Course:
Bioinformatics
Name:
Mark de Groot
College number:
0455253
Version:
1.0
Emailaddress:
martinus.degroot@student.uva.nl

BTB or not to B(TB)?

30 januari

2007

[That's the question… an edit of a famous quote of the English poet William Shakespeare. During this course we have set up a project which has to do something with Bioinformatics and Medical Informatics. In this paper I will tell about these two areas and there (possible) connections. In order to complete and execute this project it was necessary to use GRID and e-Bioscience, so I will mention these terms in this paper]

Part of the course bioinformatics

TABLE OF CONTENTS

1. INTRODUCTION

Through this paper I will tell you about Bioinformatics and a couple of topics which are related to Bioinformatics. The references I used to write this document are stated in chapter 8: References. The information extracted from the literature is combined with my own interpretations of the topics which are discussed in this paper.

In the conclusion I will state what I have learned this month and what I think will happen in the (near) future. I will discuss the way the education has taken place and give some advices to the instructors if necessary.

Chapter 2 will be about the course Bioinformatics and its topics. It is a global overview of the four weeks. Chapter 3 will be about Bioinformatics and Medical Informatics; what do these two areas have in common, what are the (main) differences between these two areas and – last but not least nevertheless important – is it possible to combine the efforts to serve a higher purpose (like: the curation of cancer and so on)? In this chapter will not be extensively written about these three areas, it is almost point wise, because I found the other chapters more interesting.

Nowadays it is not possible to accomplish (major) breakthroughs without any help of the information technology, so chapter 4 handles the topic of E-Bioscience; what is it, and why is it needed?

There are some new ways to interact across the globe like the internet and e-mail. Would it not be interesting to connect scientific centers which do the same research as other centers across the globe? Provide these centers with a network so there are able to share resources and not only data, everyone can have its own data but will share it across a network and can use highly sophisticated systems and/or machines to perform a search or a calculation which were not possible with a single computer, but may be a GRID can make this happen. Chapter

5 will discuss about this topic; what is a GRID, why is it needed and can it help – not only scientists – to serve a higher purpose?

During this course we – we is here defined as: Marianne Heling, Michelle Niekoop and me – had to execute a project. We have chosen the implement the BTB at a nationwide scale. Last year a group of our students had written a new BTB for the neurosurgery department of the AMC and the clinicians where delighted with the new BTB and they were looking forward to it. It was made clear when the BTB was up and running; Sieger Leenstra would pass the positive experience to the other centers. So our project choice is fundamental.

This paper has to be written in English; may English is not my best language, so there is a possibility that some constructions are not correct. Nevertheless I find it a good way to practice my skills for the upcoming Masters.

2. THE COURSE BIOINFORMATICS

The first week after a two week break was a long week. Five days from nine to five or half past five. Nevertheless I found it interesting to hear something about bioinformatics and some of her topics. This week was an introduction to bioinformatics and I would not have minded if there was some sort of a refreshing hour; to speed across some primary cell processes so people would immediately know what you are talking about.

Unfortunately the lecture about the Human Genome Project was cancelled and that was one of the topics I was particularly interested in. Other topics I like were: Bioinformatics of sequencing, Transcriptome map, Gene finding, Sequence alignment and DNA microarray analyses.

Lectures were alternated with practice. After a lecture there was a practice about the topic that we were are lectured about. I found it very handy, because you learned about a specific topic and directly afterwards you do some practice to get familiar with the topic, the activities around this topic and the activities that are close related to the topic.

The rest of the course exists of two meetings a week to talk about topics which are closely related to Bioinformatics and what Medical Informatics has to do with it. The second week started with a meeting which evaluated the first week. Some remarks and suggestions were made for the next year.

In the second week the student group was divided into three subgroups. Each group had to talk about a subject the second meeting later that week. The subjects were: Bioinformatics & Medical Informatics, E-Bioscience and GRID. We had to read a lot of articles before we could even talk about these subjects. These presentations were necessary – according to me – before each group could continue with the forth exercise.

The second meeting of that week the three different groups gave their presentations. After each presentation there was a discussion of the topic. These discussions required that you had read the provided articles, so you could participate in those discussions. It was interesting to share your opinion with the other students and the teachers. It is a way to talk about the topics and exchange some ideas and sometimes someone has a point you never thought about. These discussions were lead by a discussion leader who had the task to guide the discussion and fuel the discussion with new statements and so on. This was a good way to get familiar with the topics of this course.

During the third week we interviewed two experts in the area of E-Bioscience and GRID. These experts provided us with some insight information about their area and the influence of E-Bioscience and GRID. This was interesting because these people know a lot about these terms and they can tell you what they think about the possible advantages of E-Bioscience and GRID – and disadvantages if they encountered any. These interviews somehow have influenced my opinion and the definition of the two terms – I have become more critical and for the project I have read more to find a way to implement our project the best we could reach.

Later this week we finalized our project and what the main purpose – and benefits – will be when our solution will be implemented. We perform a literature study to get more information about the GRID and E-Bioscience. We also search for projects that exist for information and when we found projects we decided if our project was the same and in that case we could connect or link our project to a project that exists.

The last week we completed our project and we prepared a presentation. We will explain what our project is, the main goals, what the differences are between nowadays and the – near – future, costs, benefits and many more. For more information see chapter 6. The evaluation of this course can be found in the conclusions of this paper.

3. BIOINFORMATICS & MEDICAL INFORMATICS

3.1. WHAT IS BIOINFORMATICS?

This question is not as easy to answer as it seems; when you search the internet or read articles you will come across different – and not always the same – definitions. Here are some definitions:

Bioinformatics has evolved to handle large amounts of sequence and structural data, generated in the laboratory[1].

The term bioinformatics could refer to: informatics involving genomics, informatics involving the biosciences, informatics involving the biosciences and clinical research or all biomedical and health informatics[2].

A scientific discipline that comprises all aspects of the gathering, storing, handling, analyzing, interpreting and spreading of biological information. Involves powerful computers and innovative programmes which handle vast amounts of coding information on genes and proteins from genomics programmes. Comprises the development and application of

[1] Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvarova J, Hermosilla I, Vicente FJ. (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform*. **37(1)**, 30-42.

[2] Miller PL. (2000) Opportunities at the intersection of bioinformatics and health informatics: a case study. *J Am Med Inform Assoc*. **7(5)**, 431-8.

computational algorithms for the purpose of analysis, interpretation, and prediction of data for the design of experiments in the biosciences[3].

When Google is asked to find definitions of Bioinformatics it will give you 25 definitions, this gives an idea about how difficult it is to define Bioinformatics. They all describe these stated definitions – or a part of it. Combining the three definitions above will create a definition like this:

> *A scientific discipline that comprises all aspects of the gathering, storing handling, analyzing, interpreting and spreading of biological information, sequenced and structured data generated in the laboratory. Comprises the development and application of computational algorithms for the purpose of analysis, interpretation, and prediction of data for the design of experiments in the biosciences.*

I combined these three definitions to a definition of what I think it is Bioinformatics. The domain of Bioinformatics is – mainly – genomics and informatics. The area where these people are active could be one of those[4,5]:

- Sequence analysis of proteins and DNA
- Analysis of microarrays
- Proteomics
- Genomics
- Molecular Structure
- Analysis and simulation of metabolic networks

[3] European Commission Research. Website: http://ec.europa.eu/research/biosociety/library/glossarylist_en.cfm?Init=B
[4] http://studiegids.uva.nl/web/sgs/nl/c/116.html
[5] Detmer Don E. (2003) Building the national health information infrastructure for personal health, health care services, public health, and research. *BMC Medical Informatics and Decision Making* 2003, **3(1)**. 1-12.

The most challenging project of the past decennium – or the most known project – was the Human Genome Project. The purpose of this project was the map the entire human genome. The human genome map is now freely available on the internet. Later on – subsection 3.3 – I will talk about the combination of this area with Medical Informatics.

3.2. WHAT IS MEDICAL INFORMATICS?

The only definition I found correct is the one which is always used during promotion weeks and instruction presentations at schools:

> *Medical Informatics has traditionally been focused on the development of computer applications for representing and implementing health care*

The – main – domain were Medical Informaticians are active are medicine and informatics. They form a liaison – a bridge – between the medical staff and the die-hard ICT personnel. Because these two areas do not understand each other very well, it was needed to create a new study like Medical Informatics. There are four major areas where Medical Informatics is active:

- University research and education
- Health care: clinical departments and ICT services of hospitals, but also foundations like Stop AIDS Now and Hartstichting
- Trade and Industry and the ICT industry
- Government

Projects where Medical Informatics plays a role such as the nationwide database of the cardio surgery centers in the Netherlands – NICE registration – and SNOMED, AMC Zorgdesktop, BTB AMC and so on.

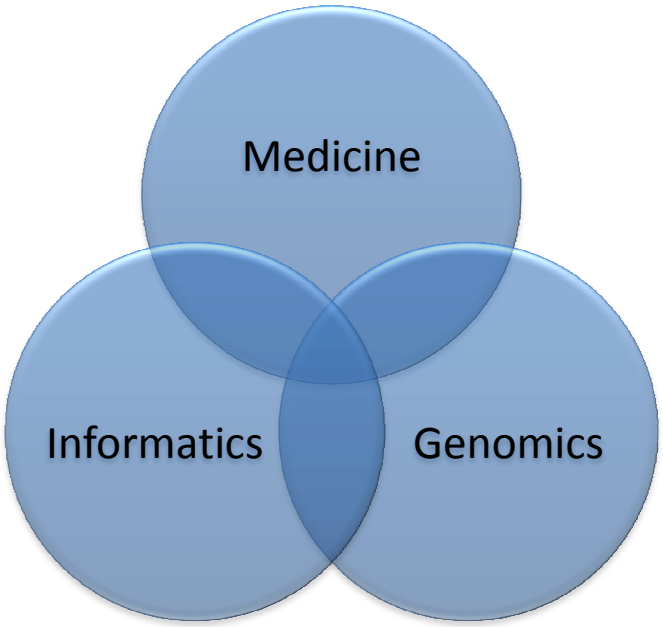
3.3. WHAT IS BIOMEDICAL INFORMATICS?

The most interesting part of this chapter is this subsection, because is it possible to create a new area which connect the Bioinformatics with the Medical Informatics? Because there are some differences, but are these differences unbridgeable? The user groups are different; Bioinformatics deals with almost only one user group, the researchers, and Medical Informatics deals with different user groups, administrative personnel, medical staff, nurses, surgeons, administration and so on. In this section it will become clear if the differences are bridgeable.

3.3.1. AREA OF BIOMEDICAL INFORMATICS

The area of Biomedical Informatics is much easier to explain with a schema – see the next page. As you can there are three areas; medicine, informatics and genomics. The overlap between informatics and genomics is called Bioinformatics, the overlap between medicine and informatics is called Medical Informatics; and the overlap between Bioinformatics and Medical Informatics – the overlap between medicine, informatics and genomics – is called Biomedical Informatics.

The figure made it clear that none of the areas that nowadays exist will disappear, however there is a possibility that some areas will grow and some areas will get smaller. The area which will definitely grow is the Biomedical Informatics area, if the there is willingness of the people who are involved.



3.3.2. SIMILARITIES AND POSSIBILITIES

When we approach the combination Bioinformatics and Medical Informatics from the other side – similarities instead of differences – we can see that they really do have similarities both areas create applications to do research or applications that speed up the process of doing things. That is just one example, there will follow more.

The health care is moving from patient/hospital-centered systems to citizen/community systems[6,7]. Maojo stated that eventually Medical Informatics will be driven to Bioinformatics because of the genetic data that becomes widely available.

Some of the other similarities are that both areas use machine learning, natural language processing, image analyses, database research, information retrieval, database integration, and decision support systems.

[6] Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvarova J, Hermosilla I, Vicente FJ. (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform*. **37(1)**, 30-42.

[7] Maojo V, Kulikowski CA. (2003) Bioinformatics and medical informatics: collaborations on the road to genomic medicine? *J Am Med Inform Assoc*. **10(6)**, 515-22.

As stated in the article of Martin-Sanchez, we need computers not only to store the data we collect, but also to verify and expand the interpretations where are constructing, because the dataflow grows – there is more data collected from apparature than a couple of years earlier. We are able to reach this when; (1) stimulating integration at the informational exchange level, (2) initiating collaborations between Bioinformatics and Medical Informatics, (3) training a new generation of scientists that speak both languages. This last remark demands scientists that speak the Medical Informatics language and speak the Bioinformatics language – in other words – the language of Biomedical Informatics.

4. E-BIOSCIENCE

The term e-science was introduced by John Taylor. Taylor saw that many areas of science were becoming increasingly reliant on new ways of collaborative, multidisciplinary working. The term e-science is intended to capture these new modes of working:

> *E-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it*[8].

The reasons why e-bioscience – e-science in a biological area – becomes needed is because of there will be more and more data generated. The exchange which takes place nowadays causes an increasing dataflow which generates more and more data.

There has to be a network that is reliable, secure and capable to handle the dataflow, data storage and computational power. That is where e-bioscience and GRID – chapter 5 – become needed. GRID provides the network and e-bioscience provides in tools to manipulate the data.

The e-bioscience part of the network handles data handling, preprocessing and fusion, data integration and knowledge presentation, process modeling and dynamic simulation and complex systems approach. Without this part it is not possible to run the entire network and infrastructure, because you cannot make it without protocols, programs and so on. Visa versa is also true; an e-bioscience idea cannot survive without an infrastructure and network.

[8] Hey T, Trefethen A. (2003) E-Science and its implications. *The Royal Society*. **361**, 1809-25.

5. GRID

The GRID is a popular term according to Google. When you search for GRID you will get over 1.000 pages which all contain the term GRID. First of all we need a definition, and then I will explain shortly what a GRID is and why it should be used. The name GRID has been derived from the power GRID because of the similarities they both try to accomplish.

> *The short answer is that, whereas the Web is a service for sharing information over the Internet, the Grid is a service for sharing computer power and data storage capacity over the Internet. The Grid goes well beyond simple communication between computers, and aims ultimately to turn the global network of computers into one vast computational resource*[9].

This definition comes close to the definition of Ian Foster[10]. Ian Foster also stated three checkpoints; when all three are fulfilled than the observed system or infrastructure is a GRID[11]. These three points are:

- GRID is a system that coordinates resources that are not subject to centralized control. A GRID coordinates and integrates resources and data from different domains, for example: different companies, users, and different administrative units

[9] http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html
[10] Foster I, Kesselman C, Tuecke S. (2001) The Anatomy of the GRID. *Intl J Supercomputer Applications*. 1-25.
[11] Foster I. (2003) What is the GRID? A Three point checklist. *Argonne National Laboratory & University of Chicago*. 1-4.

of the same company or other companies and addresses the issues of security, authentication, policy, payment and membership.

- GRID is a system that uses standard, open, general purpose protocols and interfaces. A Grid is built from multi-purpose protocols and interfaces that address such fundamental issues as authentication, authorization, resource discovery, and resource access.
- GRID is a system that delivers nontrivial qualities of service. A Grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, relating for example to response time, throughput, availability, and security, and/or co-allocation of multiple resource types to meet complex user demands, so that the utility of the combined system is significantly greater than that of the sum of its parts.

Typical for a GRID is that you provide access to users with a certificate. The certificate is something like a passport which gives you or doesn't give you access to certain data. This goes by the use of encryption and public/private keys. Later on more will be said about.

Why using a GRID instead of internet? Internet is a medium which provides information. It is a PUSH mechanism, because it also gives you unnecessary data, data you did not ask for. A GRID is a medium where you can find specific information by using the right query. It is a PULL mechanism, because it only gives you the information you need. A GRID is also very intelligent, it knows exactly where the data is set and by using the right query it gives you the right information.

There are five different kinds of GRIDs:

- Computational GRID is about sharing computer forces. Create a supercomputer to calculate the most memory draining operations. A computational GRID can calculate a submission which was not possible on a single computer because of the time needed and/or a memory draining operation.
- Data GRID is basically about storing large amounts of data and share and manage this data across the GRID.

- Information GRID provides information for a certain domain; Google, Yahoo and YouTube are good examples.
- Knowledge GRID is about semantic networks.
- Sensor GRID is connected to equipment – for example – a telescope and the surrounding GRID is used to control the equipment remotely and to analyze the data produced.

When using a GRID it is possible that a service of the GRID is temporarily unavailable. Therefore the GRID has replicas of the data; these replicas are stored across the GRID network. The GRID has a catalog which stores the information where the data is stored. Which location the GRID visits for the specific data you requested depends on your physical location. When you are closer situated to a replica you receive the replica data, if the replica data is down – the element does temporarily not work – you will receive the original data (or an another copy).

The main advantages of a GRID are that you do not have to know where the data is stored. You just sent your job to the GRID and the GRID finds the solution. All participating users are still in full control of their data and still have to update and maintain it periodically. Another advantage is that the GRID has near infinite storage capacity or computing capacity – or both – which makes it possible to perform researches which were not possible before.

Some data is valuable because it contains personal information; scientists and medical staff are afraid that this data will fall in the wrong hands. A GRID is a closed network; you have to get an account to get access, a security service checks all information before you get access to the GRID – also by using security certificates. Another option which can be used is the distributed encryption; the decryption/encryption key is divided and spread across the network, when a hacker hacks an element he gets only a part of the decryption/encryption key.

6. THE PROJECT – BTB OR NOT TO B(TB), THAT'S THE QUESTION…

6.1. THE BRAIN TUMOR BANK

In this chapter I will write about our project. Before it is possible to talk about how we would like to implement our project it is necessary to explain what the BTB is and how it looks like. It is also necessary to state the purpose of the BTB and what kind of system the BTB is, because supporting direct patient care or indirect patient care are two different things. The BTB supports the indirect patient care.

Both Bioinformatics and Medical Informatics will have a role, Bioinformatics to research the LIMS data and Medical Informatics the liaison between ICT and the Medical Staff, so the system will be a success – the ICT and medical staff communicate through the Medical Informatics with each other when it is necessary to modify the system. Besides these two things it is necessary to explain what the (main) advantages will be when a nationwide BTB has been implemented. In section 6.2 I will write about the technical aspects of the project and some considerations.

6.1.1. WHAT IS THE BRAIN TUMOR BANK [BTB]?

The Brain Tumor Bank holds specific data about every patient who had visited the neurosurgery department of the AMC. This database has been set up for the analyses of the data to answer the research questions. Unfortunately this is not possible with the current database; there is too much distortion among the data fields; there are unknown fields (it is not known what the specific field means), some fields are used for a select – small – group of patients. This problem causes that only a select user group have been able to use the

database and for the most interesting research questions it is currently not possible to use the database.

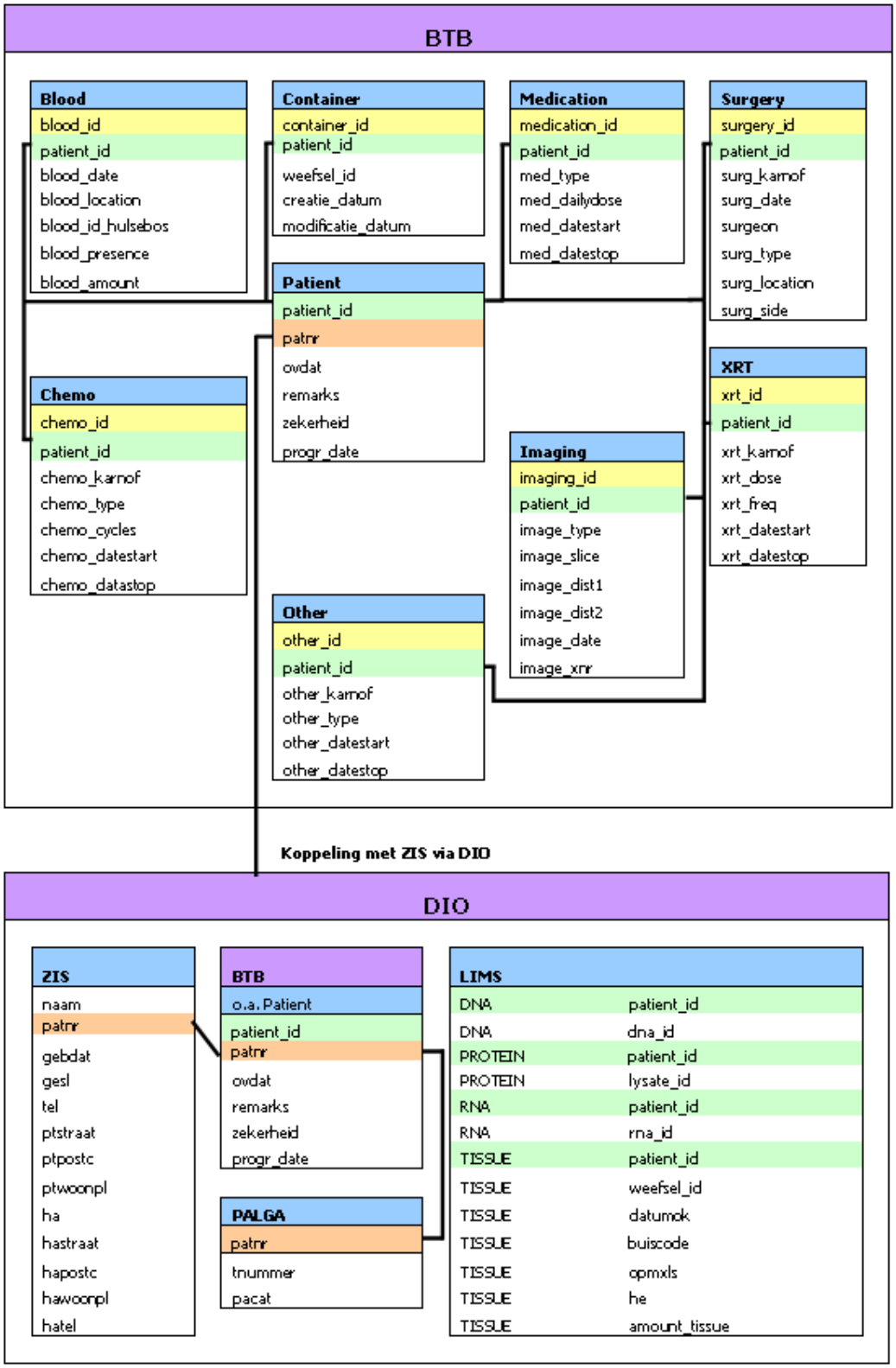
So last year – May to June 2006 – we started to restructure the BTB. We decided it has to be Oracle or SQL based (the AMC has to decide which of the two languages will be used), so it become possible to answer the most complex queries / research questions. For the other questions it is possible to create something like a button which holds a particular question, like: “How many patients have been diagnosed with tumor X in the past Y” where X and Y will be defined by the user input. Only the most complex research questions require a user who is capable of using SQL.

We also decided that the BTB would have a major advantage when it is incorporated with the LIMS and DIO. These two terms are explained below:

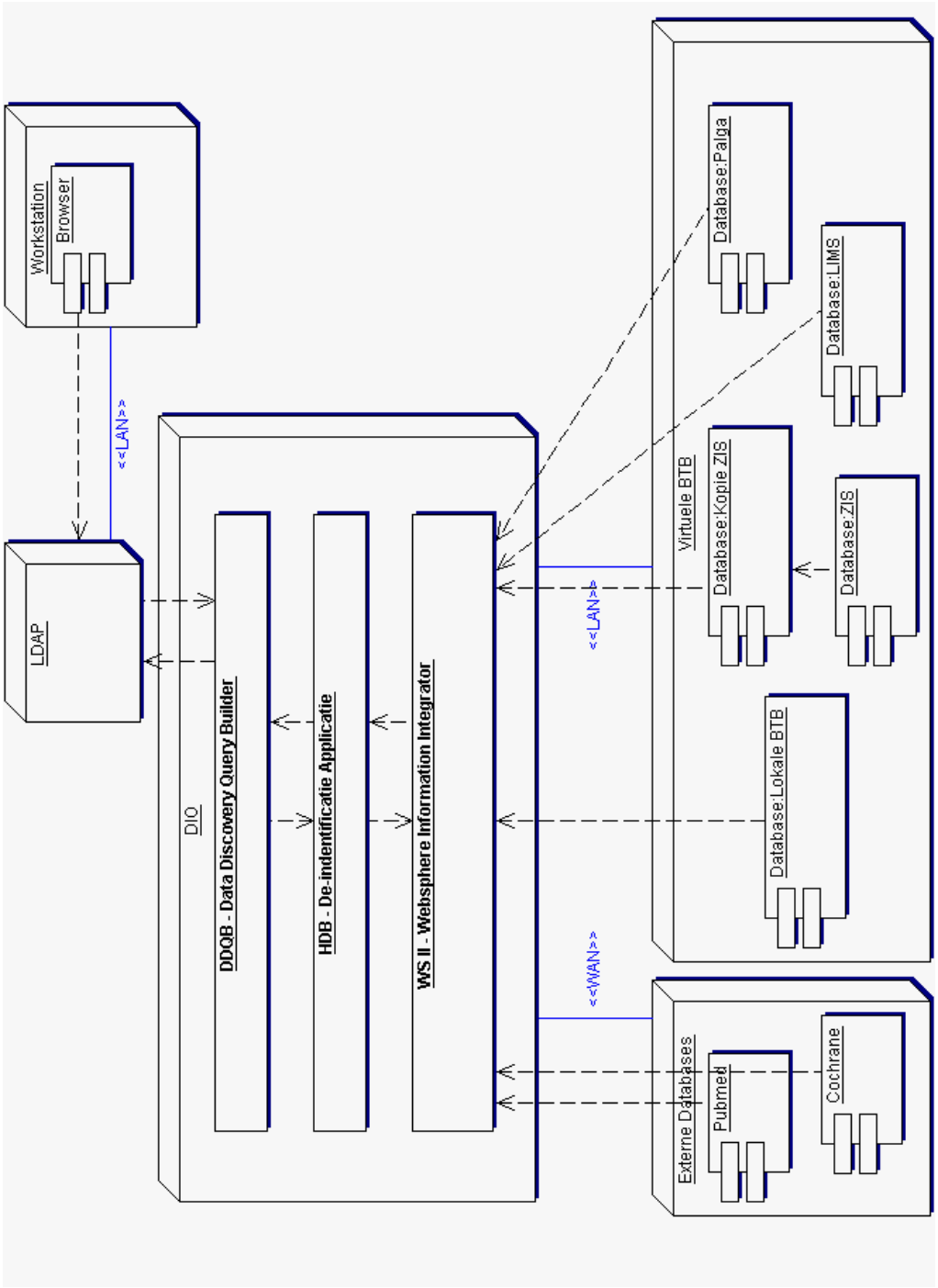
DIO is an abbreviation of ‘Data Integratie ten behoeve van Onderzoek’ [Data integration in favor of research]. It is a quite similar project like the IBM project which is currently implemented at the United States. This project connects different databases with each other. It provides a user with an interface where she/he has been able to query all databases which are linked in the project. The user does not have to know where the data is saved, because the DIO can handle this. One very important remark is that all data fields have to be unique across the databases which are linked through the DIO-project.

LIMS is an abbreviation of ‘Laboratory Information Management System’. This is software which is used in the laboratories to manage samples, laboratory users, instruments, standards / protocols and (many) other laboratory functions. The LIMS is flexible; users are able to expand or restructure / change the system (to certain limits of course).

In this project we take it for granted that the BTB is implemented like this (it is already partially implemented with no major drawbacks). The BTB contains – see next page:


BTB
Blood
blood_id
patient_id
blood_date
blood_location
blood_id_hulsebos
blood_presence
blood_amount
Container
container_id
patient_id
weefsel_id
creatie_datum
modificatie_datum
Medication
medication_id
patient_id
med_type
med_dailydose
med_datestart
med_datestop
Surgery
surgery_id
patient_id
surg_karnof
surg_date
surgeon
surg_type
surg_location
surg_side
Patient
patient_id
patnr
owdat
remarks
zekerheid
progr_date
Chemo
chemo_id
patient_id
chemo_karnof
chemo_type
chemo_cycles
chemo_datestart
chemo_datastop
XRT
xrt_id
patient_id
xrt_karnof
xrt_dose
xrt_freq
xrt_datestart
xrt_datestop
Imaging
imaging_id
patient_id
image_type
image_slice
image_dist1
image_dist2
image_date
image_xnr
Other
other_id
patient_id
other_karnof
other_type
other_datestart
other_datestop
Koppeling met ZIS via DIO
DIO
ZIS
naam
patnr
gebdat
gesl
tel
ptstraat
ptpostc
ptwoonpl
ha
hastraat
hapostc
hawoonpl
hatel
BTB
o.a. Patient
patient_id
patnr
owdat
remarks
zekerheid
progr_date
PALGA
patnr
tnummer
pacat
LIMS
DNA patient_id
DNA dna_id
PROTEIN patient_id
PROTEIN lysate_id
RNA patient_id
RNA rna_id
TISSUE patient_id
TISSUE weefsel_id
TISSUE datumok
TISSUE buiscode
TISSUE opmxls
TISSUE he
TISSUE amount_tissue

In combination with DIO:



When all of this has been realized it is possible to feed the database – by supplying the DIO with a query – with complex research questions. The BTB holds very interesting data to answer very interesting research questions. The neurosurgery department is particularly interested in the course of a glioblastoma, the most aggressive tumor among the brain tumors.

6.1.2. THE PURPOSE OF THE BTB

The main purpose of the BTB is doing research at the area of neurosurgery and research on glioblastomas in particular. The reason why the neurosurgery department is so interested in the glioblastoma is because of the lethality of this brain tumor. Within five years 90% of the patients diagnosed with a glioblastoma are dead. The second – nevertheless very important – reason is that there is not much known about this tumor.

These purposes do not serve direct patient care; therefore it is not possible to use this database directly in the daily patient care. This was also a demand to participate in the DIO project, only research databases are allowed to join.

This database will be particularly used by researchers for research, or by neurosurgeons who are interested in some clinical research questions, like: what is the progression of tumor X, what is the top ten of tumors the last year and so on. Because of the research setting of the database it is not possible to use this database in direct patient care, because there is no patient specific output, the analyses made by the database (BTB and DIO) are means and medians of an selected population in a particular area – like province Noord Holland. These outcomes or results of the queries are not directly map able – from the verb mapping – to an individual. This remark has to be always remembered when this database is used.

6.1.3. ADVANTAGES OF THE BTB

When everything is working fine it is possible to acquire new insights at the area of neurosurgery. With this information it is possible to create new models of diseases or make better or new prognoses about a specific disease.

It is possible to do a more pinpointed research at particular variables if these variables deviate regular in tumor X. These researches acquire new data and knowledge; these two can pursue new research and/or open doors.

These acquired insights could help to improve the fight against diseases and tumors, or new research for new medication or improved medication. These are just some examples of in which the BTB could play an important role (I would say should instead of could), because there is still not much known about the human brain and its disorders.

6.1.4. WHY SHOULD THE BTB GO NATIONWIDE?

The BTB of the AMC (department neurosurgery) contains relatively a few patients. The more patients (patient data) are stored in a database, the more trustworthy the outcomes of a particular research question will be.

Beside this reason there is an another reason about the size of the database; the more patients are stored in the database, the bigger is the chance that seldom disorders will be in the database and therefore it is possible to query the database about these seldom disorders.

A third reason is that the most interesting tumor – the glioblastoma – does not occur frequently in the BTB of the neurosurgery department of the AMC, and nationwide neither. So it recommendable to link all the neurosurgery databases (BTB like) to acquire more patient data so the outcomes will be more reliable. There are some more advantages to link the databases, for example compare the hospitals with each other (which department has the most survivors e.d.) and so on.

When these hospitals save the data in a BTB like database, then it is possible to connect these databases. These databases should be connected by a data GRID and federated databases. The main advantages will be the access to nearly infinite computer power and infinite storage capacity and sharing among (distant) colleagues. By doing a literary research

it will become clear which way is be the best (or the less worse) option to implement a nationwide BTB by connecting all the relevant databases.

We are talking about patient data so security, authentication, authorization and integrity are important issues which clearly will influence our choice to implement a data GRID and a federated database structure and maybe it is nowadays not possible, but we are going to find it out.

6.2. THE PROJECT

What is a GRID without participants? That will not be a GRID, so the first subsection will explain who will be the collaborators and where those collaborators geographically are located. The Second subsection will explain the structure and infrastructure of this project and explains how a job will be executed and how the results will be returned. The third subsection will show the resource allocation, which will be located where and a brief explanation about this allocation. What about the financial part? Is it reachable to implement the BTB? These questions will be handled in the forth subsection.

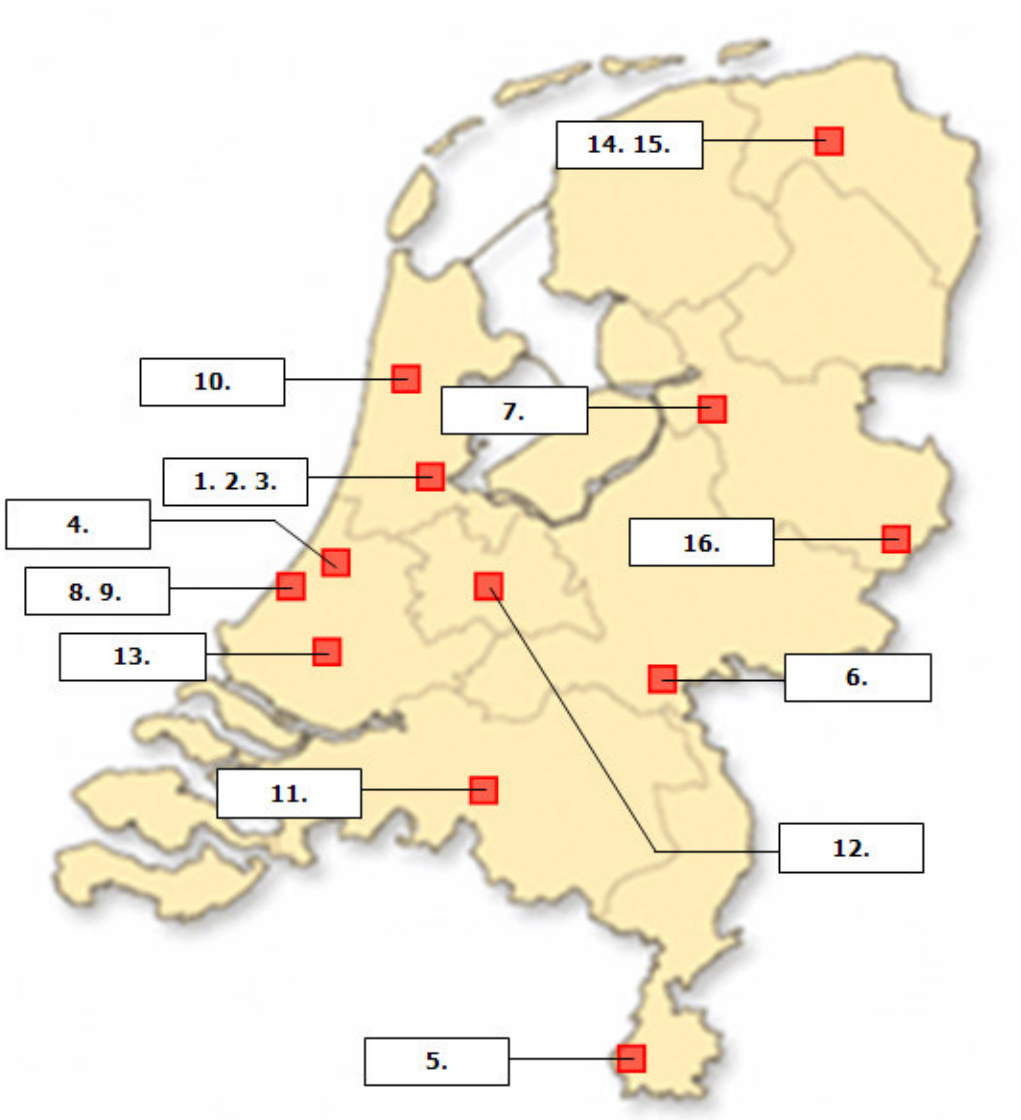
6.2.1. THE COLLABORATORS

Last year we have revisited the BTB for the neurosurgery department of the AMC. Before the project can go nationwide it is necessary to collaborate with other neurosurgery departments across the country. It should be promoted by the 'Dutch Society of Neurosurgeons', because the thirteen academic centers and the three biggest peripheral centers are members of this society.

Last year we spoke with Sieger Leenstra and he and the neurosurgery department of the AMC were very enthusiastic about the new BTB. When it was implemented he would advice the 'Dutch Society of Neurosurgeons' to do the same, so it would be possible to do relevant research across the country.

The collaborators are – also see the figure on the next page:

#	Collaborator	#	Collaborator
1	Academisch Medisch Centrum – Amsterdam	9	Westeinde Ziekenhuis – Den Haag
2	Vrije Universiteits Medisch Centrum – Amsterdam	10	Medisch Centrum – Alkmaar
3	Slotervaart Ziekenhuis – Amsterdam	11	Neurochirurgisch Centrum Tilburg – Tilburg
4	Leids Universitair Medisch Centrum – Leiden	12	Universitair Medisch Centrum – Utrecht
5	Academisch Ziekenhuis Maastricht/Heerlen – Maastricht	13	Erasmus Medisch Centrum – Rotterdam
6	Academisch Ziekenhuis Nijmegen – Nijmegen	14	Universitair Medisch Centrum – Groningen
7	Neurochirurgisch Centrum Zwolle – Zwolle	15	Martini Ziekenhuis – Groningen
8	Leyenburg Ziekenhuis – Den Haag	16	Medisch Spectrum – Enschede



6.2.2. THE STRUCTURE

After a literature research we decided that a data GRID and a federated database structure were possible. For the understanding of the paper it is necessary to give a definition of federated databases.

> *A federated database is a logical association of independent databases that provides a single, integrated, coherent view of all resources in the federation. The federation architecture makes several distinct physical databases appear as one logical database to end-users[12].*

The reasons why we have chosen these structure:

- Nearly infinite computer power – needed for complex queries and jobs, comparison of MRI images for example.
- Nearly infinite storage power – needed to save large amounts of data and information about patients.
- More patient data available – the researches performed will have a higher power and it is possible to do research on seldom diseases.
- To get information the clinicians would like to have – produce status rapports, like the top ten diseases the last year.
- It is a closed network, so it is more difficult for unauthorized users to access the GRID. The encryption/decryption key is divided and distributed across the network; in case of an unauthorized access the data is not readable.
- More chance of major breakthroughs.

[12] http://www.b-eye-network.com/view/2164

With this list we have to convince the future users to support the nationwide BTB and use it to benefit from it. Let's have a closer look on how it works. The figure on the next page is a simplified version of our project; the green arrows show the way from the user throughout the GRID and the blue arrows show the way from the GRID to the user. The figure globally shows the path from sending a job to the GRID to report the results back to the user.

There are mainly two groups who are going to use the nationwide BTB; clinicians and researchers. They have installed the necessary software so they are able to use the GRID. Before they can connect or contact the GRID they have to verify who they are and if they have the rights to perform the job they would like to perform. This is handled by the security service. It checks the certificates and validates the user – authorization, verification and authentication.

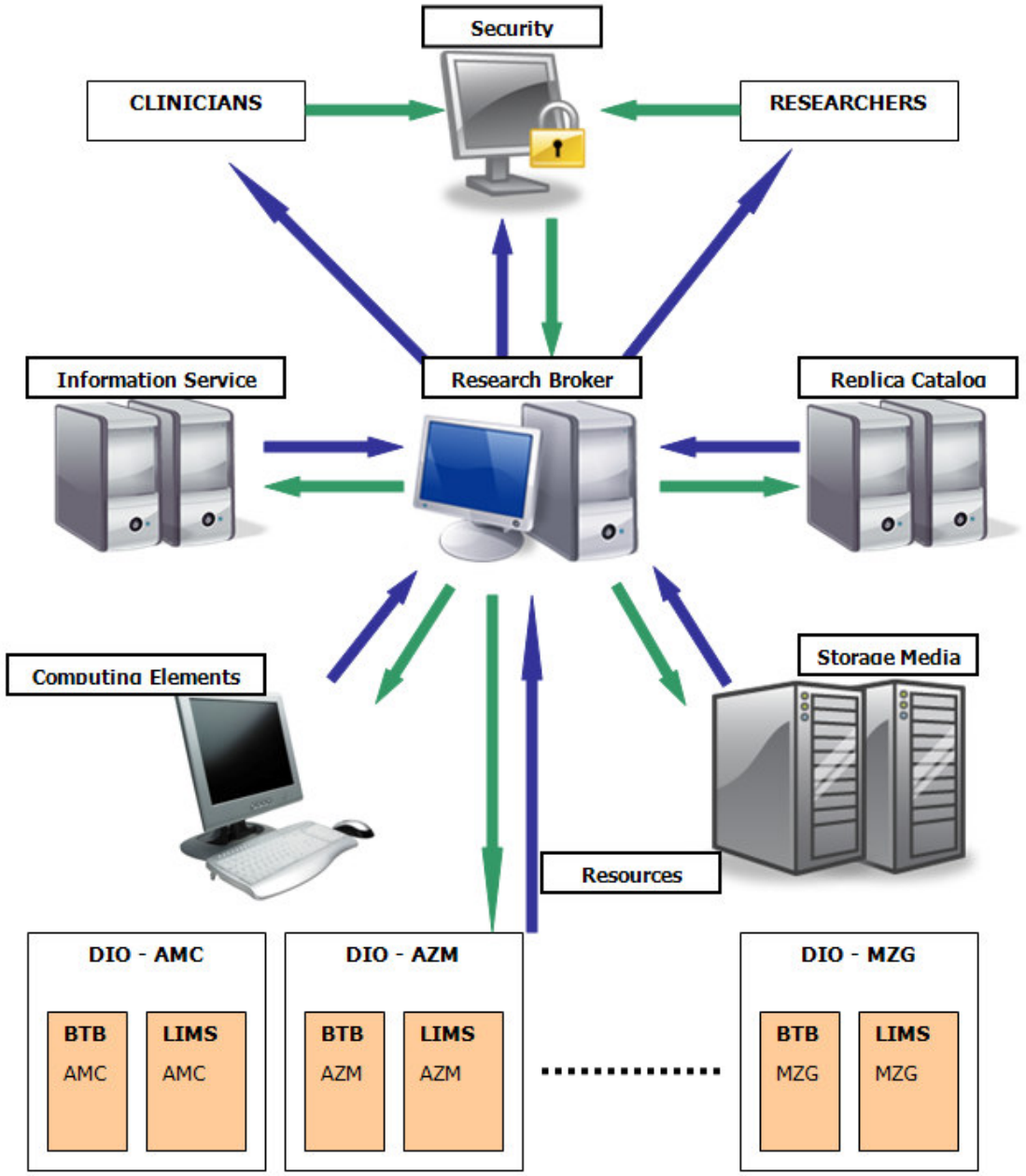
Once the user has been cleared, the user is able to communicate with the Research Broker. The Research Broker is the heart of the GRID and it will find the best resources to execute your job – if your job contains a computing part and a part that request data from the BTB, the Research Broker will allocate computing power to perform the job and collects the necessary data for your job.

How does the Research Broker know where all the data is stored and where there is computing power available? Therefore the Research Broker queries the Information service to know which software and hardware are currently available, meanwhile the Replica Catalog has been contacted so the Research Broker knows all locations of the existing data.

When all appropriate resources have been located the job will be executed. After the job is completed the Resource Broker sends the results back to you. These results can be modified to certain layouts by the user. The Storage Media saves the job so that information will not be lost if the user its computer crashes.

The federated database component in the GRID is that all standalone databases are connected through a mediator. The Resource Broker will contact the mediator and the mediator extracts all information out of the standalone databases. The mediator looks like a

huge storagefacility, but it provides only the connections between the standalone databases and extract the data from the standalone databases when it is required for a specific job.



6.2.3. RESOURCE ALLOCATION

The next figure will show the resource allocation. There are two computing elements; these computers only perform computational jobs. The Resource Broker can decide that other facilities in the GRID network will serve as computing elements. There are five storage elements, but the Resource Broker decides were jobs are stored – thus it is not up to the user where jobs are saved and which computing elements are used and/or other elements function as computing element.

Legenda:	
16	Resource Sites
2	Security Servers
1	Resource Broker
1	Information Service
1	Replica Catalog
2	Computing Elements
5	Storage Elements [To store the jobs, preformed researches and so on]

6.2.4. COSTS

It sounds all right, but what about the financial part? To answer this question we performed a literature research – again. We found information about a lot of projects, but the two we have used for our calculation of expenses are the Parelsnoer project and the UK e-Science Programme.

The Parelsnoer project will get € 35 million. According to the website of the UK e-Science Programme they get £ 213 million in five years – £ 42,6 million a year. That amount of money is equivalent to € 64,4 million a year.

Our project is not as big as the UK e-Science Programme; it is a bit smaller so the expenses are not as high as this project.

The Parelsnoer project contains all UMC's of the Netherlands, our project contains more centers – 16 in total – but contains only data of the brain, the Parelsnoer project contains data of diabetes, leukemia and other diseases. Our project contains a lot of data of the brain and patient data. Therefore it will be a little bit more expensive than the Parelsnoer project because it has – also – more participants. Combining these participants will cost more time than the Parelsnoer project, because there are more. There are also more changes required, because of the number of participants. Our budget will be € 40.000.000,-- and it will take approximately five years to implement a nationwide BTB. So it is reachable to implement the project, because there are many advantages – mentioned earlier in this paper – to implement the nationwide BTB.

MILJOENEN VOOR INNOVATIE

Voor innovatieclaims is de komende jaren 210 miljoen euro beschikbaar. In het najaar werkt het kabinet de plannen hiervoor uit, maar zeker is dat 35 miljoen euro wordt ingezet voor het project Parelsnoer. In dit onderzoeksproject bundelen alle Nederlandse universitaire medische centra hun expertise over ziekten zoals diabetes, reuma en leukemie en maken het toegankelijk voor medisch gebruik, wetenschappelijk onderzoek en commerciële toepassingen.
Bovendien gaat 15 miljoen euro naar de ontwikkeling van hightech instrumenten voor een internationale proefcentrale voor fusie-energie in Frankrijk binnen het project ITER. ITER is een internationaal samenwerkingsproject dat de wetenschappelijke en technische haalbaarheid van kernfusie als energiebron wil aantonen.

7. CONCLUSION

First I will discuss the course bioinformatics and afterwards I will discuss the project and finally I will discuss what I have learned and do some suggestions for the next year.

I decided to participate in the course bioinformatics instead of the course medical knowledge technology, because the bioinformatics topics are much more interesting. The course exists of lectures, practice, project and discussion meetings.

I liked the discussion meetings because you have to read the articles to participate in the discussion. A discussion is a perfect way to share opinions and statements with each other. Everyone his/her knowledge ads information to the discussion, that information is always in some way useable. The other participants could say something you never thought about or open a new perspective on the subject or change your perspective of the subject with good arguments.

The introduction week was a long week; it was a bombardment of information. I learned a lot about the different subjects in Bioinformatics. The practice during this week was handy to get familiar with some jobs, programs and the topics across the Bioinformatics spectrum.

The second week was an important week to me, because of the three presentations and the discussion meetings afterwards I understood the subjects better. I do not think I would have known so much as I do know now without this week.

I do not think that Bioinformatics and Medical Informatics will disappear, it is possible that the figure – subsection 3.3.1. – will change a bit, but the three areas will still exist and the overlap between these three areas too. Maybe the influence of some areas will decrease; nevertheless they will still exist, because the areas are needed.

I like doing projects in small group of students. I think that two-and-a-half week is too short to get the information you supposed to have – according to the bioinformatics course program – but nevertheless it was interesting to think of a useful application which both parties can use; Bioinformatics and Medical Informatics.

I do not regret my choice to participate in the course Bioinformatics; the two suggestions I would like to make are (1) shorten the first week – if that is possible – and (2) a fewer demands regarding to the projects.

Hopefully it was worth reading,

Mark de Groot

8. REFERENCES

8.1. FOOTNOTES

1. Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvarova J, Hermosilla I, Vicente FJ. (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform*. **37(1)**, 30-42.
2. Miller PL. (2000) Opportunities at the intersection of bioinformatics and health informatics: a case study. *J Am Med Inform Assoc*. **7(5)**, 431-8.
3. European Commission Research. Website: http://ec.europa.eu/research/biosociety/library/glossarylist_en.cfm?Init=B
4. http://studiegids.uva.nl/web/sgs/nl/c/116.html
5. Detmer Don E. (2003) Building the national health information infrastructure for personal health, health care services, public health, and research. *BMC Medical Informatics and Decision Making* 2003, **3(1)**. 1-12.
6. Martin-Sanchez F, Iakovidis I, Norager S, Maojo V, de Groen P, Van der Lei J, Jones T, Abraham-Fuchs K, Apweiler R, Babic A, Baud R, Breton V, Cinquin P, Doupi P, Dugas M, Eils R, Engelbrecht R, Ghazal P, Jehenson P, Kulikowski C, Lampe K, De Moor G, Orphanoudakis S, Rossing N, Sarachan B, Sousa A, Spekowius G, Thireos G, Zahlmann G, Zvarova J, Hermosilla I, Vicente FJ. (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform*. **37(1)**, 30-42.

7. Maojo V, Kulikowski CA. (2003) Bioinformatics and medical informatics: collaborations on the road to genomic medicine? *J Am Med Inform Assoc*. **10(6)**, 515-22.
8. Hey T, Trefethen A. (2003) E-Science and its implications. *The Royal Society*. **361**, 1809-25.
9. http://gridcafe.web.cern.ch/gridcafe/whatisgrid/whatis.html
10. Foster I, Kesselman C, Tuecke S. (2001) The Anatomy of the GRID. *Intl J Supercomputer Applications*. 1-25.
11. Foster I. (2003) What is the GRID? A Three point checklist. *Argonne National Laboratory & University of Chicago*. 1-4.
12. http://www.b-eye-network.com/view/2164

8.2. WEBSITES USED

- http://studiegids.uva.nl/web/sgs/nl/c/116.html
- http://www.b-eye-network.com/view/2164
- http://gridcafe.web.cern.ch/gridcafe/
- http://www.belnet.be/dyn/SC_gridcomputing_nl.pdf
- http://web.datagrid.cnr.it/pls/portal30/GRID.RPT_INTRODOCS.show
- http://www.rcuk.ac.uk/escience/default.htm [UK e-science programme]
- http://www.minocw.nl/documenten/ocwcourant.pdf [Parelsnoer Project]
- http://www.regering.nl

8.3. ARTICLES USED

- Requirements Document BTB.
- Elkin. (2003) Primer on Medical Genomics. Part V: Bioinformatics. *Mayo Clin Proc.* **78**,57-64.
- Collins FS et al. (2003) A vision for the future of genomics research. *Nature.* **422**,935-47.

- Sheth AP, Larson JA. (1990) Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Computing Services*. **22(3)**,183-236.
- Segal B. (2005) GRID Computing – The European Data GRID Project. 1-6.