Genes Nutr (2009) 4:113–122
DOI 10.1007/s12263-009-0123-8

ORIGINAL PAPER

Owner controlled data exchange in nutrigenomic collaborations: the NuGO information network

Ulrich Harttig · Anthony J. Travis · Philippe Rocca-Serra · Marten Renkema · Ben van Ommen · Heiner Boeing

Received: 9 December 2008 / Accepted: 16 April 2009 / Published online: 30 April 2009

Abstract New 'omics' technologies are changing nutritional sciences research. They enable to tackle increasingly complex questions but also increase the need for collaboration between research groups. An important challenge for successful collaboration is the management and structured exchange of information that accompanies data-intense technologies. NuGO, the European Nutrigenomics Organization, the major collaborating network in molecular nutritional sciences, is supporting the application of modern information technologies in this area. We have developed and implemented a concept for data management and computing infrastructure that supports collaboration between nutrigenomics researchers. The system fills the gap between "private" storing with occasional file sharing by email and the use of centralized databases. It provides flexible tools to share data, also during experiments, while preserving ownership. The NuGO Information Network is a decentral, distributed system for data exchange based on standard web technology. Secure access to data, maintained by the individual researcher, is enabled by web services based on the the BioMoby framework. A central directory provides information about available web services. The flexibility of the infrastructure allows a wide variety of services for data processing and integration by combining several web services, including public services. Therefore, this integrated information system is suited for other research collaborations.

Keywords Nutrigenomics · Data management · Data integration · Distributed information system · Web services

Abbreviations

NIN	NuGO Information Network
DBMS	Database management systems
LIMS	Laboratory information management systems
DNS	Domain name service
NBX	NuGO Black Box
IPR	Intellectual property rights

U. Harttig (✉) · H. Boeing
Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany
e-mail: harttig@dife.de

A. J. Travis
University of Aberdeen Rowett Institute of Nutrition and Health, Greenburn Road, Bucksburn, Aberdeen AB21 9SB, UK

P. Rocca-Serra
EMBL-EBI, Wellcome Trust Genome Campus, Cambridge Hinxton CB10 1SD, UK

M. Renkema
Topshare International BV, PO Box 240, 6700 AE Wageningen, The Netherlands

B. van Ommen
TNO Quality of Life, PO box 360, 3700 AJ Zeist, The Netherlands

Introduction

The emergence of new analytical technologies in conjunction with high-throughput tools allows scientist to ask questions and conduct experiments of new scopes and quality. The new technologies not only open up new research possibilities, but also require to change the ways how the acquired data are managed and processed. The transformation of the obtained data into information and, through integration with other information, into

knowledge, always a key to success in science, requires more than ever a determined and cross-discipline approach to cope with associated challenges.

The successful sequencing of the human genome by the Human Genome Project [19, 32] is the prime example of a large scale effort driven by technological advances in high-throughput gene sequencing, computing technology and specifically biocomputing as the bioinformatics tool for assembling of sequences, discovering and annotating genes and proteins [4, 8] This example shows the importance of cooperation and the usefulness of a well-structured technological infrastructure to handle and integrate large amounts of data.

As in other life science areas, the genomics and related 'omics' technologies (transcriptomics, proteomics, metabolomics) have recently been introduced in nutritional sciences [9], stimulated by the success of the Human Genome Project. They promise to help nutritional sciences to adopt the new discipline of nutrigenomics that ultimately strives to develop a systems biology view of the relationship and interaction between diet, genes and human health [18, 20, 23]. To support this adoption, NuGO, the European Nutrigenomics Organization (http://www.nugo.org), was established. NuGO is formed by 23 academic partner organizations in ten European countries, with more than 750 researchers involved. NuGO is funded by the European Commission as a Network of Excellence under the "Food Quality and Safety Priority" of the Sixth Framework Programme. NuGO aims to develop and integrate genomic technologies for the benefit of European nutritional science, facilitate the application of these technologies in nutritional research world-wide, create the world-leading virtual centre of excellence in nutrigenomics and train a new generation of European scientists to use post-genomic technologies.

This paper describes the concept that NuGO has developed to tackle the data exchange and data integration challenges that arise from creating a nutrigenomics research network of excellence and thus facilitate multi-center and multi-omics layer collaborations with complete maintenance of local data ownership. We also describe the steps taken within NuGO towards the implementation of a sustainable and flexible infrastructure for exchanging and integrating nutritional and nutrigenomics data. In this work, the general applicability to other scientific collaboration was considered. Finally, we invite others to join this initiative.

General strategies of data handling

A prerequisite of data exchange is a suitable organization of the data at the source, which also determines what types of access to data are feasible.

Our analysis of current procedures shows that a widely employed strategy for local data organization is to store data files on an ad hoc, as-needed basis without a specific storage management system. This strategy is centered on individual data files and uses the basic file handling capabilities of the computer's operating system as basis. Modern software tools such as indexing programs can help with the search and retrieval of data from such data collections. However, such structures are prone to problems of data integrity (files can be deleted or altered easily), data provenance (the source and alterations to data items are not recorded) and the lack of suitable annotations, which makes data exchange and further data processing difficult.

Database management systems (DBMS) are designed to support the organization of data and are therefore, a preferable choice for organizing scientific and nutrigenomics data. DBMS provide means to efficiently store, index and retrieve data as well as mechanisms to properly preserve the integrity of data, to record alterations and to backup data. Access mechanisms control who and when data can be accessed and contribute therefore to the integrity and the privacy of data. Laboratory information management systems (LIMS), commercial systems as well as academically developed systems such as BASE2, an open-source, generic system for microarray data management [28], exploit these strengths of database management systems. The structured storage scheme enabled efficient data retrieval and export for the purpose of exchanging data. The inherent communication capabilities of databases is another advantage for data exchange.

A prerequisite for turning stored data into information and enable intelligent data exchange is the proper annotation of data. Efforts are underway within NuGO to implement capabilities to annotate experiments with information relevant to nutritional science by using a specific controlled terminology, the NuGO-CV. This semantic layer is being combined with an accessible syntax using MAGE-TAB/ISA-TAB formats. While MAGE-TAB format is used for annotating microarray data [25], ISA-TAB [30] also allows reporting of multiomics studies.

General strategies for data exchange

Different scenarios are available for the exchange and distribution of data within a project or collaboration. For our analysis of potential strategies, the characteristics of the following four major scenarios were investigated: ad hoc distribution, central database, database federation and distributed, and web-based information system (Fig. 1). Each scenario has different strengths and weaknesses which makes them a suitable choice under specific circumstances.

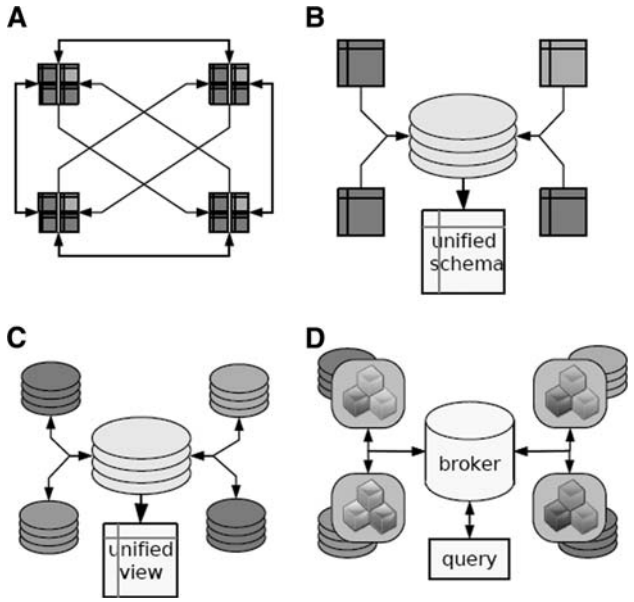

Fig. 1 Schematic representation of the major data exchange and data distribution schemes **a** ad hoc exchange, **b** central database, **c** database federation, **d** integrated web service based information system

Ad hoc distribution

A widely used strategy is the distribution of data files and information via e-mail. This strategy is convenient and flexible, since it allows ad hoc connections between different and varying partners. It is easy to implement, since the approach requires only the ubiquitous email connection. The strategy, though, has severe limitations when a project consist of several partner that all contribute with data to the project. If this is the case, multiple data transfers are necessary for data distribution and the system becomes quickly difficult to manage when data changes (Fig. 1a). Moreover, there is no mechanism to control the data integrity and no mechanism of recording the changes made to specific data items, so data provenance will become an issue in this setting.

Centralized database

For a project dedicated to a specific scientific question or a specific area of information, a common strategy is to establish a central data storage facility based on a database management systems (DBMS). A database schema tailored to the requirements of the project is generated and data items are stored according to this schema. The participants of the project enter their data into the common structure and have, within the limits and specifications of the database management system employed, a common view of and access to the data (Fig. 1b). The most demanding tasks in such a setup are (a) to integrate the projects in different sub areas by capturing the project data and information flow into a suitable database schema and (b) the installation of user-friendly facilities for data entry for different aspects of the project. Examples are large public database projects like the Protein Data Base, PDB [3] or smaller, local LIMS systems. Once entered into the data store, the original creator of the data has no further control over data processing steps, except through the agreements made for the purpose of the project.

Federation of databases

A federation of databases is commonly employed in integrated projects. In a federation, local databases are created and maintained by the project partners and store partner specific data (Fig. 1c). These databases are then organized in a 'hub and spoke' model, with a central project database as the hub and the connections to the local databases as the spokes. An advantage here is being able to rely on the domain expertise of the local partners so that only the crucial task remains to create a suitable schema that capture data relevant to the overall work flow of the projects in the central data hub. Examples are integrated structural genomics projects, such as the Protein structure factory PSF [12, 13] or the Southeast Collaboratory for Structural Genomics SCSG [1].

Distributed information system

A distributed and integrated system (Fig. 1d) consists of locally maintained databases and a layer of software mediators that facilitate the connection between the individual databases. The mediators are computer software services that communicate with the participating databases, access specific information and provide a common view of the data to the user of the mediators. This system resembles to some extend a federated system but without a central data hub. The role of the data hub is performed by mediator services. These services are web-based software programs that expose defined parts of data to external requests from users or other services, mediating the interaction between user and data. The use of mediators allows a flexible approach, since multiple mediators can be generated that are tailored towards specific needs and show only parts of the available data. Because of its flexibility it is similar to the ad hoc connection system, however, it provides a structured approach to data exchange and integration. The mediators provide a 'live' view of the data, but do not itself modify data. Full advantage is taken of the local data management, so that the individual researcher maintains maximum control over data generated and assures the accuracy, integrity and maintenance of data. The distributed system can consist of a multitude of services. A directory service forms a central part of the system and

allows an overview of what services exist and what data they provide. The directory service, analogous to a phone book, stores information about mediator services after they have been registered with the directory. The directory service is the starting point for searches and interaction with the mediators services.

The prime example of a successful, distributed information system is the Internet itself. Web sites and web pages are the data sources participating in the network. The knowledge about what sources are available and where they are available are provided either by manually curated directories (for example, Open Directory Project, http://dmoz.org, LookSmart: http://www.looksmart.com) or by automated web indexing services (such as Google, http://www.google.com). The connections are made using a standard communication protocol (HTTP) and are facilitated via the domain name service (DNS), a directory service that tells a computer to which other computer it needs to connect to via its unique internet number, in order to access a specific web page.

General requirements for a collaborative data exchange system

Intensive collaborations, which make use of data generated by individual partners, require the setup of an intelligent system for exchanging data and the information derived from experiments. Our analysis shows that researchers and data managers require a set of characteristics, listed in Table 1, to be met by such a system in order for further consideration.

To combine the requirements of local data storage and data handling with the goal of exchanging and sharing data among collaborators is one of the challenges researchers and data managers face within projects such as NuGO. Based on the described requirements and the strengths and weaknesses of potential data management scenarios, we developed the NuGO concept for a distributed information network.

Information network concept

For a research network such as NuGO with its emphasis on collaboration and highly diverse and data-intense research topics, it seems difficult to establish a common infrastructure for data exchange and data integration with traditional concepts. The size of the collaborative network and therefore the logistics, as well as sustainability issues, make a simplistic ad hoc strategy of sending data files back and forth difficult to manage. Even when the network is separated into more focused research teams with fewer partners, this strategy remains error-prone and difficult as outlined earlier. The loss of data control, the limited flexibility and the high demand of resources make the creation of a common, central database a lesser choice for the network. The same is true for a system of a federated database, even if in this system the data handling favors the individual researcher and his/her intellectual property (IP) rights. In our view, the best concept for a collaborative network is a distributed web-based information system, which mimics the research network itself with its inherent flexibility. It builds on concepts and structures that have

Table 1 User requirements for a data exchange and integration system

Data integrity	The system needs to make sure that is difficult to alter data accidentally or maliciously. If data items are changed, then a record of the changes needs to be made (data provenance, security)
Data provenance	The system needs to make sure that the origin of the data is know and that any subsequent modifications are recorded (data integrity)
Costs	Limited resources prohibit cost intensive solutions, the (re)use and adoption of existing of existing standard procedures is favored (open standards)
Open standards	Open, public standards for data processing (annotations, reporting) and for information technology, such as computing protocols and software tools, need to be employed. This eases data exchange through common mechanisms and structures and prevents dependencies on proprietary solutions and licensing costs (see costs)
Flexibility	The system needs to be able to cope with changing projects and user requirements
Autonomy	The autonomy of the individual researcher needs to be preserved by allowing individual, local solutions for data handling and management. It should give the researcher maximal control over what data and information are made available (IPR)
Intellectual property rights/privacy	Intellectual property rights (IPR) of the individual researcher need to be protected within the scope of the collaborative agreement. The privacy of sensitive data needs to be preserved (security)
Security	The system needs to provide procedures to flexibly control access to data, preventing unauthorized access, and record its use (see IPR)
User interface	Acceptance by scientists is based on the ease of use and, therefore, the design and structure of the user interface

been proven successful in other areas of networking. Such a system preserves all the advantages of locally maintained and structured data sources and therefore preserves the individual researchers autonomy and protection of the researchers IP rights. Its flexibility enables the possibility for new collaborations and data connections that can not be foreseen at present.

Compared to the other approaches, a distributed system has no technical disadvantages, since it makes use of existing and proven computing technologies and protocols. In fact, due to its inherent flexibility, the inclusion of new techniques and procedures in data management and data exchange can be much easier accomplished compared to other, more rigid strategies.

The NuGO concept for a distributed information network is based on three corner stones:

- Partners are willing to share and exchange their data as part of the collaboration with other members, and have a need to cross-link and integrate their results also with publicly available information.
- A structured system of data and information exchange and data integration is more efficient in a collaborative scientific setting compared to simplistic but difficult to manage ad hoc approaches.
- The intellectual property rights, ownership and, therefore, the autonomy of the individual researchers and their data need to be preserved.

Therefore, the structure and implementation of the NuGO Information Network (NIN) that follows from our concept must ensure that (a) data and information sources stay within the research group or institution that are generating these data. Therefore the system builds on sources that are 'distributed' throughout the network. It must assure that (b) local data sources are securely accessed on a 'as-needed' basis through computer programs (web services) and that (c) the data sources and data services are linked through common mechanisms to form a network of data and information.

The basic structure of the NuGO Information Network is, reflected by its name, a network of information services and data sources (Fig. 2). The basic components of the network are the local data sources. In the next layer, web services mediate the access to the data components and make these available within the network. The third layer consists of 'secondary' services for integration and aggregation that combine data from primary data services. On top of all layers, one or more service directories provide overall information on what services are available in the network.

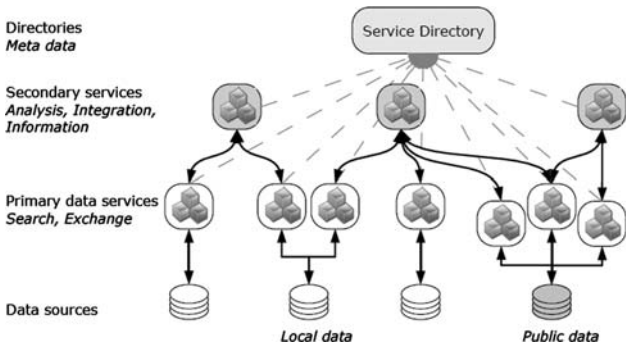

Fig. 2 Concept schema of the NuGO distributed information system: *dashed lines* service registration with the service directory

Data sources

The basic data sources participating in the network are locally maintained databases. The primary targets for the NIN are databases with experimental data from nutrigenomic studies managed by LIMS systems such as nutriBASE installations. Data sources can also be derived from structured data collections for a particular purpose such as gene-phenotype associations from the mouse gene-obesity database [38]. Also less structured data, such as spreadsheets or other data file collections can be used by web services as source of exported data. A major requisite for turning stored data into information and enable intelligent data exchange is the correct annotation of data. This includes the accurate capture of experimental data. Therefore, efforts are underway within NuGO to implement capabilities to annotate experiments with information relevant to nutritional science using a specific controlled terminology, the NuGO-CV. This semantic layer is being combined with an accessible syntax (MAGE-TAB/ISA-TAB formats) to deliver a biologist-oriented tool, ISACreator (http://isatab.sf.net/isacreator.html). While the MAGE-TAB format is restricted to microarray data [25], ISA-TAB [30] allows reporting of multiomics studies.

Data sources within the network need not to be restricted to locally maintained data sources. Publicly available databases and data services, such as GenBank [2] (http://www.ncbi.nlm.nih.gov/Genbank/) for sequence data, ArrayExpress [24, 27] (http://www.ebi.ac.uk/arrayexpress) for gene expression data or KEGG [15, 16] (http://www.genome.jp/kegg) for biochemical pathways, can be made part of the network. In fact, these public sources are essential for the integration of data since they provide the resources to cross-link data from different sources and to existing knowledge.

Communication protocol

The NIN concept uses the infrastructure from the most successful distributed information system to-date, the world-wide infrastructure known as the Internet. The procedures and protocols employed for the communication between parts of the network are therefore proven standards for a distributed system.

Local computing networks and therefore the data and information sources they contain, are usually protected by so-called 'firewalls'. In computer networking, 'firewalls' are rules that govern the communication between computers from the outside and the local network. These rules usually prevent most of the direct connections to computer resources and services, such as databases, of a local network. One widely used exception is, because of the overall importance and popularity of the world-wide-web, the unimpeded access to web servers. Web servers present local web sites to the outside world. The NIN concept uses this readily available communication channel by employing web services as the basis for the communication between local data sources and the network. Web services are a standard internet technology [33] (Fig. 3) for computer-to-computer communication. Web services can make any data available in a form suitable for transport via HTTP, the protocol web servers use for communication. However, web services are not only a communication vehicle but also a tool to control data access. The owner of a web service can define what types and subsets of data are accessed by the web service and are therefore made available. No other data is therefore accessible to others, protecting the researcher's property rights and the privacy of data.

Connections

Web services have a role analogous to regular web sites for communication with the important exception that web services are designed for the communication between computer programs, whereas web pages are designed for the communication between humans. Similarly to a directory for phone numbers or web pages, the NIN's central directory provides the information about the available web services (Fig. 2). Its repository stores meta data of the services, where they are located, what data or information they provide and under what condition they are accessible. This information is provided by the creator of the service upon registration of the service at the directory. The directory can then be queried by other computers or other services. Other web services within the system can subsequently use the information about the services to combine data into an aggregated view or a new integration schema.

Integration

Providing structured access to data for a group of collaborating researchers is the first major goal of the NIN. The network concept, however, allows also the extension of the network through 'secondary' services (Fig. 2). Secondary means that these services make use of the primary data services for further processing. This can include either analysis of data such as normalization of gene array

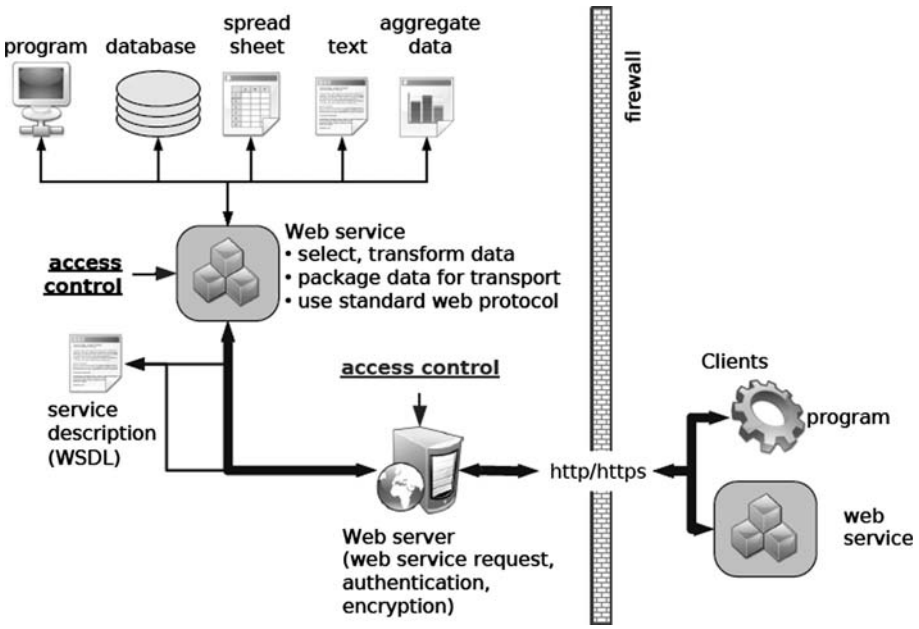

Fig. 3 Web service schema: Web services export defined data from local data sources and package these data items for transport through firewalls using standard internet mechanisms. Web services are accessed by other computer programs or web services that handle display or further processing

expression data or the integration of the data horizontally, by forming aggregates for similar data items from different sources, or vertically, by connecting different types of information into a chain of information, for example connecting information about gene sequence and SNPs, gene-expression and metabolomic data.

Data integration, especially computer supported integration, requires the suitable annotation of data items. A common vocabulary for describing the meaning (semantics) of experimental conditions and resulting data items is therefore a prerequisite for data integration in the network. While the NIN supports the development and use of standardized procedures, it cannot provide standardization of data types and annotations on its own. These standardization efforts need to be undertaken concurrently with the network development to make data exchange feasible and useful. Standardization effort within NUGO as well as other organizations. Sansone et al. [29, 30] show that the questions on what level data integration and combination needs to happen to create useful information are fundamental questions of science and a challenge for collaborating scientists. The NIN as an information management tool provides the procedures and tools that enable and encourage the efficient use of common vocabularies and annotations.

Implementation of the NuGO Information Network (NIN)

The concept of a flexible, collaborative data exchange network is an innovative approach for the nutritional sciences and nutrigenomics. However, similar concepts and solutions have already been developed in other areas of science, especially in computer sciences. For the implementation of our concept, NuGO is therefore reusing existing solutions and adapts them to its specific needs.

The implementation of the NIN concept is based on the BioMOBY system (http://www.biomoby.org), an open-source system for interoperability between biological data hosts and analytical services [21, 36, 37]. In BioMOBY, web services based data services and analytical services are registered with a central repository. A query for a particular type of service to the repository returns the machine readable information about the service and how to connect to web service.

To protect sensitive and unpublished data, the control over the services registered and the access to the information contained is necessary. The NuGO version of BioMOBY, NuGOMOBY, is physically separated from the public BioMOBY system. This separation has the additional advantage to avoid possible collisions of name spaces with services registered in BioMOBY. NuGOMOBY uses its own repository server (NuGOCentral), whose database contains the central service directory where participating services are registered.

The service information provided by NuGOCentral is made accessible to NuGO members. The authentication for this access is done using a NuGO wide central authentication service based on a standard lightweight directory access protocol (LDAP) server. The NuGOLDAP server, implemented using OpenLDAP (http://www.openldap.org) software, stores the NuGO member and NuGO team database. This information enables the fine-tuning of access rights based on criteria such partner institution, role within NuGO or a specific research team membership. The LDAP based authentication procedure can also be used by the providers of the data web services to control access to their own services.

Additional protection of the data access and data transfer can be achieved by encrypting the communication and using digitally signed certificates for authentication. The encryption of the data is done by the web server, using the standard protocols (Transport Layer Security TLS/SSL) [11, 34]. This procedure combines easily with the authentication using certificates. Digital certificates are the computer equivalent of ID cards and are issued to NuGO partners through the NuGO certification authority, which is part of the NIN (http://nugo.dife.de/nugoca). Computers of the NIN that use encryption also use certificates to provide identity information.

Although systems like BioMOBY or NuGOMOBY are designed mainly for the communication between computers, user interfaces have been established for human interaction with the repository. A web-based front end, based on the Genome Browser from the Generic Model Organism Database project, GMOD (http://gmod.org) [31], is available for searching the service data repository of the NIN, NuGOCentral. This tool can also be used to initiate the communication with available services.

To assist with the creation of web service that participate in the NuGOMOBY system, a web-based, interactive service generator, adapted from the BioMOBY code generator, has been made available through the NuGOMOBY web site. For the batch creation of services and maintenance tasks for the NuGOCentral repository several Perl programs have been developed.

NuGO Black Box (NBX) system

Communication and data exchange within a distributed bioinformatics infrastructure benefits in particular from a well-defined computing environment. A common computing platform, with a common software configuration and a standard set of bioinformatics tools pre-installed streamlines collaboration between partners. NuGO is

therefore supporting a standardized computing environment within the NIN through the NuGO Black Box (NBX) system. This system is adapted from 'Bio-Linux', developed by NEBC (UK, Natural Environment Research Council, Environmental Bioinformatics Centre, http://nebc.nox.ac.uk). The NBX system provides the physical backbone of the NIN. The system consists of a network of preconfigured, standardized bioinformatics servers distributed to the partners of the NuGO network and connected to the Internet via WAN (Wide Area Network) connections at each partner. An NBX contains server-grade hardware and is configured for high availability. The NBX servers can survive a single disk failure without any loss of data by using a RAID (Redundant Array of Independent Disks) system, and are backed up automatically every 24 h to avoid loss of data by accidental file deletion or file system corruption.

The NBX operating system is based on the Ubuntu distribution (http://www.ubuntu.com) of Linux, the free, open source UNIX-like operating system [35]. The NBX servers provide a standardized set of software for data management, analysis, and exchange. This includes the Base2 microarray LIMS for local data management and the NuGO-adapted BioMOBY framework for web-service-based data exchange. The preinstalled bioinformatics software includes Genepattern from the Broad Institute [26], and the pathway analysis tools Eu.Gene [7] and Pathvisio [22], developed by NuGO members. NuGO specific software is distributed to the NBX via a Debian (http://www.debian.org) software package repository. Software updates are downloaded automatically by the NBX servers. User access to the NBX system is possible either using the web interfaces provided by the bioinformatics applications or via direct login to an NBX server. The NBX is configured as a terminal server, providing a remote X11 graphical desktop environment tailored to the requirements of NuGO scientists. Web access and direct logins are restricted and controlled using the NuGO LDAP directory service. The NBX system enhances the benefits of the NIN for the individual partners by providing state-of-the-art computing facilities that might otherwise only slowly find their way into nutritional science laboratories.

Discussion

The Human Genome Project has not only spawned the application of 'omics' technologies in many areas of biological science. Its success has stimulated the application of advances in computing sciences towards the life sciences. The way data are handled and integrated in collaborative projects is now receiving considerable interest as is the establishing of new computing infrastructures, such as GRID systems [6, 10], captured by the term cyberinfrastructure [5, 14]. NuGO has recognized the importance and potential of these areas for the advancement of nutritional sciences and nutrigenomics. It has establishing work packages, which participates in the development with the distributed information network described in this paper.

A crucial aspect of research integration is the integration of data, either horizontally, by pooling and aggregation of similar data, or vertically, by the combination of different data types into a biological system model or biological pathways. The task of data integration itself depends on access to relevant data, a fact that emphasizes that the structure of data exchange is another crucial activity for a successful research cooperation. Therefore, a successful cooperation in a network and the development of integrative approaches also depends on a fruitful interaction of scientists with data managers and bioinformatics experts.

Finally, apart from the logical aspects of data integration mentioned above, the psychology aspect is crucial: modern biology becomes multidisciplinary and data flow grows exponentially, but the ownership of data by individual researchers or groups needs to be respected, at least until all results are published.

The NIN concept and its subsequent implementation with the setup of the NuGOMOBY system provides the partners of a network with an easy, flexible, yet structured way of collaboration and data exchange. The experience gained during this process should be of general value for other collaborative projects in nutritional sciences and other life science areas. Although often underestimated, data management and information management procedures, whether for local purposes or cooperations, are already an important part of the way science is conducted. Their importance will increase with the mounting complexity of scientific questions and increased sophistication of scientific experiments. Scientists will need to stay informed about issues and technical possibilities in data and information management [17], just as a grasp of modern analytical methods are necessary for successful science.

The NuGO Information Network introduces nutritional scientists to this area and provides a tool for information management. Many problems still need to be addressed in cooperation between scientists and data managers. Standardization, annotations and the use of common vocabularies in data integration is one of these important areas. Using appropriate and standard terms when describing experimental conditions is key to successful, computer supported data integration. Otherwise manual intervention with its high demand on resources is necessary to make date items comparable. The effort by NuGO for standardization of describing nutrigenomics experiments by developing a specific controlled vocabulary is a large contribution towards a solution for the data integration challenge.

The improvement of computing infrastructure that is directly relevant to bench scientists is another critical area where improvements are necessary. Despite tremendous advances in computing technologies and computing power, many, especially smaller nutritional sciences groups, lack the necessary resources, expertise and computational support to profit from a productive data management environment. Standardization efforts in these areas by NuGO advocate the use of standard description of nutrigenomics experiments and a common LIMS system (the 'nutriBASE' system) deployed through a set of standardized, networked computing workstations (NBX system) that serves as the physical backbone of the NIN. These efforts can serve as blueprints for further collaborative projects.

The flexibility of the NIN concept and infrastructure allows the creation and interconnection of a wide variety of information services that access and analyze specific data sources. We continue to work with NuGO partners to define and implement such services and their integration into the NuGO Information Network. Last, as the NIN/NBX infrastructure serves nutrigenomics collaborations, we invite research teams with likewise scopes to join this effort.

Acknowledgments This work was funded by the European Commission's Research Directorate General under the "Food Quality and Safety Priority" of the Sixth Framework Programme for Research and Technological Development, Grant Number FOOD-CT-2004-506360. We thank Prof. Peter Gray, Aberdeen University and Chris Burnett, Aberdeen University for their valuable contribution to the concept development.

Conflict of interest statement The authors declare that there are no conflicts of interests involved in this work and that they have no financial relationship to the funding organization.

References

1. Adams MWW, Dailey HA, DeLucas LJ, Luo M, Prestegard JH, Rose JP, Wang B-C (2003) The Southeast Collaboratory for Structural Genomics: a high-throughput gene to structure factory. Acc Chem Res 36:191–198
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. Nucleic Acids Res 36:D25–D30
3. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) The protein data bank. Nucleic Acids Res 28:235–242
4. Birney E, Bateman A, Clamp ME, Hubbard TJ (2001) Mining the draft human genome. Nature 409:827–828
5. Buetow KH (2005) Cyberinfrastructure: empowering a "third way" in biomedical research. Science 308:821–824
6. Burrage K, Hood L, Ragan MA (2006) Advanced computing for systems biology. Brief Bioinform 7:390–398
7. Cavalieri D, Castagnini C, Toti S, Maciag K, Kelder T, Gambineri L, Angioli S, Dolara P (2007) Eu.Gene analyzer a tool for integrating gene expression data with pathway databases. Bioinformatics 23:2631–2632
8. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science 300:286–290
9. Corthésy-Theulaz I, den Dunnen JT, Ferré P, Geurts JMW, Müller M, van Belzen N, van Ommen B (2005) Nutrigenomics: the impact of biomics technology on nutrition research. Ann Nutr Metab 49:355–365
10. Coveney PV (2005) Scientific grid computing. Philos Transact A Math Phys Eng Sci 363:1707–1713
11. Dierks T, Rescorla E (2008) The transport layer security (TLS) Protocol Version 1.2, RFC 5246 (Proposed Standard), http://www.ietf.org/rfc/rfc5246.txt
12. Heinemann U, Büssow K, Mueller U, Umbach P (2003) Facilities and methods for the high-throughput crystal structural analysis of human proteins. Acc Chem Res 36:157–163
13. Heinemann U, Frevert J, Hofmann K, Illing G, Maurer C, Oschkinat H, Saenger W (2000) An integrated approach to structural genomics. Prog Biophys Mol Biol 73:347–362
14. Hey T, Trefethen AE (2005) Cyberinfrastructure for e-science. Science 308:817–821
15. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30
16. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30:42–46
17. Lemay DG, Zivkovic AM, German JB (2007) Building the bridges to bioinformatics in nutrition research. Am J Clin Nutr 86:1261–1269
18. Mariman ECM (2006) Nutrigenomics and nutrigenetics: the 'omics' revolution in nutritional science. Biotechnol Appl Biochem 44:119–128
19. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, Sekhon M, Wylie K, Mardis ER, Wilson RK, Fulton R, Kucaba TA, Wagner-McPherson C, Barbazuk WB, Gregory SG, Humphray SJ, French L, Evans RS, Bethel G, Whittaker A, Holden JL, McCann OT, Dunham A, Soderlund C, Scott CE, Bentley DR, Schuler G, Chen HC, Jang W, Green ED, Idol JR, Maduro VV, Montgomery KT, Lee E, Miller A, Emerling S, Kucherlapati, Gibbs R, Scherer S, Gorrell JH, Sodergren E, Clerc-Blankenburg K, Tabor P, Naylor S, Garcia D, de Jong PJ, Catanese JJ, Nowak N, Osoegawa K, Qin S, Rowen L, Madan A, Dors M, Hood L, Trask B, Friedman C, Massa H, Cheung VG, Kirsch IR, Reid T, Yonescu R, Weissenbach J, Bruls T, Heilig R, Branscomb E, Olsen A, Doggett N, Cheng JF, Hawkins T, Myers RM, Shang J, Ramirez L, Schmutz J, Velasquez O, Dixon K, Stone NE, Cox DR, Haussler D, Kent WJ, Furey T, Rogic S, Kennedy S, Jones S, Rosenthal A, Wen G, Schilhabel M, Gloeckner G, Nyakatura G, Siebert R, Schlegelberger B, Korenberg J, Chen XN, Fujiyama A, Hattori M, Toyoda A, Yada T, Park HS, Sakaki Y, Shimizu N, Asakawa S, Kawasaki K, Sasaki T, Shintani A, Shimizu A, Shibuya K, Kudoh J, Minoshima S, Ramser J, Seranski P, Hoff C, Poustka A, Reinhardt R, Lehrach H, Consortium IHGM (2001) A physical map of the human genome. Nature 409:934–941
20. Müller M, Kersten S (2003) Nutrigenomics: goals and strategies. Nat Rev Genet 4:315–322
21. Navas-Delgado I, del Mar Rojano-Muñoz M, Ramírez S, Pérez AJ, León EA, Aldana-Montes JF, Trelles O (2006) Intelligent client for integrating bioinformatics services. Bioinformatics 22:106–111
22. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C (2008) Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 9:399
23. van Ommen B, Stierum R (2002) Nutrigenomics: exploiting systems biology in the nutrition and health arena. Curr Opin Biotechnol 13:517–521
24. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I,

Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone S-A, Sklyar N, Zhao M, Sarkans U, Brazma A (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37:D868–D872

25. Rayner T, Rocca-Serra P, Spellman P, Causton H, Farne A, Holloway E, Irizarry R, Liu J, Maier D, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert C, White J, Whetzel P, Wymore F, Parkinson H, Sarkans U, Ball C, Brazma A (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7:489
26. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0. Nat Genet 38:500–501
27. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Sansone S-A (2003) ArrayExpress: a public database of gene expression data at EBI. C R Biol 326:1075–1078
28. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C (2002) BioArray software environment: a platform for comprehensive management and analysis of microarray data. Genome Biol 3: software0003.1-0003.6
29. Sansone S-A, Rocca-Serra P, Tong W, Fostel J, Morrison N, Jones AR, RSBI Members (2006) A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. OMICS 10:164–171
30. Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S, and the RSBI Working Group M (2008) The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" OMICS 12:143–149
31. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610
32. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351
33. Wikipedia (2008a) Web service—Wikipedia, The Free Encyclopedia. Online; accessed 3 November 2008. http://en.wikipedia.org/w/index.php?title=Web_service&oldid=249016692
34. Wikipedia (2008b) Transport Layer Security—Wikipedia, The Free Encyclopedia. Online; accessed 30 October 2008. http://en.wikipedia.org/w/index.php?title=Transport_Layer_Security&oldid=248218492
35. Wikipedia (2008c). Linux—Wikipedia, The Free Encyclopedia. Online; accessed 4 November 2008. http://en.wikipedia.org/w/index.php?title=Linux&oldid=248966060
36. Wilkinson M, Schoof H, Ernst R, Haase D (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. the PlaNet exemplar case. Plant Physiol 138:5–17
37. Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform 3:331–341
38. Wuschke S, Dahm S, Schmidt C, Joost H-G, Al-Hasani H (2007) A meta-analysis of quantitative trait loci associated with body weight and adiposity in mice. Int J Obes (Lond) 31:829–841