State-of-the-Art and Future Challenges in the Integration of Biobank Catalogues

Heimo Müller[1(✉)], Robert Reihs[1], Kurt Zatloukal[1], Fleur Jeanquartier[2], Roxana Merino-Martinez[3], David van Enckevort[4], Morris A. Swertz[4], and Andreas Holzinger[2]

[1] Institute of Pathology, BBMRI.at, Medical University Graz, Neue Stiftingtalstraße 2/B61, 8036 Graz, Austria
heimo.mueller@medunigraz.at
[2] Institute for Medical Informatics, Statistics and Documentation Research Unit HCI-KDD, Medical University Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria
[3] Department of Medical Epidemiology and Biostatistics (MEB), Karolinska Institutet, PO Box 281, 171 77 Stockholm, Sweden
[4] Genomics Coordination Center, University Medical Center Groningen, PO Box 30.001, 9700 RB Groningen, The Netherlands

Abstract. Biobanks are essential for the realization of P4-medicine, hence indispensable for smart health. One of the grand challenges in biobank research is to close the research cycle in such a way that all the data generated by one research study can be consistently associated to the original samples, therefore data and knowledge can be reused in other studies. A catalogue must provide the information hub connecting all relevant information sources. The key knowledge embedded in a biobank catalogue is the availability and quality of proper samples to perform a research project. Depending on the study type, the samples can reflect a healthy reference population, a cross sectional representation of a certain group of people (healthy or with various diseases) or a certain disease type or stage. To overview and compare collections from different catalogues, we introduce visual analytics techniques, especially glyph based visualization techniques, which were successfully applied for knowledge discovery of single biobank catalogues. In this paper, we describe the state-of-the art in the integration of biobank catalogues addressing the challenge of combining heterogeneous data sources in a unified and meaningful way, consequently enabling the discovery and visualization of data from different sources. Finally we present open questions both in data integration and visualization of unified catalogues and propose future research in data integration with a linked data approach and the fusion of multi level glyph and network visualization.

Keywords: Biobank catalogue · Linked data · Minimum information about biobank data sharing (MIABIS) · Knowledge discovery · Visualization · Glyph

1 Introduction

Biobanking is a relatively new concept that has been evolving over the years to become an essential part of biomedical research. Thousands of biobanks worldwide have been

A. Holzinger et al. (Eds.): Smart Health, LNCS 8700, pp. 261–273, 2015.
DOI: 10.1007/978-3-319-16226-3_11

collecting *bio-specimens*, clinical and research data from millions of individuals in different stages of their lives, before, during and after disease. All this information is a great source of knowledge for fundamental biomedical research and has the potential to dramatically contribute to the development of better predictive, preventive, personalized and participatory (P4) healthcare.

The biobanking landscape is evolving from insulated local biospecimen repositories to robust organizations providing services that cover a large part of the biomedical research cycle, from the biobanking processes up to large scale molecular profiling. High-throughput technologies are more accessible to research-biobanking and the number of biobanks providing services that require large storage capability and parallel data analysis is increasing.

One of the major challenges in biobank research is to close the research cycle in such a way that all the data generated by one research study can be consistently associated to the original samples and hence data and knowledge can be reused in other studies.

Another challenge is to achieve a real informatics integration of biobanks. Even when the technical conditions are created to establish networks of biobanks where bio-resources can be visible to clinicians and researchers regardless of the geographical location, the harmonization process is still in a very early state, not only due to the heterogeneous representation of biobank data but also and most importantly, to the lack of standards for representing and implementing governing policies for ethics and regulation involving sharing of biobank human samples and data (Fig. 1).

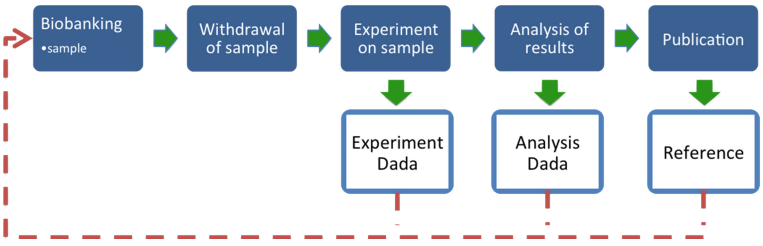

Fig. 1. Example of biomedical research cycle involving biobanking. Data generated in several steps of the research process should be associated to the original samples in the biobank for further reuse.

A high level of collaboration is necessary in the biobank community to gather sufficient resources to be pooled in order to reach statistical significance and hence derive consistent associations and meaningful knowledge. At the same time, data disclosure should be carried out in compliance with legal and regulatory issues at different levels: national, institutional, biobank, study participant, etc. Harmonization, standardization and regulations need to be in place in order to stimulate the development of infrastructures for biobank interoperability in such a way that all these resources can be visible to the biomedical research community.

Several initiatives are on-going in that direction. For instance, MIABIS: Minimum Information About Biobank data Sharing [1] defines guidelines where several components

represent different actors in the biomedical research process involving biobanking. Each component has a minimum list of attributes required to provide valuable information.

At the European level, an increasing number of countries and projects are implementing biobank catalogues aiming to make their bioresources visible and hence stimulate biobank data sharing. MIABIS is being implemented by several of these initiatives as part of their data models.

The *key question* when searching a biobank catalogue is where one can get the proper samples to perform a research project. Depending on the study type, the samples should reflect a healthy control population, a cross sectional representation of a certain group of people (healthy or with various diseases) or a certain disease type or stage. To support this first level of a query, harmonized disease and phenotype ontologies are needed. The typical next step of a query is which samples are available. Different study types require different types of samples (e.g., blood, serum plasma, tissue, urine, isolated biomolecules, such DNA, RNA, or proteins), and depending on the planned analytical challenge since currently there is no unique description of sample quality available. However the level of compliance with ISO and CEN standards as well as results from spot check quality testing will provide a common description of key quality-relevant parameters. Most research projects not only require access to samples but also access to detailed information on the sample and donor. To provide this information in an internationally standardized manner requires an enormous international collaborative effort addressing many as yet unsolved issues of health care informatics. The next level of information required to initiate a biobank-based research project refers to ethical and legal conformity and terms of access. All this above mentioned information should be provided in an aggregated manner to avoid privacy issues. However, the level of detail should be appropriate in order to allow the definition of a research project and to obtain approval by a research ethics committee or to pass scientific review. Finally, after successful approval of a research project by the respective bodies and after signature of a material transfer agreement, coded data related to individual samples and donors should be made available to users. All these different steps should be efficiently supported by an integrated biobank catalogue, thereby minimizing the time period from the first query to a biobank catalogue to the actual release of samples and data to start the research project. This time period is the most important performance indicator for biobanks.

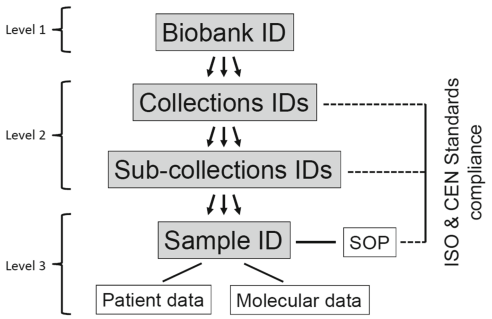

Fig. 2. Different levels of a biobanking catalogue. These levels correspond to the steps of user access needs.

To overview and compare collections from different catalogues and to search for new hypotheses, we must find unexpected patterns and interpret evidence in ways that frame new questions and suggest further explorations. Multilevel glyphs and visual analytics methods will help us to (1) overview collections within a catalogue as the human visual sense is optimized for parallel processing, (2) connect the global view with detail information, (3) provide different contextual views depending on users' needs and experience levels and (4) deal with heterogeneous data sets and different levels of data quality.

Current research highlights the need for interactive data visualization of biobank catalogues, while first approaches of visualization are already emerging [2, 3]. To further address this need it is essential to combine results of knowledge discovery in biobank catalogues and make use of visualization to benefit from the high visual data analysis capacities of humans in order to achieve new fundamental findings for predictive analytics in the medical domain.

2 Glossary and Key Terms

BBMRI-ERIC: is a pan-European distributed research infrastructure of biobanks and biomolecular resources. BBMRI-ERIC facilitates the access to biological resources as well as biomedical facilities and support high-quality biomolecular and medical research.

Biobank: is a collection of biological samples (e.g. tissues, blood, body fluids, cells, DNA etc.) in combination with their associated data. Here this term is mostly used for collections of samples of human origin.

BioSampleDB: The BioSamples database of the EMBL-EBI aggregates sample information for reference samples (e.g. Coriell Cell lines) and samples for which data exist in one of the EBI's assay databases such as ArrayExpress, the European Nucleotide Archive or PRoteomics Identificates DatabasE.

Glyph: In the context of data visualization, a glyph is the visual representation of a piece of data where the attributes of a graphical entity are dictated by one or more attributes of a data record [4].

Linked Data: describes a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried.

MIABIS: Minimum Information About Biobank data Sharing is an attempt to harmonize biobank and research data for sharing. MIABIS defines guidelines where several components represent different actors in the biomedical research process involving biobanking.

P4 Medicine: Preventive, Participatory, Pre-emptive, Personalized, Predictive, *Pervasive (= available to anybody, anytime, anywhere).*

TNM staging: The TNM Classification of Malignant Tumours (TNM) is a cancer staging notation system that gives codes to describe the stage of a person's cancer, when this originates with a solid tumor: T describes the size of the original (primary) tumor and whether it has invaded nearby tissue: N describes nearby (regional) lymph nodes that are involved: M describes distant metastasis.

Sample Collection: A collection of biological specimens (tissue, blood, blood components, cell lines, biopsies, etc.) having at least one common characteristic.

3 State-of-the-Art

3.1 Tools and Data Structure for Catalogue Harmonizations

3.1.1 State of the Art Publications

Data integration is about combining heterogeneous data sources in a unified and meaningful way, enabling the discovery and monitoring of data from different sources. Data integration is synonymous with sharing. When it comes to biomedical data, complexity, diversity and sensitivity are major factors driving the modelling of the integration process. An additional factor is the need to comply with the ethics and regulations for sharing clinical data or research data involving human samples.

The DataSHaPER (Data Schema and Harmonization Platform for Epidemiological Research) is both a scientific approach and a suite of practical tools. Its primary aims are to facilitate the prospective harmonization of emerging biobanks, provide a template for retrospective synthesis and support the development of questionnaires and information-collection devices, even when pooling of data with other biobanks is not foreseen. [5].

The integration of biomedical data is preceded by harmonization and standardization processes. The *BioSHaRE* project [6] demonstrated how retrospective harmonization could make it possible to perform complex statistical analysis on distributed data without compromising personal data protection when using DataSHIELD method [7]. Another interesting sharing tool is eagle-i [8] that allows bio-resources discovery among research institutions. Eagle-i uses the ontology approach to model research resources as instruments, platforms, protocols, bio-specimens, etc. in a distributed environment.

MIABIS is the BBMRI-ERIC's approach to harmonize biobank data for sharing. MIABIS 1.0 standardized high-level biobank data. The main components were "Biobank" and "Sample Collection" [1] (level 1 and level 2 as in Fig. 2). MIABIS paved the way for the creation of the first ontology for biobanking: omiabis [9]. These two steps in the biobank harmonization process have raised interest from the biobank community. Projects such as BiobankCloud (http://www.biobankcloud.com/), BioMedBridges (http://www.biomedbridges.eu/) and RD-Connect (http://rd-connect.eu/) are implementing MIABIS in their data models. Several catalogue initiatives from BBMRI-ERIC member states and BCNet (http://bcnet.iarc.fr/) are also implementing MIABIS. MIABIS 2.0 is currently being designed and a widespread adoption of this standard in Europe is expected.

In the biomedical research domain, integration and interoperability strongly depend on good methods and open source tools that facilitate the adoption of standards and

hence stimulate the sharing culture. Harmonization is frustrating hard work that requires significant human intervention. Biomedical informatics systems are not easily modifiable or adaptable to new standards. We need best practice guidelines for semantics and data formats which at the same time, allow biobanks and researchers to continue using their own idiosyncratic semantics. Biobank management systems should be queried to discover what data they can offer and they should return references to data in the form of URIs (Uniform Resource Identifiers). In that way, biobanks and researchers will use their own semantic annotations rather than imposed specific labels and attributes. At some point the biomedical research domain will need to embrace the Internet of Thing (IoT) concept which is perfectly adaptable when it comes to cataloguing biobank and research bioresources.

Started in the BBMRI Netherlands, the MOLGENIS/catalogue tool was developed as a unified framework to create and federate local and national biobank catalogues. The result is an open source software that is now collaboratively developed between BBMRI, CTMM/TraIT, LifeLines, BioMedBridges and RD-Connect to name a few. The catalogue can host four levels of information: (1) biobank/study descriptions using custom or MIABIS standard format; (2) data schema/data dictionary of data elements; (3) aggregate data/sample availability counts and (4) the individual level data ready for analysis. Increasingly bigger datasets are required for epidemiological and genetic analysis and hence it is important to enable pooling of data from multiple biobanks.

The MOLGENIS/catalogue is building on the open source MOLGENIS platform [10]. This platform was chosen because its data structure can be completely configured using a meta-data definition in the Excel file. It offers pre-build components that allow users to (i) upload data (ii) visualize the data in aggregated or tabular form (iii) securely share the data through a comprehensive security model (iv) integrate data from different domains. In addition there are programmatic interfaces in R for statisticians and in javascript for systems integrators, which also allows data federation of multiple MOLGENIS/catalogues as demonstrated in the BioMedBridges project.

In collaboration with BioSHaRE, Biobank Standardisation and Harmonisation for Research Excellence in the European Union, MOLGENIS also addresses the challenge of data harmonization and integration via the BiobankConnect [11] toolbox aids, designed to assist with this arduous task.

3.1.2 Advantages and Disadvantages of Data Exchange Scenarios

With the growing possibilities of biobank data sharing, (associative) studies can achieve the power necessary to unveil biologically relevant associations for complex traits or diseases. However, sharing phenotypic and genotypic information opens up a complex world of regulatory compliance and privacy concerns. When this information is shared and can be linked back, re-identification of the subject becomes a real concern. Minimal information models help to reduce the risk of re-identification by reducing the available parameters. Study subject selection for most associative studies can be performed on coarser information that aggregates subjects in larger cohorts of similar patients and therefore protects their privacy.

Standardisation of the data items in the models improves the ability to find and reuse biobank data, but simplification of complex phenotypic information in coarse data

vocabularies can lead to a loss of precision. Data vocabularies and ontologies need to be extensive and up to date with the current insights into the biology of diseases. A linked data model provides the ability to maintain the precision of rich ontologies by using linkage instead of tying a data model to a specific ontology.

3.2 Visual Analytics for Biobank Catalogues

3.2.1 State of the Art Publications

When dealing with the integration of biological data for analysis, visualization plays a major role in the process of understanding and sense-making [12, 13]. An overview about the state of the art in the visualization of multivariate data is given by Peng and Laramee [14] as well as Bürger and Hauser, where they discuss how different techniques take effect at specific stages of the visualization pipeline and how they apply to multivariate data sets being composed of scalars, vectors, and tensors. Moreover they provide a categorization of these techniques with the aim of a better overview of related approaches [15], with an update published 2009 [16].

Visual data exploration methods on large data sets were described by several authors, and particularly Keim [17], Hege et al. [18], Fayyad, Wierse and Grinstein [19], Fekete and Plaisant [20], and Santos and Brodlie [21] provide a good introduction to this topic. A recent state-of-the-art report on glyph based visualization and a good overview on theoretic frameworks, e.g. on the semiotic system of Bertin, was given by Borgo et al. [22].

Krzywinski et al., [23] introduce a network structure called hive plots, a graph visualization with nodes as glyphs in the context of systems biology. The layout and format of their glyphs is extensible and editable. Genes that connect cancer subsystems to other systems are represented differently. Another interesting application of glyphs for a visual analytics is an approach for understanding biclustering results from microarray data that has been presented by Santamaria, Theron and Quintales [24] and another one by Gehlenborg and Brazma [25] and Helt et al. [26] and a recent work by Konwar et al. [27]. The closest work to using glyphs with an adaptive layout is the work of Legg et al. [28] in the application domain of sport analysis. Here the data space is event based, and the adaptive layout strategy is focused on overlapping events with so called “macro glyphs”, which combine several glyphs into one. In the “macro glyph” approach only scaling and no level of detail suitable for different screen spaces are applied. Maguire describes a taxonomy based glyph design with an application of biological workflow analysis [29, 30]. Last but not least Müller et al. [2] also show the usage of data glyphs in a visual analytics application and provide an outlook to a biomedical web visualization scenario, the combination of focus and context principle and different level of details are shown in Fig. 3.

A glyph visualizes the mortal state, disease free survival time and the T-Staging of a diagnosis, related to a sample, see Fig. 4.

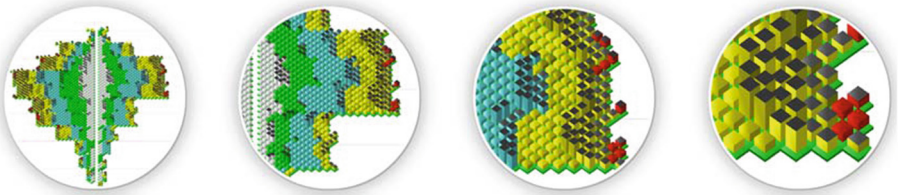

Fig. 3. Visualization of a collection of approx. 10.000 colon cancer samples, shown in 4 zoom levels [2]

Fig. 4. Mapping of sample attributes to a 3D glyph

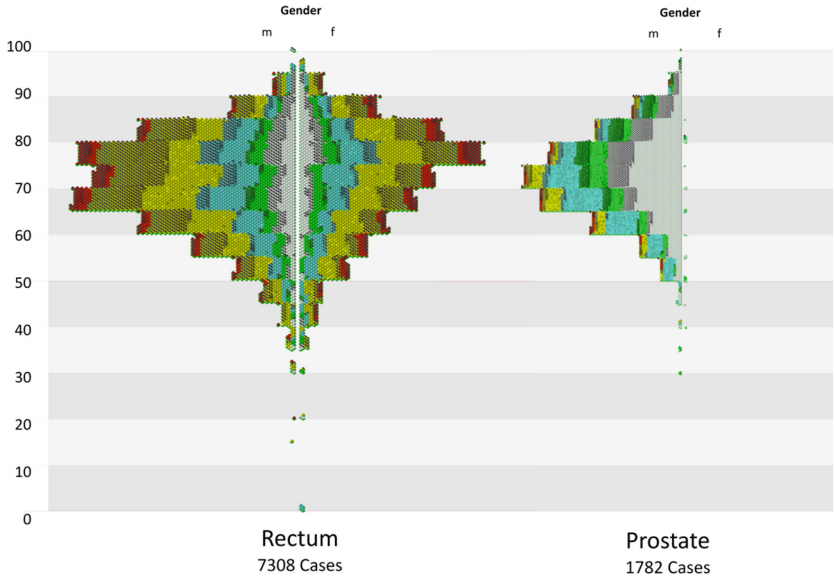

Fig. 5. Comparison of two sample collections of the Biobank Graz

In Fig. 5 the comparison of two sample collections of the Biobank Graz, each covering the same time range of 25 years is shown. The spatial arrangement of the glyphs is done in an age pyramid. All male cancer patients are on the left side and female patients on the right side. The vertical position of a glyph is determined by the patients' age and the horizontal position by the T-staging. We can clearly see the differences in the overall number of rectum and prostate cancer cases as well as the

different distributions of T-Staging. Beside of the overview and comparison of two medium size groups, outliers and data errors can be identified easily, e.g. patients with age of 0 years and female prostate cancer cases.

3.2.2 Advantages and Disadvantages of Glyph Visualization

Glyph visualization is a natural way to map information about a sample to a visual symbol (glyph). However, the data density in this application area is very high, therefore we propose higher dimensional (2,5D and 3D) visualization methods. There are certain advantages and disadvantages when using 2D, 2.5D and 3D glyph visualizations, including different possibilities for placement strategies, linking and brushing, mapping low- to high-dimensional data, projection and interaction, up to benefitting from depth perception while dealing with issues such as occlusion and overlapping glyphs. A comprehensive comparison of 2D, 2.5D and 3D glyphs are still a matter to be researched: Systematically comparing visual complexity levels of 2D, 2.5D and 3D glyph visualization and methods for smooth transitions between different levels of graphical complexity are therefore fundamental research questions yet to be solved [2, 22].

Nonetheless, glyph based visualization is less abstract for effectively conveying information compared to other (visual) representations. By combining glyphs with graphs, certain visualization issues can be solved. For instance, by presenting complex glyphs as nodes in a network, the network itself shrinks and glyph visualization benefits from the graph's spatial arrangement.

Compare also Sect. 4 regarding Challenges and Sect. 5 regarding the fusion of Glyph and Network visualization.

4 Open Problems

Challenge 1: Harmonization of data is a huge challenge in the interchange of biobank data. Minimal data models such as MIABIS are a first attempt to harmonize the field, however, cannot solve the problem of harmonizing the data between different institutes. There is a **difference in the definition of data items.**

Challenge 2: tightly connected with challenge 1 there is also a **difference in the manner in which data is encoded**. Data is often encoded in non-standardized text (often called: "free text") and in the respective national language and there is a plethora of incompatible or only partially compatible ontologies and thesauri, often with merely a national scope.

Challenge 3: **Legal and ethical requirements** in the protection of patient privacy and concerns about losing control of research data lead to hurdles for sharing of data. Even though technical solutions exist to pseudonymize data, manual code lists are often used, which leads to **risk of privacy breaches.**

Challenge 4: At the same time sharing and linking data can **lead to re-identification through combination of data from different sources.**

Challenge 5: When the data elements are (well) structured and connected to ontologies we can analyse and compare collections in a catalogue. For this purpose glyph visualization techniques can be applied, i.e. for visual comparison, hypothesis generation and quality control. Here an appropriate glyph design is important, the development of glyph assessments algorithms and a **comparison of visual complexity levels of 2D, 2½ D and 3D glyph visualization** has to be done.

Challenge 6: Additionally to challenge 5, we have to find methods for **smooth transitions between different levels of graphical complexity.** The main research question here is, on how a high-density design (along with the challenges of the realization of such aspects, e.g. occlusion, depth perception and visual cluttering), indeed influence the user perception and recognition rate in glyph visualization. In particular it is necessary to look at the composition and interferences of visual variables and to carry out a systematic evaluation of shape/placement methods. There are a lot of studies comparing 2D versus 3D visualization techniques in the visualization of spatial related data, e.g. medical renderings or geographic data. However, for abstract information no inherent mapping of the data either to the 3D shape of a glyph nor the spatial position is given, which would be a natural model for visualization. In current solutions the glyph rendering method is changed due to the glyph size in the screen space. Future work should focus on methods for automatic glyph transitions (fusion of semantic and graphical zoom) and evaluate the results in a study.

Challenge 7: After the open problems of dimensionality and transitions of level of details are solved, algorithms for the **optimization of the sample/catalogue attribute mapping to visual variables** and methods for the **spatial arrangement of glyphs derived from network structures** have to be developed.

Challenge 8: In the fusion of glyph and network visualization a central research question is, **how a network topology can be mapped to glyph attributes** (e.g. relations of a sample to several studies) and/or to spatial positioning (e.g. temporal relation of a sample within a disease trajectory).

5 Future Outlook

A major task in the integration and harmonization of biobank catalogues is the provision of a terminology mapping service to overcome the non comparability of data from different sources. This results from the circumstance that institutions usually define their own best fitting data schema for sole use. As a consequence, they often omit to describe the exact meaning of their data, because they don't take into account, that it could be useful for future research performed by others. However, for the correct interpretation of data, especially for third parties, this is essential. Another problem is the fact that the partnering scientists have to consent on a common data schema, which is time-consuming and assumes willingness for compromises.

Within a terminology-mapping service attributes of structured data sets are described in a detailed both formal as well as descriptive manner. This should, in an ideal world,

be done by data creators, who usually know their domain well. Future research will develop methods to support this process as well as motivating the users in doing it. This can e.g. be done with a visual analytics application, which indicates possible matches between attributes from different data schemas already during the data creation process and supports the description of data schema by presenting possible existing metadata sets matching as a starting point.

In the terminology mapping data elements are described by a set of overlapping ontologies, which can are modelled as Linked Data objects and stored together with sample attributes in a (federated) triple store. A toolbox for the curation of a triple store is essential to describe and improve data quality and completeness of a biobank catalogue. For the visualization part of such a toolbox glyphs together with focus and context techniques can be applied. To overcome certain issues with spatial placement of glyphs and benefitting from the fact that common graphs are easy to read, glyph visualization together with network visualization of Linked Data and enrichments on linked data graphs should be combined. In addition, nesting graphs by putting related biobank data into a formal graph structure may enable further exploration. With the emerging standards in biobank data sharing, this approach can be applied to visualise unified biobank catalogues and consequently, unveil and make sense of biologically relevant associations.

Acknowledgements. The work was performed and supported in the context of BBMRI.at the Austrian national node of BBMRI-ERIC. Our thanks are due to all partners for their contributions and various discussions and to Ms Penelope Kungl for proofreading.

References

1. Norlin, L., Fransson, M.N., Eriksson, M., Merino-Martinez, R., Anderberg, M., Kurtovic, S., Litton, J.-E.: A minimum data set for sharing biobank samples, information, and data: MIABIS. Biopreservation Biobanking **10**(4), 343–348 (2012). doi:10.1089/bio.2012.0003
2. Müller, H., Reihs, R., Zatloukal, K., Holzinger, A.: Analysis of biomedical data with multilevel glyphs. BMC Bioinform. **15**(Suppl 6), S5 (2014). doi:10.1186/1471-2105-15-S6-S5
3. Huppertz, B., Holzinger, A.: Biobanks – A source of large biological data sets: open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
4. Ward, M.O.: Multivariate data glyphs: Principles and practice. In: Handbook of Data Visualization, pp. 179–198. Springer, Berlin (2008). doi:10.1007/978-3-540-33037-0_8
5. Fortier, I., Doiron, D., Little, J., et al.: Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int. J. Epidemiol. **40**, 1314–1328 (2011). doi:10.1093/ije/dyr106
6. Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B.H.R., Perola, M., Stolk, R.P., Minelli, F.L., Waldenberger, M., Holle, R., Kvaløy, K., Hillege, H.L., Tassé, A.M., Ferretti, V., Fortier, I.: Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg. Themes Epidemiol. **10**(1), 12 (2013). doi:10.1186/1742-7622-10-12

7. Wolfson, M., Wallace, S.E., Masca, N., Rowe, G., Sheehan, N.A., Ferretti, V., LaFlamme, P., Tobin, M.D., Macleod, J., Little, J., Fortier, I., Knoppers, B.M., Burton, P.R.: DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. Int. J. Epidemiol. **39**(5), 1372–1382 (2010). doi:10.1093/ije/dyq111
8. Vasilevsky, N., Johnson, T., Corday, K., Torniai, C., Brush, M., Segerdell, E., Wilson, M., Shaffer, C., Robinson, D., Haendel, M.: Research resources: curating the new eagle-i discovery system. Database (Oxford). 2012 Mar 20;2012:bar067. doi:10.1093/database/bar067
9. Brochhausen, M., Fransson, M.N., Kanaskar, N.V., Eriksson, M., Merino-Martinez, R., Hall, R.A., Litton, J.-E.: Developing a semantically rich ontology for the biobank-administration domain. J. Biomed. Semant. **4**(1), 23 (2013). doi:10.1186/2041-1480-4-23
10. Swertz, M.A., Dijkstra, M., Adamusiak, T., van der Velde, J.K., Kanterakis, A., Roos, E.T., Lops, J., Thorisson, G.A., Arends, D., Byelas, G., Muilu, J., Brookes, A.J., de Brock, E., Jansen, R.C., Parkinson, H.: The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. BMC Bioinform. **11**(Suppl 1), S12 (2010). doi:10.1186/1471-2105-11-S12-S12
11. Pang, C., Hendriksen, D., Dijkstra, M., van der Velde, K.J., Kuiper, J., Hillege, H., Swertz, M.: BiobankConnect: software to rapidly connect data elements for pooled analysis across biobanks using ontological and lexical indexing. J. Am. Med. Inform. Assoc. 2014 Oct 31. doi:10.1136/amiajnl-2013-002577. [Epub ahead of print] PubMed PMID: 25361575
12. O'Donoghue, S.I., Gavin, A.-C., Gehlenborg, N., Goodsell, D.S., Hériché, J.-K., Nielsen, C.B., Olson, A.J., Procter, J.B., Shattuck, D.W., Walter, T., Wong, B.: Visualizing biological data-now and in the future. Nat. Methods **7**(3 Suppl), S2–S4 (2010). doi:10.1038/nmeth.f.301
13. Turkay, C., Jeanquartier, F., Holzinger, A., Hauser, H.: On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 117–140. Springer, Heidelberg (2014)
14. Peng, Z., Laramee, R.S.: Higher dimensional vector field visualization: A survey. In: Tang, W., Collomosse, J. (eds.) Theory and Practice of Computer Graphics, pp. 149–163. The Eurographics Association (2009). doi:10.2312/LocalChapterEvents/TPCG/TPCG09/149-163
15. Bürger, R., Hauser, H.: Visualization of multi variate scientific data. In: Proceedings of EuroGraphics, pp. 117–134 (2007)
16. Fuchs, R., Hauser, H.: Visualization of multi-variate scientific data. Comput. Graph. Forum **28**(6), 1670–1690 (2009). doi:10.1111/j.1467-8659.2009.01429.x
17. Keim, D.A.: Visual exploration of large data sets. Commun. ACM **44**(8), 38–44 (2001). doi:10.1145/381641.381656
18. Hege, H.-C., Hutanu, A., Kähler, R., Merzky, A., Radke, T., Seidel, E., Ullmer, B.: Progressive retrieval and hierarchical visualization of large remote data. Scalable Comput. Pract. Exp. **6**(3), 60–72 (2001)
19. Fayyad, U., Grinstein, G.G., Wierse, A.: Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann, San Francisco (2002)
20. Fekete, J.-D., Plaisant, C.: Interactive information visualization of a million items. In: IEEE Symposium on Information Visualization, INFOVIS 2002, pp. 117–124. IEEE Computer Society (2002). doi:10.1109/INFVIS.2002.117315

21. Dos Santos, S., Brodlie, K.: Gaining understanding of multivariate and multidimensional data through visualization. Comput. Graph. **28**(3), 311–325 (2004). doi:10.1016/j.cag.2004.03.013
22. Borgo, R., Kehrer, J., Chung, D.H.S., Laramee, R.S., Hauser, H., Ward, M., Chen, M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In: Eurographics 2013-State of the Art Report, pp. 39–63. The Eurographics Association (2012)
23. Krzywinski, M., Birol, I., Jones, S.J.M., Marra, M.A.: Hive plots–rational approach to visualizing networks. Briefings Bioinform. **13**(5), 627–644 (2012). doi:10.1093/bib/bbr069
24. Santamaría, R., Therón, R., Quintales, L.: A visual analytics approach for understanding biclustering results from microarray data. BMC Bioinform. **9**, 247 (2008). doi:10.1186/1471-2105-9-247
25. Gehlenborg, N., Brazma, A.: Visualization of large microarray experiments with space maps. BMC Bioinformatics **10**(Suppl 13), O7 (2009). doi:10.1186/1471-2105-10-S13-O7
26. Helt, G.A., Nicol, J.W., Erwin, E., Blossom, E., Blanchard, S.G., Chervitz, S.A., Harmon, C., Loraine, A.E.: Genoviz Software Development Kit: Java tool kit for building genomics visualization applications. BMC Bioinform. **10**, 266 (2009). doi:10.1186/1471-2105-10-266
27. Konwar, K.M., Hanson, N.W., Pagé, A.P., Hallam, S.J.: MetaPathways: A modular pipeline for constructing pathway/genome databases from environmental sequence information. BMC Bioinform. **14**, 202 (2013). doi:10.1186/1471-2105-14-202
28. Legg, P.A., Chung, D.H.S., Parry, M.L., Jones, M.W., Long, R., Griffiths, I.W., Chen, M.: MatchPad: Interactive glyph-based visualization for real-time sports performance analysis. Comput. Graph. Forum **31**(3pt4), 1255–1264 (2012). doi:10.1111/j.1467-8659.2012.03118.x
29. Maguire, E., Rocca-Serra, P., Sansone, S.A., Davies, J., Chen, M.: Taxonomy-based glyph design – with a case study on visualizing workflows of biological experiments. IEEE Trans. Vis. Comput. Graph. **18**(12), 2603–2612 (2012)
30. Maguire, E., Rocca-Serra, P., Sansone, S.A., Davies, J., Chen, M.: Visual compression of workflow visualizations with automated detection of macro motifs. IEEE Trans. Vis. Comput. Graph. **19**(12), 2576–2585 (2013)