OXFORD

*Briefings in Bioinformatics*, 20(4), 2019, 1215–1221

**doi: 10.1093/bib/bbx140**
Advance Access Publication Date: 30 October 2017
Paper

# A generally applicable lightweight method for calculating a value structure for tools and services in bioinformatics infrastructure projects

Gerhard Mayer, Christian Quast, Janine Felden, Matthias Lange, Manuel Prinz, Alfred Pühler, Chris Lawerenz, Uwe Scholz, Frank Oliver Glöckner, Wolfgang Müller, Katrin Marcus and Martin Eisenacher

Corresponding author: Martin Eisenacher, Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, Universitätsstraße 150, D-44801 Bochum, Germany. Tel.: +49 234 32-29288; Fax: +49 234 32-14554; E-mail: martin.eisenacher@rub.de

## Abstract

Sustainable noncommercial bioinformatics infrastructures are a prerequisite to use and take advantage of the potential of big data analysis for research and economy. Consequently, funders, universities and institutes as well as users ask for a transparent value model for the tools and services offered. In this article, a generally applicable lightweight method is described by which bioinformatics infrastructure projects can estimate the value of tools and services offered without determining exactly the total costs of ownership. Five representative scenarios for value estimation from a rough estimation to a detailed breakdown of costs are presented. To account for the diversity in bioinformatics applications and services, the

**Gerhard Mayer** is a PhD student in the Medical Bioinformatics group at the Medizinisches Proteom Center (MPC) within the Medical Faculty of the Ruhr-University Bochum (RUB) and works in the de.NBI network.
**Christian Quast** is a postdoc at the Max Planck Institute in Bremen, Germany. He is the project lead of the SILVA project and manages the releases as well as the software development. Additionally, he is heading the implementation of the UniEuk taxonomy framework.
**Janine Felden** is a postdoc at the MARUM—Center for Marine Environmental Sciences, University of Bremen. She is working as data and project manager with the Data Publisher for Earth and Environmental Science—PANGAEA.
**Matthias Lange** studied computer science in Magdeburg. Since his PhD thesis in 2006, he worked as Bioinformatician at the IPK-Gatersleben. His main interests are information retrieval and research data management.
**Manuel Prinz** is a bioinformatician in the Data Management and Processing IT group in the Department Theoretical Bioinformatics at the German Cancer Research Center (DKFZ) in Heidelberg.
**Alfred Pühler** is a senior research professor at Bielefeld University for Genomics of Industrial Microorganisms. He is also a coordinator of the German Network for Bioinformatics Infrastructure (de.NBI) and Head of Node of ELIXIR-Germany.
**Chris Lawerenz** is a group leader of the Data Management and Processing IT group in the Department Theoretical Bioinformatics at the German Cancer Research Center (DKFZ) in Heidelberg.
**Uwe Scholz** is a group leader of the research group Bioinformatics and Information Technology at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben and coordinates the plant service unit GCBN within de.NBI.
**Frank Oliver Glöckner** is a group leader of the Microbial Genomics and Bioinformatics Research group at the Max Planck Institute for Marine Microbiology Bremen and Professor of Bioinformatics at Jacobs University Bremen.
**Wolfgang Müller** is a group leader for Scientific Databases and Visualization at the Heidelberg Institute for Theoretical Studies, HITS.
**Katrin Marcus** is a director of the Medizinisches Proteom Center, Ruhr-University Bochum with special expertise in proteomics focusing on neurodegenerative and neuromuscular diseases. She also serves as a steering committee member and chair of the HUPO Brain Proteome Project.
**Martin Eisenacher** is a coordinator of the research unit Medical Bioinformatics at the Medizinisches Proteom Center (MPC) within the Medical Faculty of the Ruhr-University Bochum (RUB).
**Submitted:** 29 June 2017; **Received (in revised form):** 22 September 2017

notion of service-specific 'service provision units' is introduced together with the factors influencing them and the main underlying assumptions for these 'value influencing factors'. Special attention is given on how to handle personnel costs and indirect costs such as electricity. Four examples are presented for the calculation of the value of tools and services provided by the German Network for Bioinformatics Infrastructure (de.NBI): one for tool usage, one for (Web-based) database analyses, one for consulting services and one for bioinformatics training events. Finally, from the discussed values, the costs of direct funding and the costs of payment of services by funded projects are calculated and compared.

**Key words:** bioinformatics infrastructure; de.NBI; cost factors; value-influencing factors; value estimation for offered tools and services; service provision unit

## Introduction

The provision of an effective and sustainable computing and data infrastructure is seen as a prerequisite for further developing an efficient life science industry and harvesting their economic potentials [1]. This need was identified early in the past and has led to the foundation of national or transnational bioinformatics institutes like NCBI (1988), EMBL-EBI (1992) or SIB [2] (1998). In the past couple of years, many European countries installed national bioinformatics infrastructure projects, and the most prominent of these in Europe are DTL [3] (The Netherlands, https://www.dtls.nl, accessed June 2017), IFB (France, http://www.france-bioinformatique.fr, accessed June 2017), NBIS (Sweden, https://nbis.se, accessed June 2017), INB (Spain, http://www.inab.org, accessed June 2017) and the German Network for Bioinformatics Infrastructure—de.NBI [4] (Germany, https://www.denbi.de, accessed June 2017). Most of them are partner nodes in ELIXIR [5] (https://www.elixir-europe.org, accessed June 2017), a transnational European-wide distributed life science infrastructure project. These infrastructure projects provide data repositories, software tools and resources for the data management, analysis and interoperability to be used by life science projects producing or analyzing 'big data'. This also encompasses services for knowledge transfer such as training and consulting for enabling the users to efficiently use the provided resources. Often these resources are deployed in cloud systems to offer highly scalable and high-performance computing environments to the end user.

The costs for these services are either charged to the users or the services are offered for free, for example if they are funded by funding institutions. Even if they are offered free of charge, there is a desire of infrastructure providers to estimate the value of the offered tools and services. This is necessary to (1) plan the institutional budget to provide required personnel and technical infrastructure, (2) justify research funding and get long-lasting support for hosting services from research projects and (3) render the financial resources needed for the infrastructure transparent to the stakeholders and the general public [6]. Standardized value estimation is a basic requirement to realize funding or payment models toward self-sustainability in the long term. A virtual price tag for the used service will further increase the awareness of the value of bioinformatics work and the compliance of researchers to finance bioinformatics resources.

The publicly funded bioinformatics infrastructure projects are set to the technical and scientific provisioning and support of the offered services. Usually, they have no dedicated financial accounting department to determine the exact operational costs for the offered tools and services on the basis of a full-cost pricing [total cost of ownership (TCO)] model. Consequently, there is a need for a generally applicable lightweight cost model.

In de.NBI [7], two special interest groups (SIG2 'service and service monitoring' and SIG4 'hardware infrastructure and data management') have developed such a simplified method. Some of the details described may be specific to the de.NBI network institutions concerned (e.g. overhead rates for indirect costs) or to Germany [e.g. value-added tax (VAT) rate] but should in principle be transferable to other institutions and countries.

## Methods

For a complete full-cost accounting, all elements of costs according to a TCO model have to be included. In general, one can distinguish between fixed and running costs. Common elements are as follows:

- Computer hardware and software
  - Hardware costs (computer, printer, . . .)
  - License costs for software
  - Schedule of depreciation for all tangible goods
  - Utilization rate to account for idle time (mean % CPU usage, software usage days per licensed week, . . .)
- Development costs
  - Personnel costs for development, testing and implementation (even in a productive environment with fixed functionality at least the costs for security bug fixes)
- Operational expenses
  - Direct infrastructure costs (building, i.e. floor and office space, equipment, furniture, . . .)
  - Infrastructure maintenance costs (e.g. janitorial supplies or maintenance, repair and overhaul)
  - Consumption costs (e.g. cooling, heating, electricity, phone, office supplies and consumables such as paper, advertising flyers)
  - Connectivity/data transfer costs (network, Internet, especially for big data)
  - Personnel costs for administration and general support (data backups and recovery, personnel administration)
  - Personnel costs for help desk, maintenance, consulting and training
  - Support contracts for hardware and licensed software
  - Costs for access to journals and books
- Long-term expenses
  - Replacement costs (estimation of the costs for replacing defective or old hardware)
  - Upgrade/scalability expenses (costs for nonlinear growth of service volume)
  - Decommissioning (e.g. for hardware at end of lifetime)

Other bioinformatics tool and service providers like the EBI (European Bioinformatics Institute) contracted a consultancy company to estimate the costs and the generated value of their institution [8]. For de.NBI, a lightweight minimal consensus value structure model was developed, which makes some simplifications to the TCO model. For instance, because the direct infrastructure and infrastructure maintenance costs are considered as

a lump sum ('overhead costs') in the grants of funding organizations (currently 20% for funding of the BMBF, the German Federal Ministry of Education and Research), they are also included as a lump sum in our value structure (cost Model 1). Considering that a scientist has to pay for a service via invoice (cost Model 2), this lump sum is usually higher (up to >70% depending on the research institution or organizational structure) plus minimum profit (e.g. 4%) plus VAT (currently 19% in Germany).

Because fixed costs scale up in steps and running costs scale up linearly, and for allowing the inclusion of further improvements of the offered tools and services, the value structure model should reflect such scalability issues. Therefore, we defined different scenarios for the value structure model:

- Scenario 1: Value of the status quo, where neither tool or service improvements nor growth of usage volume is taken into account
- Scenario 2: Value, which includes tool/service improvements, but no growth
- Scenario 3: Scenario 2 plus growth of the usage volume
- Scenario 4: Scenario 3 plus the retrospective development costs for the tools (may have been financed by third-party projects)
- Scenario 5: Scenario 4 plus the expected future hardware exchange or replacement and future development costs

Scenarios with higher ordinal numbers have the potential to reflect the TCO model better, but more assumptions may be necessary. Furthermore, a mix of scenarios is imaginable, e.g. a scenario incorporating expected future hardware costs but without growth of usage volume and without incorporating retrospective development costs.

Within de.NBI, we follow a best practice for estimating the Scenario 1 value of a de.NBI service. We defined service-specific 'service provision units', which are the basic units of value calculation. This could be for example 'one database query' or 'one statistical analysis day' or 'one training day'. For each offered tool/service, the specific underlying value considerations are explicitly formulated together with the related assumptions ('value influencing factors'). Only factors with financial influence are taken into account. Other measures of 'value' that are typically derived using usage statistics, like for instance citations and value perceived by the users, etc., are regularly monitored inside de.NBI but not considered here. Such factors with financial influence could be, for example, 'personnel costs for one day database maintenance assuming 100 queries per day' or 'personnel costs for a statistician assuming one analysis takes three days'; for more detailed examples, see section below. The value-influencing factors are adapted regularly (e.g. every 6 months), so that with changing knowledge, e.g. about personnel salary increase, or growing experience about the assumptions, e.g. the usage volume or average usage, the calculated values converge more and more to the 'true' values. For calculating the personnel costs, as a basis the yearly adapted personnel staff appropriation rates of the DFG are taken (the German Research Foundation http://www.dfg.de/formulare/60_12/60_12_en.pdf, accessed June 2017). If using the average scientist salaries of the institution or the work group or even the known person performing a service, this value-influencing factor will be more exact. Optionally, a 'scaling limit' for the usage volume can be specified, up to which the used value structure is reasonable.

## Results

In the following, we show for the four service examples, 'tool usage', 'web-based database query', 'bioinformatics consulting' and 'training' how their value can be determined. This is by default done for Scenario 1. This scenario does not consider the cost to implement new features, fix security bugs and to improve the overall performance of the software. Especially, this scenario does not consider the lifetime of the software (operating system, libraries or other tools) that is required to offer the service. Some of the examples nevertheless consider also hardware exchange, which is part of Scenario 5.

According to the DFG personnel staff appropriation rates 2017, a postdoctoral scientist (PostDoc) or comparable costs €68 400 in average per year, and a nonacademic technical staff member costs €47 100 in average per year. Assuming 220 working days per year (365 without weekends, working/public holidays, average sick leaves, etc.), the PostDoc working day value is €310.91, and the technician working day value is €214.09.

In the Supplemental Material, we provide Excel files for the examples to provide a starting point for own calculations.

### Example 1: Tool usage (analysis via stand-alone executable)

The service provision unit is 'one analysis via tool usage', i.e. one analysis performed with a tool, which is installed on a service provider's server and that is remotely accessible by the user. The tool usage thus comprises the upload of the data, running the tool and returning the results to the user. We assume that the tool is used this way five times per week.

As value-influencing factors, we consider a fraction X of a full-time PostDoc maintaining the tool in its environment (bug fixing, user help desk, e.g. supporting upload/download). Assuming there are five tools to maintain or five other services done by the PostDoc, a first assumption for the fraction X is 20%. We also consider 20% of a technical maintenance person for the server, the remote access and the network.

More formally, the value-influencing factors of 'one analysis via tool usage' are: (20% of PostDoc week + 20% of technician week) divided by 5 [analyses per week] + 20% overhead for indirect costs.

With numbers:

$$\frac{(0.2 \cdot €1554.55 + 0.2 \cdot €1070.45)}{5} \cdot 1.2 = (€62.18 + €42.82) \cdot 1.2 = €105.00 \cdot 1.2 = €126.00$$

As an extension of Scenario 1, we also include costs for hardware renewal. We assume that renewing the dedicated middle-range server hardware costs €3300, and it is renewed after 3 years. Thus, it serves for 780 analyses (3 * 52 weeks * 5 analyses per week) and contributes €4.23.

With numbers:

$$(€62.18 + €42.82 + €4.23) \cdot 1.2 = €109.23 \cdot 1.2 = €131.08$$

A value calculation for Scenario 2 would include efforts to add tool functionality and respective documentation. Depending on the tool, the research domain and the implemented complexity, for example personnel costs (e.g. 2–12 person months), the respective proportional prospective costs for development hardware (client and possibly server computers) plus possibly proportional costs of software licenses for the development environment would be incorporated before applying the overhead for indirect costs.

Analogously, a value calculation for Scenario 4 would include the retrospective development costs for the tools (this is

not done here because many tools provided as service have been financed by previously funded third-party projects).

Analyses by tool usages within workflow systems such as Galaxy or KNIME or even in a cloud system as Docker Container or other mechanisms have to be calculated with other assumptions for the 'value influencing factors' or even with differing 'value influencing factors', and therefore, their value cannot be directly derived from this example.

### Example 2: Web query (analysis via browser)

The service provision unit is 'one analysis against a database' using a Web-based service. The factors to estimate the value of a single 'analysis' include the value to maintain the Web presence and the value to run the analysis. Additionally, costs to create or to license the database may have to be considered. Each of these factors includes both personnel as well as hardware-associated costs. Further, it is assumed that the hardware is used exclusively to create the database and to provide the service, while the personnel may not be working full-time on the service. The following example is based on the SILVA Web service [9], a service for the analysis of ribosomal RNA sequences.

The personnel investment in the daily operation of the Web service is rather small even for a large number of analyses in case these jobs run fully automated. Nevertheless, it requires technical and scientific staff to operate the service. The technician is mostly concerned with the operation of the computer hardware, the network and the operating system implementing security fixes and system updates. These tasks take about 40% of a full-time position and, according to the above mentioned technician salary; this is €18 840 per year. A software developer is required to maintain the custom software developed to run this service, mainly to adapt it to the ever-changing Web environment (Web browser, and the fast evolving HTML, HTTP, SSL and JavaScript standards and associated libraries). This takes roughly 55% of a full-time PostDoc position, or €37 620 per year. Additionally, a domain expert is needed, mostly to support users, but also to supervise the service and to proactively check for issues that may arise during the operation of the service. These tasks cannot be accomplished by the software developer because of the lack of domain knowledge. This takes up to 33% of a full-time PostDoc position, €22 572 per year. In total, €79 032 has to be spent on wages per year to operate this service.

The cost of the corresponding hardware (Web server as well as associated storage server) is €10 000. For fail-over safety and to reduce the time the users have to wait for their results, a redundant design of both components is necessary. Together, the cost of this setup is €20 000. As previously described, these costs are written off over 3 years and the costs per year sum up to €6, 666.67.

The personnel and hardware costs to operate the analysis service sum up to €85 698.67 per year. This sum, however, does not yet include any license fees for third-party software, the preparation of the reference database and other resources nor does it include the aforementioned overhead factor of 1.2.

The reference databases used for the analyses have to be considered as indirect costs, as the databases have to be regarded as a fixed external resource used during the analyses. These costs may either result from licensing the databases or from creating and curating the databases. The calculation below assumes that the databases are maintained by the provider of the analysis service. It includes the computational costs as well as the personnel costs to supervise the preparation of the production databases. However, it does not include any costs for past development of the software to create the databases nor does it consider costs to further develop this software.

In most cases, a compute cluster is needed to create the databases. The specification of the compute nodes highly depends on the data that needs to be processed. In case of the SILVA databases, 20 computes nodes, each equipped with 12 CPU cores and 64 GB of memory are required to process the raw data in 2 weeks. Each of the compute nodes costs €4500; altogether they cost €90 000 depreciated after 3 years (€30 000 per year). We assume that this hardware is exclusively used for SILVA. In case that one can lease the idle time, one must take into account the utilization rate.

The preparation of the databases has to be supervised and an additional 2 weeks have to be invested to prepare all the data. Including the maintenance of the software, 5–6 weeks have to be spent to prepare a single release of the databases. Updating the databases three times a year, this accounts for 45% of a full-time PostDoc position, or €30 780 and 20% of a technician position (€9420). After the data have been processed, it needs to be curated by a domain expert, which takes another 2–3 weeks per release. In between the releases, the domain expert has to continue to curate the data and taxonomy and to incorporate up-to-date information published by the research community in preparation of the next database release. Over the course of a year, 66% of full-time PostDoc position (€45 144) has to be invested. In total, €85 344 has to be spent on personnel for the preparation of the reference databases required by the Web service.

Overall, the operation of the Web service and the preparation of the database cost about €241 251.20 including personnel and hardware costs as well as a 1.2 overhead factor.

To calculate the cost for the user to run a single analysis, the overall costs have to be divided by the number of jobs served per year. Each compute node can serve a maximum of 105, 120 jobs, if a job takes on average 5 min to process.

Having redundant compute nodes, in theory 210, 240 jobs can be served per year reducing the cost per analysis to about €1.15. However, this assumes that the compute nodes are used 100% of the time on every day of the year, which is a theoretical assumption, as it includes weekends and public holidays where users show less activity. The price of a single analysis must rather be based on the expected number of jobs a service serves per year than the maximum number of jobs a service may serve. As an example, the SILVA project served 86 560 jobs in 2016. Using this number and anticipating an increase in demand of 5% to about 91 000 for 2017 increases the price per analysis to €2.65.

In summary, for the SILVA project to serve 91 000 jobs and to continue to update the reference databases in 2017, two PostDoc positions (€136 800) and 60% of a technician position (€28 260) are required. Additionally, €36 666.67 has to be invested into the hardware. The overall formula to calculate the cost of the SILVA Web service is:

$$\frac{(2 \times €68400(\text{PostDoc}) + 0.6 \times €47100(\text{Technician}) + €36666.67(\text{hardware})) \times 1.2}{91000(\text{jobs})} = €2.66$$

### Example 3: Bioinformatics consulting (here: bioinformatics for proteomics)

The term 'consulting' is used here in contrast to 'analysis' because in performing this service, we support the service users to decide which analysis workflow in which workflows system should be used and or which (open-source/free-to-use or commercial) tools may be used for analysis by inspecting their experimental hypothesis, planned experimental design, existing (mass spectrometry) technologies and—potentially—already

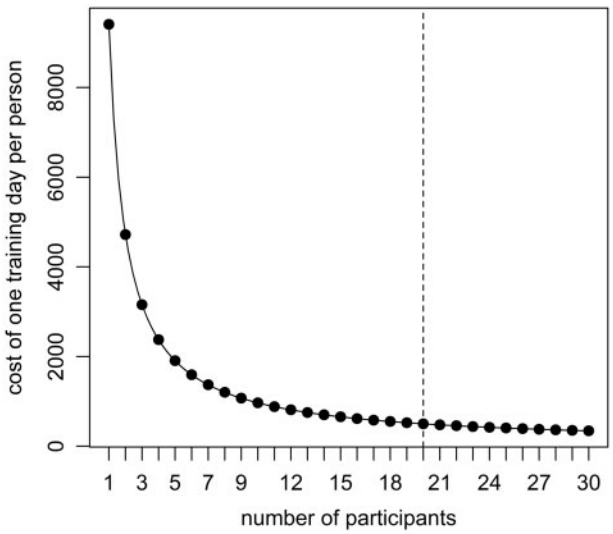


**Figure 1.** Dependence of the 'per-person value' for 1 training day from the number of participants.

existing data sets. A 'Bioinformatics analysis' service would have more value influencing factors such as software licenses.

The service provision unit is 'one bioinformatics for proteomics consulting'. It is assumed that a default 'bioinformatics for proteomics' consulting lasts 2 days including communication with the user, collection of existing hypothesis and experimental design, preparation of data files, if already produced, and literature/online search for existing public data. The value-influencing factors are the scientific personnel for a PostDoc and the renewal costs for a middle-range desktop computer plus monitor for these 2 days. We assume that a mid-range desktop computer for the consulter plus monitor, which is depreciated over 3 years, costs €1000 and therefore €1.52 per working day. Additionally, a factor of 1.2 is used to cover the indirect costs (overhead allowance for consumption costs). Therefore, one bioinformatics consulting (of the default 2 days duration) has a value of

$$2 \times (€310.91 + €1.52) \text{ x } 1.2 = €749.83,$$

i.e. €374.92 per consulting day

This shows that the personnel costs are the main value-influencing factor for bioinformatics consulting.

### Example 4: de.NBI Training

By extending the assumptions made in Example 3, for the bioinformatics training course, we additionally assume that a mid-range server has original costs of €3000 and therefore costs €4.55 per working day. Thin clients, which are used during training events, would be calculated with a purchase price of €300, what equals to costs of 300/660 = €0.45 per training day. But we assume here that—in contrast to the desktop computers and servers—these thin clients are exclusively used for 3 preparation days and 6 training days per year, so that the utilization rate must be taken into account. Then, one such thin client costs €300/27 = €11.11 per day used, if a depreciation period of 3 years is assumed.

The service provision unit is 1 training day and includes as value-influencing factors the time for preparation and for conducting the actual training, the hardware usage for the training preparation, the teaching hardware and its preparation, flyers and posters for advertising the training event (€100), printed handouts (€4 per participant) and small snacks (€10 per participant) but no travel and accommodation expenses for the participants. In addition, we assume that the training takes place at the institution of the trainers, so that there are no travel and accommodation expenses for them, and the room for the training is provided for free. The number of trainees is assumed as 20, and we assume that four scientists and one technician are involved in the preparation of the training event. The preparation time for the four scientists is assumed to be 5 working days for each of the scientists and 1 working day for the technician (to install the needed software to the teaching hardware, a mid-range server, used during the actual teaching and to the 10 thin clients available for usage by the trainees). For that preparation, the scientists use mid-range desktop computers. Again, we calculate with a factor of 1.2 for the consumption cost overhead.

Therefore, we can calculate the personnel costs as four scientists times 5 preparation days per scientist times €310.91 per preparation day plus one technician times 1 preparation day per technician times €214.09 per preparation day, which totals to 20(days) x €310.91 + €214.09 = €6432.29 for all 5 preparation days, and personnel costs of 4 (PostDocs) x €310.91 = €1243.64 for the teaching day, totalling to €7675.93 of personnel costs.

In addition, we can calculate the hardware costs during the preparation phase as four scientists times 5 preparation days per scientist times €1.52 per preparation day plus one technician times 1 preparation day per technician times (€4.55 (server) + 10 (thin clients) x €11.11 = €115.65) per preparation day. This totals up to 20 x €1.52 + €115.65 = €146.05 for all 5 preparation days.

For the 1 server and 10 thin clients used at the teaching days, the costs are calculated as €4.55 (server) + 10 (thin clients)x €11.11 = €115.65.

Then, the overall hardware costs for preparation days and training days are summed up to €146.05 + €115.65 = €261.70.

Finally, for 20 participants, there are other costs of €380 for advertising, handouts and snacks.

For a training day with 20 participants, the total costs with taking the 20% overhead into account equal to a sum of

$$(€7675.93 + €261.70 + €380.00) \text{ x } 1.2 = €8317.63 \text{ x } 1.2 = €9981.16,$$

which equals to €499.06 for 1 training day for each of the 20 participants.

Figure 1 shows that because of the fact that the total costs for a training day are dominated by the fixed personnel costs for the training preparation, the cost for a training day per participant is given by a hyperbolic curve.

Larger training events such as summer schools are separately funded in de.NBI; for those, further value-influencing factors such as presentation rooms, travel, accommodation and lunch/dinner costs for participants and trainers have to be considered.

## Taking the funding model into account

Besides using the determined values to calculate the financing needs for direct funding (funding Model 1, 'infrastructure funding'), one can also estimate the values in case that the infrastructure is paid with compensation fees by users (funding Model 2, 'contract research').

**Table 1.** Estimated rounded values of direct infrastructure funding (with an overhead factor of 20%) and contract research (with an overhead factor of 60%) for the four examples described in 'Results' section

| | Funding Model 1: Direct infrastructure funding (20% overhead) | Funding Model 2: Contract research (60% overhead + 4% minimum profit + 19% VAT) | Factor Model 2/Model 1 |
|---|---|---|---|
| Example 1: Tool usage | €126.00 | €207.92 | 1.6501 |
| Example 2: Web query | €2.66 | €4.39 | 1.6501 |
| Example 3: Bioinformatics consulting | €374.92 | €618.66 | 1.6501 |
| Example 4: 1 Training day | €499.06 | €823.52 | 1.6501 |

*Note:* The assumption is a VAT of 19% and a minimum profit margin of 4% for contract research.

These two funding models differ by requiring different overhead costs (20% for infrastructure funding, 70% or more for contract research). For the contract research, a minimum profit (to avoid unfair competition with the private sector) and VAT need to be added.

In Table 1, the estimates for the two funding models are compared for the four examples described above. We calculated

$$\text{costs x overhead factor (20\%)}$$

for funding Model 1 and

$$\text{costs x overhead factor (60\%)x minimum profit margin (4\%)}*\text{VAT(19\%)}$$

for funding Model 2, i.e. we assumed an overhead of 20% for funding Model 1 and 60% for funding Model 2, so that the estimated value for funding Model 2 is by a total factor of about 65% higher than for funding Model 1:

$$\frac{1.6 \text{ x } 1.04 \text{ x } 1.19}{1.2} = 1.65$$

## 5. Discussion

The presented value structure model does not automatically imply that an infrastructure will charge its users for the tools and services provided. A possibility to avoid charges is to apply for research grants together with researchers that want to use the infrastructure. Beyond that, a wide range of financing models is conceivable, such as 'charge all users', 'charge only commercial users' or 'support the whole research community via infrastructure funding'.

The abovementioned value components have been collected from a scientific and IT perspective only. When charging all users with payments per service is considered, diverse challenges from a financial perspective have to be solved: Who issues the legal invoices? Who tracks payments? How is the risk of non-payments calculated? Who takes care of accounting? How is the money flow organized? Can the bank account of the service providing institution be used or is an own legal entity (like a company or an association) necessary? How high are taxes in both cases? Who files the tax declaration? Solving these challenges has to be incorporated also into the respective value structure (and can significantly increase the payment costs).

An important aspect is the question on liability up to penalties for nonperformance for the tools and services offered. In theory, potential compensations for delayed or nondelivery of services can account for rather large amounts of money. In a first, maybe naïve, approach, we assume that the risk can be minimized by adding the respective disclaimers in the terms of use of the services. Owing to the complexity of the issue, we consider a detailed discussion of the problem beyond the scope of this article.

Even if tools and services are not charged to the users, it is reasonable to indicate them the value of the offered service. This helps to increase the awareness of the users that these services are not free-to-use.

For other scenarios additional challenges arise, e.g. for Web services, one has to cope with the ever-evolving HTML, CSS and JavaScript standards. Web browsers implement these new standards and at some point will stop supporting the old standards, leaving the Web service inaccessible to the users. Other problems are security bugs, which leave the Web service and the user data vulnerable to attacks. However, the largest problem when deploying software on the Web is the lack of long-term support implementing new Web standards, for older versions of the software.

As an example, the SILVA website is implemented as an extension of a content management system (CMS). The version of the SILVA CMS extension that has been implemented in is no longer supported, which means that the SILVA extension has to be completely rewritten for newer versions of the CMS. The effort to rewrite the CMS extension is far greater than has been accounted for in Example 2. It is hard to estimate the exact maintenance cost, as it is hard to estimate such breaking changes in the environment in which a Web service runs. However, it exceeds the 55% of a software developer accounted for in Example 2.

Highly significant, but less frequently used Web services will be presumably more expensive per Web query than our SILVA example. One needs always a fixed amount of money for holding available such a service. That fixed amount is mainly caused by the personnel costs for the maintenance of the (Web) software, for the curation of the data and for user support. The variable amount, which increases with the number of service users, is the mainly increased expense for the support and for a bigger and/or more powerful hardware. Less utilization of a service (because of a smaller scientific community) means higher costs per Web query for offering such a service but is of course not correlated with higher scientific impact.

**Key Points**

- A lightweight model to estimate the value structure of bioinformatics tool usage, services and training was described.
- The value model depends on assumptions made for each of the five defined scenarios. To demonstrate the application of the value model, four examples for the simplest Scenario 1 are given.

- With increased experience, the necessary assumptions reflect more precisely the reality, and therefore, the estimated values converge more and more to the real costs.
- The values should be communicated to the user community to increase their awareness that the provision of bioinformatics services must be acknowledged and rewarded.
- The value structure developed provides arguments for ensuring long-lasting support from funding organizations.

## Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

## Funding

The German Federal Ministry of Education and Research (BMBF). The BMBF grant de.NBI—German Network for Bioinformatics Infrastructure (FKZ 031 A534 A to G.M., FKZ 031A536A to U.S., FKZ 031A532A and B to A.P. and FKZ 031A539B to J.F.). The funding of Martin Eisenacher is related to PURE and VALIBIO, projects of Northrhine-Westphalia. The Max Planck Society (to C.Q. and F.O.G.). The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben (to M.L.).

## References

1. Robbins RJ. Bioinformatics: essential infrastructure for global biology. *J Comput Biol* 1996;**3**(3):465–78.
2. Stockinger H, Altenhoff AM, Arnold K, *et al*. Fifteen years SIB Swiss Institute of Bioinformatics: life science databases, tools and support. *Nucleic Acids Res* 2014;**42**(W1):W436–41.
3. Eijssen L, Evelo C, Kok R, *et al*. The Dutch Techcentre for Life Sciences: enabling data-intensive life science research in the Netherlands. *F1000Res* 2015;**4**:33.
4. Pühler A. German Network for Bioinformatics Infrastructure – de.NBI. Germany: BMBF, 2016, 8–13. https://www.systembiologie.de/
5. Crosswell LC, Thornton JM. ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol* 2012;**30**(5):241–2.
6. Chang J. Core services: reward bioinformaticians. *Nature* 2015;**520**(7546):151–2.
7. Tauch A, Al-Dilaimi A. Bioinformatics in Germany: toward a national-level infrastructure. *Brief Bioinform* 2017. doi: 10.1093/bib/bbx040.
8. Beagrie N, Houghton J. *The Value and Impact of the European Bioinformatics Institute—Full Report*. Salisbury, UK: Charles Beagrie Ltd.; Hinxton, UK: EMBL-EBI, 2016, 1–96.
9. Quast C, Pruesse E, Yilmaz P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**(D1):D590–6.