based on National Comprehensive Cancer Network (NCCN) guidelines and semantic information representation in oncology environment. In order to do this ontology for patient data, guidelines and an oncology taxonomy has been developed. A set of rules was written to build the specified guidelines and the CDS in order to get recommendations and help clinicians in their choices [73].

The architecture includes 4 layers:

- Distributed data layer representing the patient data in HL7/CDA2
- Semantic layer: representing patient ontology, oncology taxonomy, guideline model
- Knowledge service layer: extracting knowledge from ontology and run the rules to make inferences, creating guidelines and getting recommendations, Workflow system: an alternative to the single guideline
- Interface layer: a web interface and configuration interface. The tools used for building KON^3 are Protégé [74] and its plug-ins SWRLTab and Jess [75,73].

10.8 References

[1] Abassi M, Kashiyarndi S (2011) Clinical decision support systems: a discussion on different methologies used in health care.

[2] Demmer-Fushman D, Lin J (2007) Answering clinical question with knowledge-based and statistical techniques. Association for Computational Linguistics, pp. 63-103

[3] Wright A, Sittig D (2008) A framework and model for evaluating clinical decision support architectures. J Biomed Inform. 41 (6): 982-90

[4] Sittig D, Wright A, Osheroff J, Middleton B, Teich J, Ash J, Campbell E, Bates D (2008) Grand challenges in clinical decision support. J Biomed Inform. 41 (2): 387-92.

[5] Miller R (1994) Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc- 1(1):8–27

[6] Teich J, Wrinn M (2000) Clinical decision support systems come of age. MD Comput. 17(1):43–46

[7] Wright A, Goldberg H, Hongsermeier T, Middleton B (2007) A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. J Am Med Inform Assoc

[8] Sim I, Gorman P, Greenes R, et al. (2001) Clinical decision support systems for the practice of evidencebased medicine. J Am Med Inform Asso. 8(6):527–534

[9] Miller R, Waitman L, Chen S, Rosenbloom S (2005) The anatomy of decision support during inpatient care provider order entry (CPOE): empirical observations from a decade of CPOE experience at Vanderbilt. J Biomed Inform. 38(6):469–485

[10] Osheroff J, Pifer E, Teich J, Sittig D, Jenders R (2005) Improving outcomes with clinical decision support: an implementer's guide. Chicago, IL: HIMSS

[11] Wang J, Shabot M, Duncan R, Polaschek J, Jones D (2002) A clinical rules taxonomy for the implementation of a computerized physician order entry (CPOE) system. Proc. AMIA Annual Symposium. pp. 860–863

[12] Falas T, Papadopoulos G, Stafylopatis A (2003) A review of decision support systems in telecare. J Med Syst. 27(4):347–356

[13] Miller P (1986). Critiquing: a different approach to expert computer advice in medicine. Med Inform (Lond). 11(1):29–38

[14] Kim Y, Cho Y (1995) Correlation of pain severity with thermography. Engineering in medicine and biology society. IEEE. pp. 1699-1700

[15] Parker C, Rocha R, Campbell J, Tu S, Huff S (2004) Detailed clinical models for sharable, executable guidelines. Proc. Medinfo. 11(Pt 1):145–148

[16] Ram P, Berg D, Tu S, et al. (2004) Executing clinical practice guidelines using the SAGE execution engine. Proc. Medinfo. 11(Pt 1):251–255

[17] Kawamoto K, Lobach D (2005) Design, implementation, use, and preliminary evaluation of SEBASTIAN, a standards-based web service for clinical decision support. Proc AMIA Symp

[18] Balas E, Weingarten S, Garb C, Blumenthal D, Boren S, Brown G (2000) Improving preventive care by prompting physicians. Arch Intern Med 160(3):301–308

[19] Cabana M, Rand C, Powe N, et al. (1999) Why don't physicians follow clinical practice guidelines? A framework for improvement. Jama. 282(15):1458–1465

[20] Garg A, Adhikari N, McDonald H, et al. (2005) Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama. 293(10): 1223–1238

[21] Grimshaw J, Russell I (1993) Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet. 342(8883):1317–1322

[22] Hunt D, Haynes R, Hanna S, Smith K (1998) Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. Jama. 280(15): 1339–1346

[23] Johnston M, Langton K, Haynes R, Mathieu A (1994) Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. Ann Intern Med. 120(2):135–142

[24] Kawamoto K, Houlihan C, Balas E, Lobach D (2005) Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 330(7494):765

[25] Jao C, Hier D (2010) Clinical decision support systems: an effective pathway to reduce medical errors and improve patient safety. InTech.

[26] ActiveHealth Management (2005) Computerized decision support system reduces medical errors, cuts costs. http://www.medicalnewstoday.com/articles/19959.php

[27] Bakken S, Currie L, Lee N, Roberts W, Collins S, Cimino J (2008) Integrating evidence into clinical information systems for nursing decision support. Int J Med Inform 77(6): 413-420

[28] Bates D, Cohen M, Leape L, Overhage J, Shabo M, Sheridan T (2001) Reducing the frequency of errors in medicine using information technology. J Am Med Inform Assoc 8(4): 299-308

[29] Bates D, Gawande A (2003) Improving safety with information technology. N Engl J Med 348(25): 2526-2534

[30] Bates D, Kuperman G, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B (2003) Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 10(6): 523-530

[31] Bates D, Leape L, Cullen D, Laird N, Petersen L, Teich, Burdick E, Hickey M, Kleefield S, Shea B, Vliet M, Seger D (1998) Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. JAMA 280(15): 1311-1316

[32] Kaushal R, Barker K, Bates D (2001). How can information technology improve patient safety and reduce medication errors in children's health care? Arch Pediatr Adolesc Med 155(9): 1002-1007

[33] Kaushal R, Bates D, Landrigan C, McKenna K, Clapp M, Federico F, Goldmann D (2001) Medication errors and adverse drug events in pediatric inpatients. JAMA 285(16): 2114-2120

[34] Kaushal R, Shojania K, Bates D (2003) Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 163(12): 1409-1416

[35] ActiveHealth Management (2005) Computerized decision support system reduces medical errors, cuts costs. Retrieved July 31, 2009.

[36] Berner E (2007) Clinical decision support systems: theory and practice. New York, Springer

[37] Chaudhry B (2008) Computerized clinical decision support: will it transform healthcare? J Gen Intern Med 23 Suppl 1: 85-87

[38] Trowbridge R, Weingarten S (2001) Clinical decision support systems. Making health care safer: A critical analysis of patient safety practice. Shojania K, Duncan B, McDonald K, Wachter R (eds.) Rockville, MD, Agency for Healthcare Research and Quality

[39] Kawamoto K, Lobach D (2007) Proposal for fulfilling strategic objectives of the U.S. Roadmap for national action on clinical decision support through a service-oriented architecture leveraging HL7 services. J Am Med Inform Assoc. 14(2):146–55

[40] Berner E, Moss J (2005) Informatics challenges for the impending patient information explosion. J Am Med Inform Assoc. 12(6):614–7

[41] Boyd C, Darer J, Boult C, Fried L, Boult L, Wu A (2005). Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. 294(6):716–24

[42] Hicks J (2003) The potential of claims data to support the measurement of health care quality. PhD diss. RAND Graduate School

[43] Safran C, Bloomrosen M, Hammond W, Labkoff S, Markel-Fox S, Tang P, Detmer D (2007) Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc. 14(1):1–9

[44] Improving Quality and Reducing Cost with Electronic Health Records: Case Studies from the Nicholas E. Davies Awards. Healthcare Information and Management Systems Society (HIMSS) (2007)

[45] Ash J, Sittig D, Campbell E, Guappone K, Dykstra R (2007) Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc. in press

[46] Osheroff J, Teich J, Middleton B, Steen E, Wright A, Detmer D (2007) A roadmap for national action on clinical decision support. J Am Med Inform Assoc. 14(2):141–5

[47] Wright A (2007) SANDS: A service-oriented architecture for clinical decision support in a national health information network. Oregon Health & Science University; Portland, OR. Ph.D. Dissertation

[48] Certification Commission for Healthcare Information Technology Certified Ambulatory Electronic Health Record (EHR) Products. http://www.cchit.org/certified/products.htm

[49] Healthcare Information Technology Standards Panel (HITSP). Healthcare Information Technology Standards Panel; 2009. Available at: www.hitsp.org

[50] Draft roadmap for expansion of health IT certification. Certification Commission for Healthcare Information Technology; 2009 January 14. Available at: http://www.cchit.org/files/Expansion/CCHITExp ansionRoadmapDRAFT20090115.pdf

[51] Metzger J, Welebob E, Turisco F, et al. (2008) The Leapfrog Group's CPOE standard and evaluation tool. Patient Safety and Quality Healthcare;. http://www.psqh.com/julaug08/cpoe html

[52] Clinical decision support work group. Health Level Seven, Inc. (2009) http://www hl7.org

[53] Field T, Rochon P, Lee M, et al. (2008) Costs associated with developing and implementing a computerized clinical decision support system for medication dosing for patients with renal insufficiency in the long-term care setting. J Am Med Inform Assoc 15(4):466-72.

[54] Dick R, Steen E, Detmer D (1997) The computer-based patient record: An essential technology for health care, revised edition. Washington, DC: The National Academies Press

[55] Health Level 7 (2005) Patient Evaluation Service Draft Standard. Ann Arbor, MI

[56] Peleg M, Boxwala A, Tu S, et al (2004) The InterMed approach to sharable computer-interpretable guidelines: a review. J Am Med Inform Assoc 2004 11(1):1–10

[57] OpenClinical (2006) Summaries of guideline representation http://www.openclinical.org/gmmsummaries.html

[58] Ohno-Machado L, Gennari J, Murphy S, et al. (1998) The guideline interchange format: a model for representing guidelines. J Am Med Inform Assoc 5(4):357–372

[59] Kuperman G, Gardner R, Pryor T (1991) HELP: a dynamic hospital information system. New York: Springer-Verlag

[60] McDonald C, Murray R, Jeris D, Bhargava B, Seeger J, Blevins L (1977) A computer-based record and clinical monitoring system for ambulatory care. Am J Public Health. 67(3):240–245

[61] Wright A, Sittig D (2008) A four-phase model of the evolution of clinical decision support architectures. Int J Med Inf. 77: 641-649

[62] Friedlin J, Dexter P, Overhage J (2007) Details of a successful clinical decision support system. AMIA Annu Symp Proc: 254-258

[63] Weingart S, Toth M, Sands D, et al. (2003) Physicians' decisions to override computerized drug alerts in primary care. Arch Intern Med. 163(21):2625-31

[64] Hsieh T, Kuperman G, Jaggi T, et al. (2004) Characteristics and consequences of drug allergy alert overrides in a computerized physician order entry system. J Am Med Inform Assoc. 11(6):482-91.

[65] Berner E (2002) Ethical and legal issues in the use of clinical decision support systems. J Healthc Inf Manag. 16(4):34-7

[66] Berner E (2008) Ethical and legal issues in the use of health information technology to improve patient safety. HEC Forum. 20(3):243-58.

[67] Amarasingham R, Plantinga L, Diener-West M, Gaskin D, Powe N (2009) Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med. 169(2):108-14

[68] Stead W, Lin HS (2009) Computational technology for effective health care: immediate steps and strategic directions

[69] Jones P, Meleis A (1993) Health is empowerment. ANS Adv Nurs Sci. 15:1-14

[70] Farion K, Michalowski W, Wilk S, O'Sullivan D, Rubin S, Weiss D (2009) Clinical decision support system for point of care use: ontology driven design and software implementation. Methods Inf Med 2009: 48 (4): 381-90

[71] Wilk S, Michalowski W, Farion K, Sayyad Shirabad J (2010) MET3-AE system to support management of pediatric asthma exacerbation in the emergency department. Stud Health Technol Inform, vol. 160, no. Pt 2, pp. 841-5

[72] Wilk S, Michalowski W, O'Sullivan D, Farion K, Matwin S (2008) Engineering of a clinical decision support framework for the point of care use. AMIA Annu Symp Proc. pp. 814-8

[73] Ceccarelli M, Donatiello A, Vitale D (2008) KON^ 3: a clinical decision support system, in oncology environment, based on knowledge management, IEEE. pp. 206-210

[74] Siddiqi J, Akhgar B, Gruzdz A, Zaefarian G, Ihnatowicz A (2009) Automated diagnosis system to support colon cancer treatment: MATCH. Fifth International Conference on Information Technology: New Generations. pp. 201-205

[75] Tu S, Campbell J, Musen M (2004) SAGE guideline modeling: motivation and methodology. Kaiser K, Miksch S, Tu S (eds.) In: Computer-based support for clinical guidelines and protocols: Proc symposium on computerized guidelines and protocols. IOS Press. pp. 167-171

11 Conclusion

The intention of this deliverable was to provide the reader with a comprehensive overview of the current state of the art of the methodologies and technologies that are applied at the moment within the various areas of research that constitute both the foundations as well as the targets of the p-medicine project. As can be clearly observed when going through all of the sections of this deliverable, a tremendous variety of highly challenging topics exist: It is the respective task of the different partners to take them on and to tackle them within the project and finally to transform and integrate the gained results into one single running system platform – as this is the defined final goal of this project.

Therefore the given overview of the state-of-the-art is essential in that it supplies the required fundamental understanding of what the present status to be found within the different fields constituting the overall p-medicine project.

This understanding in turn is necessarily combined with the results gained in task 2.2 where the needs and requirements of the users are analysis in a Scenario based fashion and subsequently provided in combination to both the architects of the envisaged system as well as its implementers: The deliverables D2.1 together with D2.2 describe in-depth how the researchers (and also the technology partners) currently work in their respective environments, what tools, technologies and methodologies exists and how those are actually applied – and then describe where there are deficiencies and how a new system with novel components can better fulfil their respective day-to-day needs.

Appendix - Abbreviations and Acronyms

ACGT	Advancing Clinico-Genomic Trials on Cancer
AD	Architectural Description
ADaM	Analysis Data Model
ADM	Architecture Development Model
ANSI	American National Standards Institute
ASL	Arterial Spin Labelling
AWS	Amazon Web Services
BAC	Bacterial Artificial Chromosomes
BPEL	Business Process Execution Language
CA	Certificate Authority
CDASH	Clinical Data Acquisition Standards Harmonization
CDE	Common Data Element
CDISC	Clinical Data Interchange Standards Consortium
CDS	Clinical Decision Support
CDMS	Clinical Data Management System
CDSC	Clinical Decision Support Consortium
CDSS	Clinical Decision Support System
CGI	Common Gateway Interface
CITMAS	Clinical International Trial Management System
CRF	Case Report Form
CRC	Clinical Research Centre
CRO	Clinical/Contract Research Organisation
CTA	Clinical Trial Authorisation
CTU	Clinical Trials Unit
DAWG	Data Access Working Group
DNA	Deoxyribonucleic acid

DRLS	Drug Registration and Listing System
DRMAA	Distributed Resource Management Application API
DSL	Domain Specific Language
DSMB	Data Safety and Monitoring Board
EBI	European Bioinformatics Institute
ECRIN	European Clinical Research Infrastructure Network
EC	European Commission
EDC	Electronic Data Capture
EGI	European Grid Infrastructure
EHR	Electronic Health Record
EMA	European Medicines Agency
eSDI	eSource Data Interchange
ESTRI	Electronic Standards for the Transfer of Regulatory Information
EVS	Enterprise Vocabulary Services
EWG	Expert Working Group
FDA	Food and Drug Administration
FMA	Foundational Model of Anatomy
GAARDS	Grid Authentication and Authorization with Reliably Distributed Services
GAE	Google App Engine
GAS	Gridge Authorization Service
GBM	Glioblastoma Multiform
GCP	Good Clinical Practice
GEDM	Gene Expression Data Mining
GLIF	Guideline Interchange Format
GME	Global Model Exchange
GMP	Good Manufacturing Practice
GO(A)	Gene Ontology (Annotation)
GRMS	Grid Resource Management System

GT4	Globus Toolkit 4.0 Release
GTS	Grid Trust Service
GUI	Graphical User Interface
HCS	High Content Screening
HL7	Health Level Seven
HPC	High-Performance Computing
HITSP	Healthcare Information Technology Standards Panel
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
ICH	International Conference on Harmonisation
IEC	Independent Ethics Committee
	International Electronical Commission
IRB	Institutional Review Board
ISF	Investigator Site File
ISO	International Organization for Standardization
JADE	Java Agent DEvelopment Framework
JSDL	Job Submission Description Language
JSNP	Japanese Single Nucleotide Polymorphism
KEGG	Kyoto Encyclopedia of Genes and Genomes
LGPL	Lesser General Public License
MAPPER	Multiscale Applications on European e-Infrastructures
MDA	Model-Driven Architecture
MIME	Multipurpose Internet Messaging Extensions
miRNA	Micro RNA
MOM	Message-Oriented Middleware
MTM	Multidisciplinary Team Meetings
NBOCAP	National Bowel Cancer Audit Programme
NCCN	National Comprehensive Cancer Network

NDC	National Drug Code
NoE	Network of Excellence
OASIS	Organization for the Advancement of Structured Information Standards
OBO	Open Biomedical Ontologies
ODM	Operational Data Format
OGF	Open Grid Forum
OID	Objects Identifier
OMG	Object Management Group
OMIM	Online Mendelian Inheritance in Man
OWL	Web Ontology Language
PDM	Platform Definition Model
PIM	Platform Independent Model
PIR	Protein Information Resource
PKI	Public Key Infrastructure
POX	Poor-Old-XML
PRACE	Partnership for Advanced Computing in Europe
PSM	Platform Specific Model
QoL	Quality of Life
RDF	Resource Description Framework
REST	REpresentational State Transfer
RNA	Ribonucleic acid
ROA	Resource Oriented Architecture
RPC	Remote Procedure Call
SAGA	Simple API for Grid Applications
SAML	Security Assertion Markup Language
SCC	Study Coordination Centre
SDV	Source Data Verification
SEI	Software Engineering Institute

SIB	Swiss Institute of Bioinformatics
SMTP	Simple Mail Transfer Protocol
SNP	Single Nucleotide Polymorphisms
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
SOA	Service Oriented Architecture
SOP	Standard Operating Procedure
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SVM	Support Vector Machines
TMF	Trial Master File
TOGAF	The Open Group Architecture Framework
UDDI	Universal Description, Discovery and Integration
UE	Usability Engineer
UEP	Usability Engineer Process
UML	Unified Modelling Language
UMLS	Unified Medical Language System
UNICORE	Uniform Interface to Computing Resources
URI	Uniform Resource Identifier
VO	Virtual Organizations
VPH	Virtual Physiological Human
WfMC	Workflow Management Coalition
WSDL	Web Service Description Language
WSRF	Web Services Resource Framework
XACML	Extensible Access Control Markup Language