Importance of Access to Biomedical Information for Researchers in Molecular Medicine

Annikki Roos; Turid Hedlund

Information Systems Science, Department of Management and Organization, Swedish School of Economics and Business Administration, Arkadiankatu 22, 00100 Helsinki, Finland
e-mail: annikki.roos@ktl.fi; turid.hedlund@hanken.fi

Abstract

In this paper, we analyze and describe the information environment of biomedicine from the point of view of the researchers in molecular medicine, which is a sub branch of biomedicine. We shall describe the nature of the discipline and its reflections to the information environment. A survey concerning the most important information resources in one molecular medicine research unit was conducted, and in this paper the main results of the survey is reported. The role of scholarly journals in the research process will also be analyzed. Special attention will be given to the possibilities of open access to the research process.

Keywords: information environment; information resources; databases; research process; molecular medicine

1 Introduction

The aim of this paper is to analyze and describe the information environment of biomedicine from the point of view of the researcher in molecular medicine (MM), a sub branch of biomedicine. Our target group is a research group containing researchers at different stages of their research career and the focus of study is on their daily work using information resources as part of the research process. The discipline is a rapidly growing and developing new research methods and processes which can be observed by the fact that pure laboratory work is to a growing degree transformed to computerized techniques. We argue that the change of the discipline from mainly laboratory based work to data based work has thoroughly changed the research processes. This has natural implications also to the information environment, as well as information retrieval, sharing practices and usage of information.

In this study the focus of research and our main research questions deal with the information environment of molecular medicine and firstly what are the main changes it has undergone. Secondly we investigate by conducting a survey, which are the most important information resources for researchers at different stages of their research career and thirdly what is the role of scholarly journals in the research process? For example, what is the publishing strategy and the criteria for choosing a journal to publish in.

We selected one research unit working in MM in Finland as a case. A web survey was conducted and qualitative information about researchers, their current work tasks, used information resources, publishing strategies and practices were gathered. A presentation and a feedback session concerning the results of the enquiry were given to the researchers. In this session important and explaining comments were given by the researchers in the target group about the use of information resources which have been taken into account when analysing and reporting the results of this study.

The outline of the paper is as follows: In Section 2, we describe the nature of the discipline and its reflections to the information environment. In Section 3, the effects of the changes in the environment will be analyzed against research process and scholarly communication practices. Special attention will be given to the experienced possible effects of open access in its different forms to the process. In Section 4, the results of the study are reported and in Section 5 we come to the conclusions and discussion.

2 Molecular Medicine as a Discipline

The discipline of biomedicine is growing exponentially. There are many factors behind the growth, of which the most important might be substantial increase in government support, the continued development of biotechnology industry, and the increasing adoption of molecular-based medicine. [1]. It has been pointed out in many sources that the nature of biomedicine has changed. It has transformed from laboratory based science to an

information science, science “in silico”. [e.g. 2, 3, 4], which means mainly the computerization of the research process.

Specialization to different research domains, fields and sub-disciplines qualifies biomedicine. As Buetow felicitously remarks each of these “speak its own scientific dialect”. Like in many other scientific fields, “big science” (i.e. big budget, big staff, big machines etc.) is a growing challenge to the discipline. Research equipment and technology are extremely expensive and these are factors which have been leading researchers to work on teams. Biomedicine, according to Buetow is a “team science”. It is typical of biomedical research teams that many research problems in order to be solved have to cross traditional discipline boundaries. [1].

Molecular medicine, a sub-discipline of biomedicine is a practice oriented, applied science and utilizes molecular and genetic techniques in the study of the biological processes and mechanisms of diseases. It is highly reliant upon the development of techniques and technology for acquiring data. [5]. Its final, practical task is to provide new and more efficient approaches to the diagnosis, prevention, and treatment of a wide spectrum of congenital and acquired disorders [6]. The nature of MM, like biomedicine in general is interdisciplinary, it could also be seen as a hybrid of biomedicine and molecular biology. Molecular biology in turn is based on the combination of biochemistry, cell biology, virology and genetics [7].

3 Information Environment and the Changing Research Process

We define information environment in this study as the entity of information objects as well as the tools and services needed to retrieve, manage and analyze them.

A large volume of data in combination with the diversity of data types is typical for MM information environment. The characteristic of the data is that it is rapidly expanding and ever-changing. [1]. Most of the research databases, like genomic and proteomic databases are commonly updated and globally shared. A yearly updated list of online molecular biology databases is found in the website of Nucleic Acid Research [8].The January 2007 edition contained almost 1000 databases [9]. The amount of data growth could be described by for example the situation of the GenBank, a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. It doubles in size about every 18 months. At the beginning of 2007, it contained over 65 billion nucleotide bases from more than 61 million individual sequences. [10].

What is even more challenging is that there is a need to integrate different kinds of data, e.g. to move between the biological and chemical processes, organelle, cell, organ, organ system, disease specific, individual, family, community and population. [1]. Like Butler notices, there are some disciplines which already have software that allows data from different sources to be combined seamlessly. For example, a gene sequence can be retrieved from the GenBank database, its homologues using the BLAST alignment service, and the resulting protein structures from the Swiss-Model site in one step. [11]

In parallel with the growth of data, the number of different tools, developed for data retrieval and analysis is growing. An actively maintained directory of bioinformatic links lists over 1000 web servers and other useful tools, databases and resources for bioinformatics and molecular biology research in 2006 [12, 13].

PubMed, the most important bibliographic database in biomedicine consisted in 2006 of 16 million references. The growth rate of the database is about 12 000 references every week, which means yearly over 600 000 new references. The growth curve of Medline, the main database in PubMed is illustrated in Figure 1. These lines describe the growth of traditional, published material, mainly in article format in biomedicine in a condensed way. It seems that inside the growing domain, there are some really “hot topics” where the amount of literature increase is extreme.

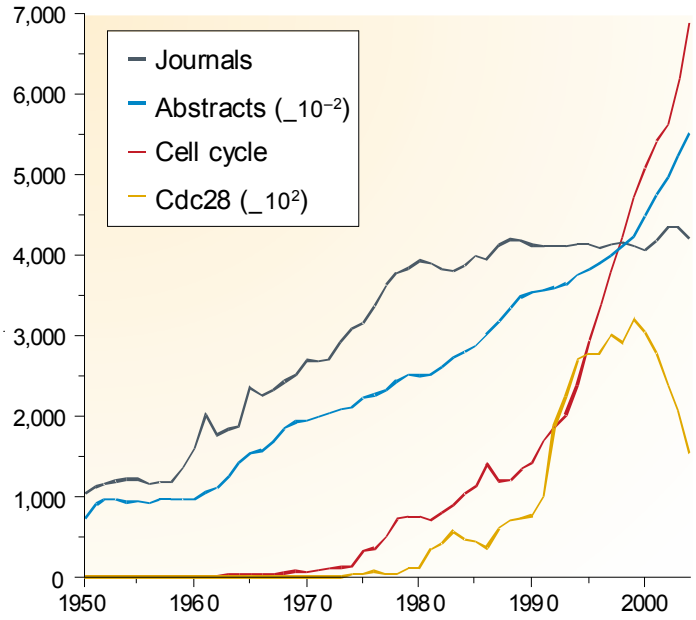

Figure 1: Growth of Medline: the number of journals, abstracts, papers on the cell cycle and papers on Cdc28 [14 published in Nature Reviews Genetics]

The typical features of MM information environment could be concluded as large volume and constantly growing number of data and published material, diversity of data types, great number of retrieval, analysis and other tools, interdisciplinary and globally shared and updated environment and team work. This is a fertile ground for the creation of new knowledge and inventions, but the lack of integration constitutes an increasing challenge to the development.

Cannata et al. have urged the organization of bioinformatics resources; data, knowledge, computational resources and services as a solution to the disintegration. They talk about "bioinformatics resourceome projcet" which would mean a process of creating a distributed system for describing resources, announcing their availability, and presenting this to the research community in an easy-to-navigate manner. The first step would be creation of an overall, distributed and collaboratively expandable ontology. [15, 16]. Mukherjea [17] has described the possibilities of using the semantic web in integrating the information resources. Grid technology has also been seen as a technical solution to the disintegration of data, information and tools. [1]

4 Results of the Survey

4.1 About the Research Unit and the Current Tasks

The research unit chosen as the case is situated in a Finnish research institute. As their aim, the unit declares to produce top level research in the molecular background of cardiovascular, immunological and neuropsychiatric diseases. At the moment of enquiry (February 2007), the unit consisted of 10 research groups with 83 researchers. From these 58 were PhD students and the rest were graduate students, group leaders and senior researchers We received totally 63 answers (75.9 %) to our web survey. 43 (68%) of those who responded were students and 20 (32%) were senior researchers, post docs and group leaders.

The research subjects of the groups were quite different, some of the groups concentrating on the genetic background of common diseases ("complex diseases"), some mainly to molecular genetics of monogenic diseases. There was also one bioinformatics group and one which specialized mainly in systems biology, one to quantitative genetics and a couple of groups mainly to the cell and molecular biology of certain diseases. We assume that the diversity of the research subjects caused some variety to reported work tasks between groups.

In the survey, all researchers were asked about their current work tasks and about information resources related to their current project or tasks and some information about usage of resources in general were asked. Respondents did get free spaces to write about their information resources, we gave only some examples for possible answers. We tried to get as broad a spectrum of possible resources, and did not want to limit or direct answers more than necessary. For current work tasks, we gave nine alternatives, from which it was possible to choose as many as were needed. Researchers were also able to add new tasks when necessary.

From the following figure (Figure 2.) the distribution of current tasks and their frequency among researchers is shown. The most common task among researchers was writing a report or an article, about totally 67 % of the researchers were doing it currently, the distribution among seniors and students is 70 % (seniors) and 67 % (students). Two-thirds of researchers were reading, 76 % of them were students. Of those working in the laboratory 74 % were students. It was more common (43 % of the respondents) to search information about literature from databases than data from data collections (25 %). Over one-third of the researchers were doing scientific computing. The researchers, who were studying the genetic background of "complex diseases" were practicing more scientific computing than most of the other groups. In two research groups where two-thirds (over 70 %) of all respondents answered that they were doing scientific computing.

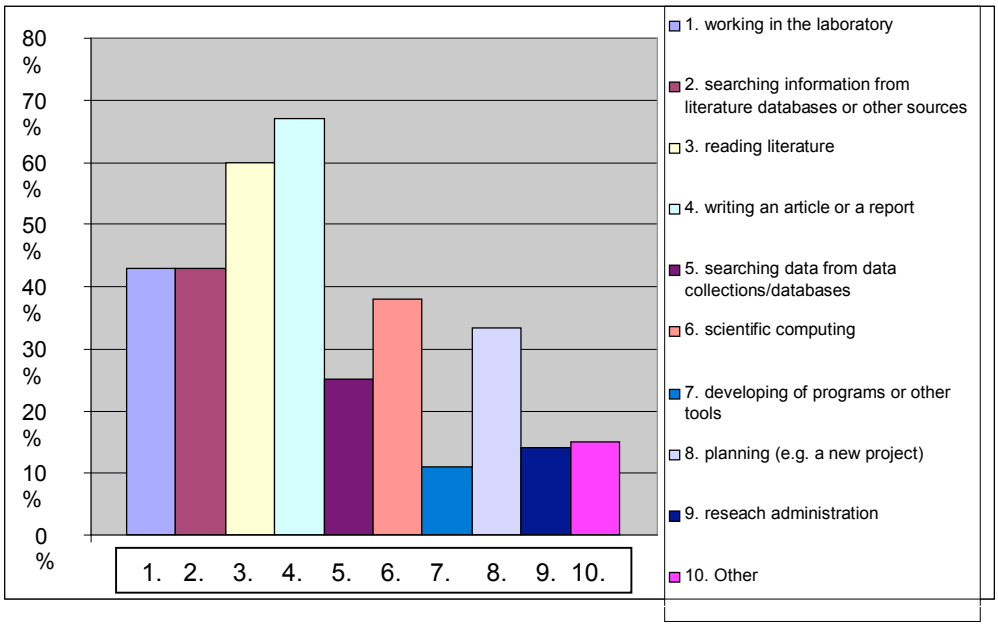

Figure 2: Current work tasks of researchers

4.2 Most Important Resources

When asked to choose at least the three most useful resources for their current research projects, doctoral and graduate (n=43) students named more resources than seniors and group leaders (n=20). PubMed got most references as the most useful resource in both groups. In the student's group UCSC Genome Browser was second and Google third as the number of references are concerned. In the seniors' group the ranking was contrary.

As their first information source 68% of the respondents named intranet/internet and in practice according to their answers, this means mainly PubMed and Google. 27 % of all researchers did prefer to contact a colleague or a supervisor. There seems to be no difference between students and seniors. In the feedback session researchers commented that the first information source depends on the nature of the issue: in practical questions and problems a colleague is preferred. It might also be possible, that some personal characters of the group leaders might at least partially explain the difference. The results indicate that in certain groups more researchers than on average in the groups favoured contacting a colleague in the first place. However, this is a speculation and needs to be observed more thoroughly.

When asked about which published material they use, the majority of respondents (53 %) answered that they use only or dominantly articles. 35 % of the researchers responded that they used articles and books equally and the rest 12 % named articles, databases and also some books.

When asked to name journals that researchers follow regularly, 23 % of the respondents reported that they do not follow any particular titles, rather their own topic from the literature databases. All of these respondents were graduate and doctoral students. Almost all graduate students belonged to this group.

91 % of the researchers said that they had used data collections during their current project. Those who did not use were juniors, who had recently started research work or researchers who were at the moment mainly working in the laboratory and writing articles. The problem with the reported data resources was that, because the question was open, researchers' answers were at very different levels. Some of them named quite general data collections, like "protein databases", or merely services or portals, like Entrez, while there were also respondents who used the detailed names of the databases or services. Totally 43 different data resources or services were named. The most common were NCBI and Entrez databases from National Centre for Biotechnology Information (by NIH and NLM) and UCSC Genome Bioinformatics resources, especially one tool, namely UCSC Genome Browser.

53 % or the researchers replied that they had used some research tools during their current project. The selection of tools and programs was also very diverse, from programs developed in their own laboratory to the commercial products. Totally 67 different tools were named. Students were naming more tools than seniors. The most often mentioned tool was Primer3, which is a PCR (Polymerase chain reaction) primer designer tool. The largest group in our survey as a whole was proteomics and sequence databases and analyzes tools.

It was noteworthy that tools for data mining seem to be common, but none mentioned text mining tools or tools for hypothesis creation. A tool called iHOP was familiar to the researchers. It's interesting, because it integrates gene and protein data from different collections with scientific literature.

Social bookmarking tools like Nature's Connotea were not named, neither any blogs. When asked why not, the answer in the feedback session was that they did not find those useful because their research problems were so specific: "they are only a waste of time". According to some opinions published in Nature researchers in general have not been eager to accept these tools because they might have been afraid of the poor image of the new tools and might have suspected the tools might damage their career [see 18].

4.3 Role of Scholarly Journals in the Research Process

Writing and publishing articles in scientific journals are seen as an important part of the research process in biomedical sciences and molecular medicine. This is shown among others in [14] but also in this present case study of the research group on MM in Finland. When asked about their current work tasks about 67 % of the researchers in the case group answered that they were writing an article or a report.

Since the research group constitutes of senior researchers as well as doctoral and graduate students this can be seen as a high percentage. The amount of work and the importance of article writing is also to be seen in the results presented in Table 1., where we were asking the researchers questions about their publishing strategy for the coming year. All of the senior researchers and group leaders are going to publish at least 1 article, most of them (87.5%) are going to publish at least two articles and 75% of the group leaders and 43% of the senior researchers are planning to publish at least three articles. We have counted as main authors, the first and second author and the last author. In this case study most of the senior researchers and group leaders are acting as supervisors to younger researchers, why it seems appropriate that the last author is counted as important.

Group leaders	100% minimum 1 article as main author (1 & 2 or last)
Senior Researchers	100% minimum 1 article as main author
Post doc	83,3% minimun 2 articles as main author
Doct.students	88% minimum 1 article as main author
Graduate stud.	73,3% minimum 1 article as main author

Table 1: Publishing strategy regarding scientific articles of researchers for the coming year of the researchers in MM

When looking at realized results (from 2006) for publications from the research group, 71 research articles in A-class journals and a total of 79 scientific articles were published. Of these 13 articles were in open access hybrid journals (applying some type of embargo) and 2 articles were in purely open access journals.

Regarding the choice of where to publish the researchers were presented the following criteria: impact, the speed of publishing, scope, open access or some other criteria, of which they were asked to name the one they regarded as most important.Impact was named as the most important by 58% of the researchers and scope by 39%. A few of the researchers named a combination of scope and impact. Open access as the main criteria was named by only 3% of the researchers.

The researchers were also asked to name journals with a suitable scope for publishing. On the top of the list of journals with suitable scope (Table 2.) was Nature genetics (named by 15). The impact factor for Nature genetics is also very high (25.797).

Journal title	**Number of nominations**	**Impact factor the journal**
Nature genetics	15	25.797
Human molecular genetics	11	7.764
Molecular psychiatry	10	9.335
American journal of human genetics	9	12.649
European journal of human genetics	6	3.251
Nature	6	29.273

Table 2: Top listing of journals with suitable scope for publishing

However, even though journals hold an established position in scholarly communication, there has appeared comments and viewpoints which have suggested that because scientific publications are slow and access to them is limited they act more as barriers to the development of new knowledge and science. [19].

In fact, traditional journals have very seldom made it possible to attach data files containing research data to the article. However, digital publishing and open access initiatives have opened up new possibilitities for scientific publishing (Björk 2007). In a study by Hedlund and Roos (2007) on publishing practices among biomedical researchers, the authors found that there is a growing rate of research publications in BioMed Central by Finnish researchers during the years 2003-2004. Cockerill & Tracz (2006) name fileds like bioinformatics, genomics and systems biology as possible success fields for open access.The initiative from the open access journal publishers BioMed Central is to put up a structured XML version of each full text article for data mining. There is also an increasing number of institutional repositories that allow researchers to upload data files linked to their published articles, which then serve as a possible source for data mining. Cockerill and Tracz (2006) argue that in the future the potential reader of a research article may not be only human beings but instead software agents looking for data to extracted and processed for a knowledge base. Therefore open access is important for work that involves multiple disciplines, as for example computer scientists, mathematicians and biologists collaborating in the areas of systems biology and bioinformatics.

5 Conclusions and Discussion

The information environment of researchers in MM could be summarized in the following diagram (Figure 3.)

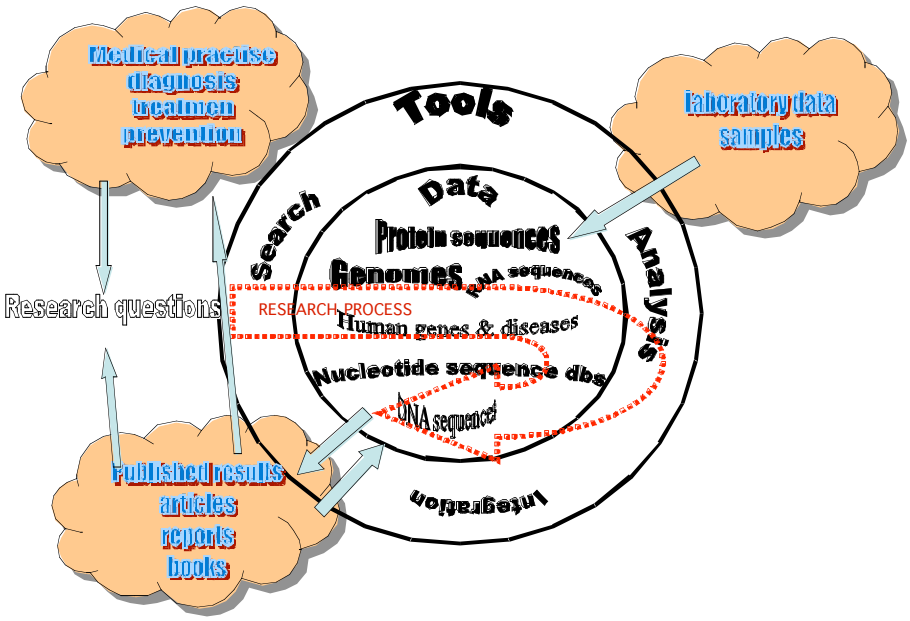

Figure 3: The research process and the information environment of molecular medicine

It can be concluded that access and use of data resources is an important and integral part of the research process in MM. The amount of different data collections, searching and analysis tools is huge. The disintegration of the environment seems also to be quite problematic.

We noticed that a more thorough analysis would be needed to make any conclusions about the relationship between the different work tasks in the research process and the used resources. We assume that many of the tasks might consist of several levels all of which might be worked out via different resources. The reason for this being for example in the varied complexity of the research problems.

The number of published articles is growing exponentially, especially in the "hot topics" of the domain. Researchers might find it difficult to follow even the development in their own research area. Maybe this is the reason why students do not follow particular journals, rather topics. The amount of literature is growing so fast that they are not able to do anything else than to follow the most recent and important articles from reference databases like PubMed. The disinterest to follow particular journals might also be due to the fact that they are not so well integrated into the domain yet, or it could be possible that their research subjects are so interdisciplinary that at least at the beginning of their career they are not able to follow any particular titles.

Journal publishing is still seen as the prominent way of distributing research results in molecular medicine. It has been shown in the case study that writing articles and reports is occupying the researchers as an important part of the research process. Even though many attempts to introduce open access, e.g. by providing institutional and national licences to cover authorship fees in BioMedCentral journals there still seems to be a strong reliance on traditional journals and especially journals with high impact.factors. Publishing in journals with high impact factor and the right scope is a strong base in the prevailing publishing strategy. However, it could be possible that the importance of traditional publishing channels and particularly articles might be on their way to change in the future if the text mining and hypothesis creation tools will be developed, and if the technical cyberinfrastructure with semantic web tools will be developed to integrate the environment. Open access will be helpful and a natural part of this development.

Notes and References

[1] BUETOW, KH. Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research. 2005:821-4.

[2] LENOIR, T. Shaping Biomedicine as an Information Science. *Conference on the History and Heritage of Science Information Systems*: Information Today 1999.

[3] LENOIR, T; ALT, C. Flow, Process, Fold: Intersections IN. In: Picon A, Ponte A, eds. *Science, Metaphor, and Architecture*. Princeton: Princeton University Press 2003:314-53.

[4] HINE, C. Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work. *Social Studies of Science* 2006:269-98.

[5] MACMULLEN, W. Information problems in molecular biology and bioinformatics. Journal of the American Society for Information Science and Technology. 2005;56(5):447-56.

[6] GOOSSENS; M. Principles of molecular medicine. New England Journal of Medicine. 1999;340(20):1601-2.

[7] MIETTINEN, R; TUUNAINEN, J; KNUUTTILA, T; MATTILA, E. Tieteestä tuotteeksi? Yliopistotutkimus muutosten ristipaineessa. Helsinki: Yliopistopaino 2006.

[8] Nucleic Acids Research. Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Alpha List. 2007 [cited 10 April 2007]; Available from: http://www.oxfordjournals.org/nar/database/a/

[9] GALPERIN, M Y. The Molecular Biology Database Collection: 2007 update. Nucleic Acids Research. 2007;35(Database issue):D3.

[10] BENSON, DA; KARSCH-MIZRACHI I; LIPMAN D J; OSTELL J; WHEELER D L. GenBank. *Nucleic Acids Research* 2007:D21-5.

[11] BUTLER, D. Mashups mix data into global service. Nature. 2006;439(7072):6-7.

[12] UBiC. NAR Web Server Issue (July 1, 2006) - UBC Bioinformatics Centre. 2007 [cited 10 April 2007]; Available from: http://bioinformatics.ubc.ca/resources/links_directory/narweb2006/

[13] FOX, J A; MCMILLAN, S; OUELLETTE, B F. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Research*: Oxford Univ Press 2006:W3.

[14] JENSEN, L J; SARIC, J; BORK, P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006 Feb;7(2):119-29.

[15] CANNATA, N; CORRADINI, F; MERELLI, E. A Resourceomic Grid for bioinformatics. Future Generation Computer Systems. 2007;23(3):510-6.

[16] CANNATA, N; MERELLI, E; ALTMAN, RB. Time to Organize the Bioinformatics Resourceome. *PLoS Computational Biology* 2005:e76.

[17] MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform. 2005 Sep;6(3):252-62.

[18] BUTLER, D. Science in the web age: Joint efforts. Nature. 2005;438(7068):548-9.

[19] INSEL, T R; VOLKOW, N D; LI, T K; BATTEY, J F; LANDIS, S C. Neuroscience Networks. PLoS Biology. 2003;1(1):e17.