## M·3: Molecuul, Mens en Maatschappij / From Molecule to Society and Back

prof. dr. D.I. Boomsma, Vrije Universiteit, Amsterdam
prof. dr. C.W.A.M. Aarts, Rijksuniversiteit Groningen/Universiteit Twente
prof. dr. F. van Harmelen, Vrije Universiteit, Amsterdam
dr. P.K. Doorn, Data Archiving and Networked Services (DANS) - KNAW
dr. K. Zeelenberg, Organization: Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS)
dr. A. Abdellaoui, Vrije Universiteit, Amsterdam

(a full list of all applicants is attached)

### Summary

The Netherlands is a small country with a well-characterized population where every citizen can be relatively easily reached. Human subject research in the Netherlands is performed across a large number of cohorts, often across different disciplines of social and medical sciences. Our vision for the next decades is to set up an infrastructure for large-scale representative and longitudinal interdisciplinary research from molecule to person to society and back: M·3: *Molecuul, Mens en Maatschappij*; enabling the study of the effects of genomics on society and the feedback from culture and society to the expression of the Dutch genome.

M·3 tackles questions about behavior, lifestyle, and health, on an individual and societal level, based on assessment of population variation across spatial, biological, environmental, social, historical and cultural levels. Characterizing pathways across multiple levels and time-periods is realized because of technological and methodological innovations in different disciplines, including phenotyping at individual, group, and higher aggregation levels by new big data tools; genomics, transcriptomics, and a wide range of other 'omics' fields (metabolomics, proteomics, microbiomics, etc.); breakthroughs in analysis methods and informatics (Bayesian, big data and network approaches; semantic web); data linkage innovation in survey and ambulatory data-collection; high resolution geo-data; and the mining of detailed historical databases.

M·3 will facilitate and integrate resources for data- and sample collection, offer a repository for storage, and offer tools, assessments and analyses from social and life sciences. Thereby the large cohorts from the social sciences can be enriched with genomics, transcriptomics and other omics data and cohorts across all disciplines can be enriched with historical and geospatial information. M·3 offers the Netherlands a research facility that promotes insight into how molecules shape human beings across the entire lifespan, how they shape society and processes of social inequality, and how society influences biological make-up. This infrastructure will attract a broad range of talented researchers around the globe and will promote important and novel research questions not bounded by disciplines. M·3 has the potential to transform the way we study and understand humans on multiple levels.

**Keywords**: society-genome interplay; social inequality; geographical and historical stratification; FAIR data; support facility; lifespan; omics

## A. Science and Technical Case

### 1.1. Scientific value of the M·3 Infrastructure: From Molecule to Society and Back

Facilitating interdisciplinary research is best done through shared missions in the form of scientific questions that can advance society and our understanding of it[1]. Several branches of the social sciences and humanities[*] and of the life sciences[†] aim to understand individual differences in complex human traits and social outcomes and processes. A comprehensive understanding of human behavior on a molecular, biological, individual, and societal level requires narrowing the gap between social and life sciences. We propose an infrastructure that brings together data, methodologies, means, and minds to realize the input for understanding the pathways from molecule to person to society and, equally important, the pathways from society to person to molecule (*Molecuul – Mens – Maatschappij*: M·3).

Society and science are increasingly interested in explaining social constructs with biological measures and explaining how society impacts on biology, e.g., how social context can modify the expression of the genome. M·3 will create an infrastructure which brings together expertise with respect to assessment of biological parameters, the broad exposome, and individual and social outcomes. M·3 will combine and create expertise and facilitate bringing together and harmonizing existing cohort data, collecting new data through innovative approaches, store existing and new samples, enable the application of new techniques to samples and data, and create the knowledge and means to analyze and interpret different types of data across multiple levels, and support the development of novel methodologies and education of interdisciplinary scientists.

Enriching and combining knowledge from the social and life sciences will lead to the prospect of understanding humans and the societies they create on a more fundamental level. We aim for an infrastructure with the capacity to support the major disciplines of social and life sciences in order to pave the way for a unified discipline with the goal of understanding the human species on every level.

The notion that properties of the human mind are encoded in the highly plastic physical organ that is our brain, partly through the highly stable molecular genetic code, is creating interdisciplinary fields where theories about our behavior and social functioning are based on biological mechanisms. The field of behavior genetics for example flourished because of  observations that substantial proportions of variation in behavioral and social traits are associated with variation in genetic relatedness (i.e., are heritable). The field of social neuroscience emerged because of the need to treat the nervous system as part of a social structure instead of as an isolated entity. Meanwhile, many important questions remain to be answered. What are the causal chains between biology and social behavior? What is the impact of social and political context upon the genetic properties of future generations and through what mechanisms? How many and which variables do we need to measure, and when, in order to effectively support citizens in living a healthy and prosperous life from the cradle to the grave? How can we better understand the roots of inequalities in society? How can we better understand children's developmental trajectories on a biological and social level across social strata?

Social inequality is a national and global problem and is considered, as President Obama recently put it[2], as "the defining challenge of our time". Inequalities have been observed across a range of developmental and health related outcomes, educational attainment, and societal outcomes[3,4]. Such inequalities have been ascribed to demographic and environmental factors including social deprivation, differential access to resources, exposure to harmful contaminants as well as to personal and biological factors. These factors can exert detrimental effects from early childhood on, and even prenatally[5]. The role of genomics in inequality is poorly understood and it is only since the breakthroughs in technologies and big data science that we can begin to disentangle the complex role of gene-environment interaction (i.e. the effects of environment conditional on genotype), the role of gene-environment covariation (i.e. the non-independent contributions of

[*] The areas of science concerned with society and the relationships among individuals within a society (e.g., economics, political science, human geography, demography, sociology, anthropology, archaeology, jurisprudence, psychology, history, and linguistics)

[†] The area of science (as biology, medicine, and biotechnology) that deals with living organisms and life processes.

genes and environment), the genetic consequences of assortative mating in the population (e.g., mating between spouses with a similar socio-economic status, which clusters talent and resources across generations), and the feedback loops between society and the expression of the genome. M·3 aims to build the resources and infrastructure to generate a deeper knowledge and understanding that can be incorporated into society to move towards greater equality and social harmony, by e.g. improving children's environments, lifestyles, and the educational system across all strata, thereby improving the future of the Dutch population. The Netherlands is well suited for the continuation and development of next-generation interdisciplinary endeavors. The Netherlands is densely populated, has a well-documented population with rich historical databases, one of the best digital infrastructures worldwide[6], and well-established research infrastructures from a variety of disciplines that can reach every person in the country. Our goal is to combine these elements in order to prepare the country for the rapid scientific and technological developments ahead.

### *1.1.1. Mapping of the Dutch Population*

The Netherlands is a relatively small country with inhabitants that are relatively easy to reach and have been well characterized for multiple generations on a wide variety of social constructs. For example, *Statistics Netherlands*[7] (*Centraal Bureau voor de Statistiek, CBS*) has collected, processed, analyzed and disseminated data and information on persons, households, enterprises, and the environment for more than a 100 years. The information includes detailed geo-coded individual level (and household) data on income, employment, education, health, housing, criminality, care, etc. from multiple sources. Parent-offspring relationships have been well documented in this database since 1947, making the reconstruction of family relationships feasible for the vast majority of the population. The information about persons and households is maintained in the *System of Social and Statistical Datasets* (SSD)[8,9], which is an integrated system of databases. In addition, the SSD contains information on family relationships (grandparents, parents and children, adopted children, siblings, cousins etc.), married and unmarried partnerships, the social environment (regional distribution of the population), and migration patterns of Dutch inhabitants. Since the population registers are available from 1995 onwards, the data are exceptionally well suited for life course research. A major added value of this data system is that all person data can be linked at the individual level through the use of (anonymous) linkage keys. Moreover, under appropriate conditions regarding informed consent, researchers may bring in outside data for linkage to the SSD.

Dutch inhabitants can be linked to their recent ancestors through the *Historical Sample of the Netherlands* (HSN)[10], which comprises a representative sample of about 85,000 individuals born in the Netherlands during the period 1812-1922. These individuals are followed through the archives from the cradle to the grave to construct life histories as completely as possible. The HSN includes family members, and has individual level dynamic information on family structure, occupation, marriage, religion, literacy, social network, and migration history, thus counting about 1 million persons. HSN is connected with the LINKS data, which allows for the reconstruction of all 19thand early 20th century family trees (three and four generation networks). Reconstruction is based on a digitized index of all civil certificates from the 19th and early 20th century consisting of about 25 million certificates with 80 million appearances of persons. Dutch censuses from 1795 to 1971 have been digitally archived at www.volkstellingen.nl. The *Meertens Institute*[11] manages and studies data on culture, traditions, rituals, syntactic variation, and phonological variation across time and geographic locations. Furthermore, possibilities exist to link certain cultural and linguistic properties to subpopulations or geographic areas across time.

Social scientists have collected a large variety of other data on samples from the Dutch population, very often longitudinally. Examples include internationally coordinated studies like:

- *European Social Survey* (ESS, biannually since 2002, organized as European Research Infrastructure Consortium)
- *European Values Study* (each 9 years since 1981, organized as a foundation based in the Netherlands)
- *Netherlands Kinship Panel Study* (each 3-4 years since 2003, part of the Gender and Generations Program consortium)

- *Dutch Parliamentary Election Study* (around each parliamentary election since 1971, part of the Comparative Study of Electoral Systems consortium),
- *Health Behavior in School-Aged Children* (every 4 years since 2002, coordinated by WHO)

Other cohorts include *Social and Cultural Developments in the Netherlands* (SOCON, since 1979) or *Tracking Adolescents' Individual Life Survey* (TRAILS, since 1999). These are just a few examples of the rich Dutch data resources on aspects of individual behavior. Some of these datasets can readily be connected to others (e.g. those of *Statistics Netherlands*) by using linking variables at the individual level. For other data, a connection can be made by the application of statistical imputation methods[12,13]. Some large longitudinal cohort studies have started to include a biobank component, such as the *Netherlands Twin Register* (NTR), the *Leiden Longevity Study*, and the *Groningen Lifelines Study,* and collaborate in the *Biobanking and Biomolecular resources Research Infrastructure NL* (BBMRI-NL) consortium.

These examples are by no means exhaustive and the many cohort studies with their extensive datasets can be substantially enriched with e.g. biological and geo-spatial data on an individual level. The small distances within the Netherlands, the continuous evolution of data collection and survey modes[14], the vast and rapidly increasing amount of information that can be mined from social media, and the ongoing technological advancements in the collection of biological measurements make it feasible to paint a more complete picture of people throughout their life course. We propose to invest in physically approaching groups who tend to be more difficult to reach in epidemiological studies (elderly, rural areas, lower education, males). If required for proper coverage, recruiting participants based on a large random sample from the Dutch population will be considered.

With such high multi-level, genetically informative, and longitudinal geo-coded data, one of the goals of future research in the Netherlands is to define factors that increase health, well-being, and personal and social success in the different stages of human life. Relationships between biological parameters and health and social outcomes have not yet been studied much in the context of co-variations and interactions with cultural, demographic, and geographic factors, which are major determinants of environmental exposures and lifestyle. This would generate knowledge that would make personalized medicine, but also potentially tailored advice in many other areas of life (e.g. education and labour market related outcomes), realistic and extendable beyond the clinic.

### *1.1.2. Molecule to Society*

The measurement of variation on a population scale across multiple spatial, biological, environmental, historical and cultural levels is the foundation of the interdisciplinary M·3 infrastructure. The journey from molecule to society begins at our genome. Craig Venter and Daniel Cohen, two of the world's leading genetic scientists, claim that if the 20th century was the century of physics, the 21st century will likely become the century of biology, mainly through the understanding of life at the most fundamental level of life, the genetic code[15]. Due to the realization that most of our behaviors are considerably heritable, it is now evident that the ~0.1% of the genetic code that differs between any random pair of humans contains crucial information that can lead to insights about underlying biological mechanisms. According to decades of twin studies[16,17] our genetic differences should hold substantial predictive power for most behavioral traits and diseases. Molecular genetic studies have come to similar conclusions with novel methods that are able to partition the heritability into functional genomic categories[18,19]. Humans are complex multidimensional systems, and individual outcomes are influenced by a myriad of factors besides genetic make-up, including social circles, spatial context, lifestyle, culture, history, and many other environmental influences, summarized together in the 'exposome' (the totality of human environmental exposures from conception onwards, complementing the genome)[20]. Of all these factors, our genetic code is the only factor that is nearly devoid of external influences and therefore so stable over the lifespan that the DNA sequence only needs to be measured once, and thus, we argue, should be measured on a population-scale. Genotyping methods and technology have improved radically in the last decade while getting more affordable, which led to large-scale genome-wide DNA mapping in several populations around the world.

The search for genetic variants at the source of individual differences in human behavioral traits is fully ongoing, largely through the experimental design of genome-wide association studies (GWASs) that allows the entire genome to be scanned for causal genetic variants for any heritable trait[21]. GWASs successfully identified genetic risk factors and improved the understanding of biological mechanisms underlying medical and physical traits and diseases, but this design also started to deliver robust findings for more "distal" behavioral and social traits, such as educational attainment, well-being, and entrepreneurship. The effects of single genetic loci on behavioral, social, and psychiatric traits are often small (odds ratio's < 1.1), likely because natural selection weeded out common variants with disproportionally large effects, but perhaps also reflecting our limited knowledge of human genetic variation and its measurement. Together with the large multiple testing burden, this means that exceptionally large sample sizes are needed to achieve robust findings (hundreds of thousands to millions of subjects), which is what the genetics community has been attempting in recent years by combining many cohorts in large consortia. With many cohorts supplying thousands of samples, a few of the largest studies to date have reached sample sizes of hundreds of thousands of subjects, which led to numerous robust findings. With the increasing number of population-scale genotyping efforts that are currently ongoing (e.g., deCODE genetics [N=~150k], 23andme [N=~1M], UK Biobank [N=~500k], Million Veterans Project [N=~1M], Precision Medicine Initiative [N=~1M]), and the Dutch infrastructure that we have in mind, future collaborations will likely reach sample sizes in the millions.

When a GWAS is sufficiently powered it opens up new possibilities, for example to compute polygenic scores at an individual level. In combination with environmental and social factors, polygenic scores are strong tools to better understand complex traits (personality, intelligence, talents, discipline, creativity, educational attainment, economic prosperity, and other traits related to both health and social outcomes). These polygenic scores can be used to answer questions about shared genetic etiologies[22], gene-environment interactions[23], and causal relationships of additional factors[24]. The ability to distinguish associations/correlations from causal relationships is a very welcome and much awaited addition to epidemiological studies[25-27]. In combination with the measures of behavioral, social, and demographic constructs, these methods have the potential to answer questions regarding a diverse range of potentially causal factors for behavioral and social outcomes. In addition, utilizing polygenic scores in combination with high-resolution geo-coded and longitudinal social information to investigate the geographic and social co-variations, interactions, and clustering of the genetic predisposition of a variety of traits will lead to novel and unique study-designs, not yet possible in any existing dataset.

The pace in which the field of genetics is progressing is beyond expectations, and we urge the Dutch social and life science communities to capitalize on this. Fully sequencing the first human genome (which is much more comprehensive and expensive than genome-wide genotyping on SNP micro-arrays) cost ~3 billion dollars about two decades years ago, and is currently ~3 million times cheaper at ~1000 dollars per human genome. Currently the Broad Institute of MIT and Harvard is sequencing on average one full human genome every 25 minutes[28]. Several countries have started to realize the potential of "population-scale" epidemiological studies and are investing in them accordingly. The United Kingdom has invested in a large long-term biobank study called the UK Biobank, where ~500,000 volunteers are being genotyped. The biopharmaceutical company deCODE in Iceland has measured the genomes of ~150,000 Icelanders, which is about half of their population. The Obama administration recently funded the establishment of two large-scale epidemiological projects, both good for a million genotyped participants each: the Million Veterans Program and the Precision Medicine Initiative. As of June 2015, the US based company 23andme has more than 1 million customers genotyped. The power to interrogate the entire human genome has already resulted in thousands of new potential drug targets for physical and mental disorders[29].

In the Netherlands we have the opportunity to participate in a paradigm shift that is taking place in the life sciences largely due to the rapid technological and methodological developments in genetics, and given the multi-disciplinary nature of genetics, to extend these progresses into the social sciences domain.

### *1.1.3. Society to Molecule: Historical & Evolutionary Processes*

Genetic variation with demographic, geographic, social, economic, political, and historical data offers means and possibilities to study the rich Dutch history and the impact of collective human behaviors on the genetic population structure. Combining genetic data with historical documents and current measurements on behavior, language, and culture can help further elucidate mechanisms such as natural and sexual selection underlying the evolution of systematic genetic differences and similarities within and between (sub)populations. For example, in the Netherlands the genetic consequences of the strong influence of religion on mate choice in the last couple of centuries are clearly visible when looking at the large patterns of genetic variation[30], similar to the influence of educational attainment on more recent mate choice[31]. The Netherlands is a small country in which for centuries and even today, many families remained very close to their place of birth, and therefore genetic variation shows clear geographic distributions that correlate with the major religious denominations, and cultural and linguistic differences[30,32]. These patterns of genetic variation also contain signals of natural selection pressures that are informative of human history[32]. Understanding these patterns of genetic variation does not only provide knowledge that is worth pursuing for the interest of historical knowledge itself, but is also crucial to account for in studies searching for functional genetic variants and possibly even in non-genetic studies[33,34]. Not all geographic patterns of genetic variation in the Netherlands are understood in a historical context. The Dutch demographic history itself is highly complex and not yet fully understood. Not only can documented history help us to better understand the currently observed Dutch genomes; the Dutch genomes can also help us to further reconstruct Dutch history[32,35]. Combining genetic variation in the Dutch population with the increasingly growing international datasets on modern humans and earlier hominids can be utilized to further study and understand the history and evolution of modern humans, and trace migratory events in recent and distant human history[36,37].

### *1.1.4. Society and Biology beyond Genetics*

Social and demographic processes influence the genetic make-up of a population and the genetic make-up influences social constructs. The view that causation with respect to social constructs is in a single direction (genetics influence individual outcomes and social constructs) is very much the current paradigm. Our genome, however, works through gene expression which is partly regulated through epigenetics. These are highly dynamic processes. Our current knowledge and technical possibilities to assess these processes on a large scale are limited, but this is bound to change with technological breakthroughs in e.g. epigenetics and expression sequencing and we propose combining and enriching data from social sciences with these other datatypes. The same applies to the other "–omics" fields of which some are starting to mature (e.g., the fields studying the proteome, cytome, editome, pharmacogenome, connectome, metagenome, glycome, ionome, lipidome, metabolome, metallome, proteome, physiome, mechanome, membranome, antibodyome, or microbiomes from different human organs), of which the high-throughput technologies (and thereby the science conducted on these measures) are expected to progress soon and rapidly. Combining these variables opens up the possibility to study the highly complex systems that make up a human as a whole. Adding high spatial resolution data on demographic, social, economic, political, and perhaps even climate and contaminants in the natural environment would allow us to describe someone's past and direct environments and their direct and indirect influences on biology and social outcomes.

There is a large literature on so-called neighborhood or environmental effects on individual outcomes[38]. This literature is far from conclusive with regard to the effects of environmental contexts on outcomes such as health, labor and educational outcomes. This is partly due to the fact that omitted variable- and selection bias complicate the identification of causal effects. Detailed and larger (population level) geo-coded and longitudinal individual level data are crucial in unraveling the relationships between individual behavior and environmental context, not just for one family generation, but also between generations. It is important to measure the spatial context at various levels, from the home, to the neighborhood, the city, and the wider region as person-environmental interactions are likely to take place at various spatial scales. Another reason why literature on environmental context effects is far from conclusive is because there has been limited attention for interactions between spatial variables, personality, and the wide range of people's biological characteristics and those of their direct social circles. We hypothesize that certain environmental characteristics are only important for people with certain personalities and/or a certain biological make-up. We also note that social circles increasingly may become independent of geographical location and spatial constraints. To further this field of research, integration is needed of social, economic, demographic, spatial, and biological data over long periods of time (over generations) and for a very large number of subjects.

M·3 is expected to give an impetus to research in methodology, in particular in methods for cooperation and in methods for statistical analysis of large and complex multi-disciplinary datasets that include a widely heterogeneous variety of measurements. We will need new methods for cooperation in a secure way such that privacy and information security are safeguarded. A promising technique may be Secure Multiparty Computation (SMC), which has been developed in cryptography. The combination of social and biological data and the very large sizes of the datasets give new challenges and opportunities for the development of methods for analysis of these data, for which the Netherlands with its rich tradition in psychometrics, methodology development, statistics, and epidemiology is ideally positioned.

We anticipate collecting many data types in current and future cohorts in the Netherlands within the M·3 infrastructure. Together with the nation-wide collection of biobanks that are already closely collaborating[39], we can combine, enrich, and store biological samples to be utilized in future interdisciplinary endeavors that aim to understand behavioral and social functioning of humans beyond their genetic make-up. In addition, further advances in methods and techniques of survey research are expected, including web-based interviewing (CAWI) and the use of large online panels. Among the future challenges is the further development and refinement of video-enhanced web-based interviewing[40,41], including language choice/detection, and the choice of a level of conversation sophistication. These methods will enhance the quality of survey measures, and make it more feasible to reach participants from all social strata and ethnic backgrounds of the rapidly evolving Dutch society.

### *1.1.5. The Recent Influx of Non-Western Cultures and Genomes*

The Netherlands has a long and rich history of migrations. It has been estimated that ~75% of the "native" Dutch population ("*autochtonen*") have migrated to the Netherlands in the past 2 millennia[42]. Since the Second World War, there have been five major migration waves[43]. About 20% of the Dutch population consists of recent migrants and their descendants, a large proportion that needs to be taken into account in our research. Overall health and mortality rates differ significantly between ethnic groups, which are only partly explained by differences in socio-economic status[44]. A personal or family history of migration is an important risk factor for psychiatric disorders[45-47]. Important determinants of demography, such as migration, geographic clustering, social mobility, and mate choice are also not independent from ethnic background[48,49].

Ethnic background and ancestral history can be accurately detected through one's genetic make-up. Genetic association studies are mostly done in cohorts that exist of subjects of European descent[50], while individuals with non-European descent are usually excluded from the studies to avoid false positives due to population stratification[32,34]. The investigation of more diverse populations comes with its challenges, mainly because of systematic genetic differences that have not been sufficiently accounted for in current study and micro-array designs, but that are expected to contribute substantially to mapping genetic determinants of complex traits for the human population as a whole[50]. In addition, the collection of non-western genomes in the Netherlands is likely to contain additional traces of recent and distant human history that are not detectable in Dutch genomes alone.

The comprehensive multi-level population size dataset that will be accessible through the M·3 infrastructure will provide the opportunity to elucidate mechanisms that lead to systematic differences between ethnic groups for a wide range of traits. This will lead to a deeper understanding of the traits themselves, make the Netherlands a model for studies on the effects of globalization, and may help guide policymakers to steer the country towards a more equal and social cohesive society.

### *1.1.6. Data protection*

Many of the data to be collected and shared between researchers are of a sensitive nature. Thus, data protection and information security are of prime importance throughout the entire programme and merit their own research programs, that need to be concerned with different levels of protection (linkage and research, storage of data and of biological samples; implications of (genetic) results beyond the individual person, and importantly, the involvement of participants in research programs.

### *1.1.7. Science Case: Conclusions*

Between our genetic code and individual and social outcomes, there are many biological and environmental pathways that are not independent of each other and are characterized by complex co-variations, interactions, and feedback loops. To elucidate the relationship between the multiple levels that create human societies is an enormous challenge for the future. There are many types of biological samples that need be collected and stored (across the lifespan and across tissues) to be utilized for a wide range of biological information. Data on socio-economic outcomes, ethnology, climate, ecology, family trees, language, and politics have been and are still collected on a national level. One of the goals of the M·3 infrastructure is to enrich and structure all data in such a way that they can be linked on an individual level as well as to time and geographic location in a high temporal and spatial resolution. The M·3 infrastructure will be equipped with expertise to assist social and life sciences in collecting and utilizing these different types of data, while ensuring data security and protection of privacy of the participants.

We have a unique opportunity given the time and place we are in. The Netherlands is a small country where every person is at most a few hours away from a researcher, thus collecting and combining biological data with social/demographic data is feasible. By joining forces with the high quality research infrastructures in the Netherlands from the social and life sciences we are likely to contribute to the future scientific progresses in human sciences on a molecular, individual, societal, and international level.

## 1.2. Technical Case: *The M·3 infrastructure*

We envision the M·3 infrastructure to broadly consist of two elements. The first element is a centralized facility that serves as a support *Facility and Repository* that enables social scientists to implement and carry out genomics and biomarker research in well-phenotyped cohorts, and that enables researchers from life sciences to include state-of-the art phenotyping for lifestyle, psychological, social and related traits. The second element consists of the infrastructure to facilitate record linkage among datasets, data harmonization and computing and includes new technical and data infrastructures that we label as the *Data Fusion Workbench* and *Data Reuse Vault.*

The motivation behind this central facility is the need for specialized expertise that can aid in bringing two worlds together: 1) the social sciences and humanities, with their ability to collect and analyze data reflecting social constructs, e.g., psychological, sociological, economic, political, anthropological, historical, linguistic, high resolution geographic, and demographic constructs, and 2) the life sciences, with the expertise of measuring and analyzing high-throughput biological measurements, such as genetic variants, gene and protein expression, epigenetic modifications, microbiomes from a variety of human organs, and many other informative biomarkers. We expect high-throughput technologies to have substantially progressed by the time the M·3 infrastructure will be realized, and we expect novel and major possibilities to arise by combining these diverse measurements from molecule to society. We intend this facility to serve as a partner of existing national infrastructures and their facilities from both disciplines, in the sense that they will provide knowledge and practical assistance in transforming existing monodisciplinary datasets into the richer and more informative multidisciplinary datasets that will make the next-generation scientific endeavors possible. Social scientists may not have the resources to invest in the biobanks, equipment, knowledge, and methodology needed to keep up with the rapid progresses in life sciences, and life scientists may likewise not have the resources to stay up to date with the rapid progresses of survey modes, psychometrics, and the collection of social constructs on both an individual and group level.

### *1.2.1. The Support Facility and Repository*

One of the two elements of the M·3 infrastructure consists of a central support facility and repository that ensures support for social sciences in collection of and research with biological samples and for life sciences in extended phenotyping.

M·3 will coordinate laboratory and biobanking facilities, storage facilities, and other necessary equipment, pipelines for data processing and analyses, and manpower, which have to guarantee that state-of the-art knowledge across the life-sciences field is available to scientists with different backgrounds. It is of vital importance for high-throughput biological measurements that biological samples are collected and processed in a uniform way, to avoid technical sources of variation that can cause systematic differences. These systematic differences can lead to spurious associations in downstream analyses and may eventually lead to wrong scientific conclusions[51]. We therefore propose that the M·3 infrastructure includes a centrally coordinated laboratory and repository infrastructure that is responsible for generating biological measurements in a homogeneous way for the national cohorts that are affiliated with M·3. Within this facility, samples from the different participating cohorts must be equally and randomly distributed across potential sources of technological variation (e.g., time of measurement and processing, personnel, storage, reagents, etc.) to minimize the probability of confounding between true biological and spurious batch effects. We will encourage the participating cohorts to conduct their measurements in the laboratories that will be housed at this facility. If decentralized measurements are desirable, these need to be carefully coordinated. Equally important, M·3 will coordinate phenotyping in the social science area to ensure that developmental and life-time information is optimally available, and can be linked and analyzed.

Tasks of the M·3 Support Facility and Repository include:

- Coordinating the contacts with existing facilities from social and life sciences across the Netherlands.
- Providing extensive support for the participating organizations in enriching their data with a wide range of social constructs and/or biological measures.
- Housing the logistic infrastructure required to collect new data (including e.g. cars and vans equipped to collect samples and phenotypes throughout the entire Netherlands).
- Housing a repository that offers secure storage and processing of biological samples, biological data, and phenotypic information. Handling biological samples can be done in collaboration with existing Dutch Biobank Infrastructures, such as BBMRI.
- The digital data that is generated by M·3 does not necessarily have to be stored at this physical location; data storage can be coordinated in collaboration with a large-scale and experienced ICT organizations such as SURF and DANS while maintaining high level data security and protection of privacy. M·3 should be equipped with state of the art high-throughput technologies able to map the wide range of omics necessary for a holistic view of the molecules that make up human beings.
- Coordinating the population-scale genotyping efforts mentioned in the *Science Case* section, with the long-term end-goal of genotyping millions of Dutch citizens. These genotyping efforts can be done in collaboration with large and experienced genotyping companies with nodes in the Netherlands. We propose that M·3 takes the lead in coordinating the large-scale genotyping efforts in such a way that all M·3 partners will at least have the genetic codes measured for all their subjects as one of the baseline variables. The genetic code has to be measured only once since it is not dependent on external variables, and plays the fundamental role in the creation and maintenance of life and in creating a substantial proportion of individual differences. The documentation of distant and recent historical and evolutionary events in the genetic code will open new avenues of research including international collaborations with well-coordinated population-scale genotyping efforts taking place in many other countries world-wide.
- Provide both a technical and semantical interface between the M·3 datasets and other repositories that provide data on the social and environmental context for the envisioned research questions. For example, an interface with PDOK (*Public Service on the Map*) that provided open data on key spatial characteristics

of the Netherlands, or with the historical, cultural and linguistic properties provided by the *Meertens Institute* on subpopulations and/or geographic areas across time. Links with geographic regions and/or (historical) subpopulations can be made based on high resolution geographic coordinates of birth place or living address and/or genetic ancestry estimates.

- Generating the meta-data needed to evaluate the feasibility of (external) research proposals and generating/linking the data needed for approved research proposals in close collaboration with the necessary participating M·3 organizations.

### *1.2.2. Data and computational challenges: Data Fusion Workbench and Data Reuse Vault.*

A multi-disciplinary research facility introduces a number of big computational and data challenges. To achieve the aspired innovative scientific breakthroughs, solutions need to be found for:

- Restrictions or delays in access to data because most data are not stored or shared according to "FAIR" (Findable, Accessible, Interoperable, Re-usable) principles in trusted storage environments: delays before databases are deposited in repositories or archives, impediments to discoverability because of meta-data limitations, access restrictions because of legal and licensing issues, lacks of interoperability because of incompatible formats and data types, interpretability problems because of incomplete context descriptions, etc.
- Merging heterogeneous data across discipline and time, which is still cumbersome and ad-hoc.
- Communication of scientific discoveries, which is still characterized by long delays in the publication process and their inability to be machine-interpreted.

Developing and deploying new technical and data infrastructures should address these issues. We distinguish two complementary capabilities of the M·3 infrastructure, which we label as:

- *Data Fusion Workbench*: the parts of the infrastructure that serve to optimally link data from heterogeneous sources and on different levels of aggregation: from the molecular level to that of the individual person and from there to aggregates of societal groups. Here we envisage the next generation of data interoperability, using linkage, merging and fusion techniques. The results of the data fusion workbench should be machine-actionable, complying with FAIR data principles. These results will, together with the source data, become part of an ever-growing:
- *Data Reuse Vault*: the parts of the infrastructure that serves the data storage and reuse functions, including the results of the research process. Here we outline the functions of this "digital archive of the future". It is of particular interest that this future archive supports researchers in different phases of the research cycle, and that data security and trust are well respected[‡]. In our vision, it will consist of physically distributed web-accessible data collections and repositories for which availability needs to be guaranteed over time.

The general aim of the technical infrastructure is to speed up the pace at which scientific discoveries are made by allowing new research questions to be answered and new scientific perspectives to shed light on existing data, while new insights and data will continuously be added to the vault. The general principles for potential solutions to these challenges are 'openness', 'distribution' and 'collaboration'. The following sections will explain how these apply to the different challenges.

### *1.2.3. Data Fusion Workbench*

With respect to linking, harmonizing and merging data from heterogeneous sources, state of the art developments are represented in projects such as OpenPHACTS, LOD Laundromat, CEDAR, and Nanopub.org. They provide semantic approaches to organizing and linking information from heterogeneous sources, using a particular way of expressing the data, usually applied within one particular domain or to a specific group of more or less coherent material. Challenges to be met are:

[‡] The trust is guaranteed by quality guidelines like the Data Seal of Approval

- Data fusion across a large semantic gap
- Provenance: data fusion under tracing and versioning
- Data fusion under partial exposure
- Privacy-preserving data fusion

Recent years have seen the rise of an increasing number of semantic data portals that are being used in earnest in science, industry and government. These capture scientific data (OpenPhacts), scientific literature (Semantic Springerlink), encyclopedic knowledge (DBPedia), economic indicators (World Bank), policy documents (data.gov), environmenta data, and many others. The largest search engine for semantic data (www.lodlaundromat.org, at the VU in Amsterdam) has harvested 39 billion statements from over 600,000 documents covering a large part of the scope that would be required for a full M·3 vision. However, current semantic fusion techniques are not able to provide a coherent, integrated and usable view across all of this data. We discuss the most important challenges that must be met to make the M·3 vision a reality.

Although these systems perform semantic integration across large volumes of data, the scope of the semantic diversity in the data these systems can deal with is typically limited to a particular scientific domain, such as pharmaceutical data in OpenPhacts, or historical research data in CEDAR. The M·3 domain, on the other hand, will stretch from genes to entire populations. Current well-understood ontology engineering techniques are not able to model such very large semantic scopes. It is likely that collaborative ontology engineering techniques must be developed, that go beyond the current practices that are based on either a single or a small number of intensively collaborating experts. We also expect Machine Learning techniques to play an increasing role in crossing such large semantic gaps, giving rise to fusion results which are plausible to various degrees, instead of the simple equality or subsumption links typically produced by the current state of the art techniques. The combination of these two (collaborative ontology engineering and Machine Learning techniques) in the form of large "social machines" where groups of humans and machines collaborate may well be the final direction that this challenge will send us to in the decades ahead.

Scientific reliability and reproducibility requires that datasets that were used to derive scientific results should be identifiable by 3rd parties. However, sophisticated data fusion processes and elaborate data science workflows typically combine, merge, filter and fuse datasets in sophisticated ways, making the traceability (or "provenance") of the origin of data a non-trivial requirement. Current provenance models (such as the W3C standard PROV-O) can describe data-dependencies up to a certain complexity and a certain degree of dynamics, but new provenance models need to be developed which must achieve greater expressivity, better scalability, and better usability by scientists outside Computer Science. The fact that these are three conflicting requirements already shows that the development of such new provenance models (and the methods that go with such methods) will pose a significant challenge.

Current state of the art semantic interoperability techniques rely on fully exploiting at least the ontologies, and often also the data items. But such full exposure is likely to break the privacy constraints which are paramount in many of the M·3 domains and data fusion is likely to be under partial exposure[§]. Unfortunately, current paradigms do not allow establishing semantic links between datasets that cannot be fully exposed to each other because of privacy risks. Entirely different paradigms are likely to be needed in the long run and need to be developed.

The current data fusion architectures will have to be generalized to a generic multi-agent setting (such as those that have been studied in Artificial Intelligence), where each dataset is seen as a single agent operating in a multi-agent environment, where data providing agents and data consuming agents negotiate over the terms on which data can be exchanged.

Provenance models would take the form of transaction traces between data-exchanging or data-modifying agents. Such transaction traces could take the form of block chains as used in bitcoin systems. Such systems combine full traceability of transactions with protection of privacy concerns.

[§] This section ensures compatibility of M·3's data fusion with data-interoperability architecture of Health Connect

Negotiating distributed data-agents can trade off the importance/urgency of a request against privacy concerns. They can also reward data providers (which might be individual citizens or societal organizations) with access to other data. Data fusion (semantic interoperability) is then framed as a multi-agent dialogue, where different agents negotiate over the meaning of their terms, exposing just enough data to reach sufficient semantic interoperability in order to perform a particular analytical task, while staying within the bounds dictated by privacy and ownership constraints. Dynamic consent policies are then seen as "bargaining chips" in a negotiation process, where agents provide data consent based on a "return for (data) value". This is in line with modern ideas on data ownership and "data as currency"[**].

### *1.2.4. Data Reuse Vault*

With respect to data archiving, the state of the art is that there is a growing number of local, national, and international data repositories in many domains and countries. One registry lists 1394 repositories (January 2016: re3data.org). However, most repositories contain data in independent and unconnected silos. Trust, quality and certification criteria, as well as preservation metadata standards have been formulated (OAIS, DSA, PREMIS), but are applied only scarcely. FAIR data principles have been defined and are generally agreed upon, but most data and datasets are at best F(indable) and A(ccessible). Data harvesting and federation through portals is in most areas in its infancy. Challenges of the Data Reuse Vault are to ensure principles of FAIR and trustworthiness in a fast-changing data landscape.

We envisage that the "archive of the future" will have a radically different constellation from the present-day situation. The data repository of today can be compared to a zoo or nature reserve where rare animals threatened with extinction are kept under protective conditions. Although data zoos will be necessary in the future also, for those data sets that nobody cares for (perhaps "data orphanage" is a better metaphor), it is a more desirable situation that the "animals" will be able to wander about without fences, that data sets are kept in their natural habitat. In order to guarantee the survival and accessibility of these data for future reuse, it is essential that the data will nevertheless be curated and managed consistently and responsibly. In order to make this possible, we envisage that the data vault of the future will rather be a stewardship system for the management of data over time and distributed over the Internet or "in the cloud", than a centralized place where data are deposited after they have been analyzed[52].

In the next sections we will discuss how this vision will enable the archival responsibilities in times of a fast-changing data paradigm and the increase of the quality and pace of scientific discoveries.

The future Data Reuse Vault needs to comply with the responsibilities of today's archives. These responsibilities are summarized as bridging time, disciplines and borders for research data.

- *Bridging Time:* The Data Reuse Vault will need to ensure that the current data and information can be efficiently and effectively used in the next periods, taking into account that the computational, data and research practices may have changed by then.
- *Bridging Disciplines*: The Data Reuse Vault will need to ensure that its data and information can be efficiently and effectively used by researchers from its originating discipline and other disciplines, taking into account that software, data structures, methodologies and language of different disciplines are different.
- *Bridging Borders:* The data Reuse Vault will need to ensure that its data and information can be efficiently and effectively be used by researchers outside the institutional or national domain in which they have been created, taking into account the different legislations, policies infrastructures and cultures.

The future data vault takes into consideration the evolving nature of research data, and how they are produced, managed, analyzed and published. Science and scholarship are increasingly data dependent, if not data driven, as was the core argument of the influential book "The Fourth Paradigm: Data-Intensive Scientific

[**] http://www.technologyreview.com/view/426235/is-personal-data-the-new-currency/

Discovery" edited by Toney Hey and others[53]. The Data Reuse Vault needs to accommodate the following changing data characteristics:

- *Increasing Volume & Velocity:* Ad hoc data selection and deposit will not be sufficient. Policies and support for automated data selection, curation and access are required. In addition, redundant storage and transfers can be avoided by bringing archival facilities to the data, instead of the other way around.
- *Increasing Dynamics and Complexity*: The structure of research data shifts from file-based to fact-based, such as "nano-publications"[54], in which individual statements are interconnected as knowledge graphs. These graphs are unbounded and continuously changing, which will require new preservation practices[††].
- *Increasing IPR and Privacy Complexity*: The increased complexity of privacy, as mentioned by the Data Fusion Workbench, will affect archival practices. The IPR matters will be further challenges by the collaborative setting.
- *Changing roles between Data and Software*: Traditional preservation relies on selected formats and software to interpret these. Preservation of generic fact-based data will be less focused on format selection, but on the preservation of the software and procedures to interpret the generic data.

Proposed Solutions for the Data Reuse Vault may include solutions as sketched below, but will also make use of to be developed technologies and algorithms.

- *New (r)evolutionary distributed data management models*. These will be based on *Onsite- or Cloud archiving*, and *Data Management as a Service*. Data Management will be integrated into the research process, enabling data stewardship throughout the complete research process and distribution of responsibilities to the natural community and environment of the data. In addition, the data does not need to be physically transferred to an archive, but remain onsite on the trustworthy storage of the research community, or on (shared) cloud-based storage. The Data Reuse Vault will remotely monitor these data and enforce preservation policies as a service. Data may need to be transferred to central archiving facilities once the researchers cannot host them anymore. The models support the management of the increased volume, complexity and dynamics of data, IPR and privacy.
- *New (r)evolutionary distributed graph preservation models*. These will be based on *Distributed Linked Data Archive* and *Linked Data Time Machine*. Linked (Open) data is essentially a global and dynamic graph of interrelated statements of facts. Effective and efficient archiving of this graph requires a linked- and collaborative approach whereby an archive preserves versions of those parts of the graph that are of relevance, and relies on collaborating archives to have preserved the related parts of relevance to them. This can be supported by techniques like Distributed Hash Tables (DHT) and Memento (web) time travel. The collaborative and distributed model supports the management of the dynamics and complexity of data and information, as well as fast/automated exchange and processing of knowledge and findings.
- *New (r)evolutionary models of durable research process repetition*. This will be based on *Process Documentation* and *Software Sustainability.* Research papers and data alone are not sufficient to replicate research. Repetition of a research process requires the process, algorithms and software to be documented and preserved. Today's technology, and the solutions proposed by the Data Fusion Workbench allow large parts of the processes and algorithms to be captured. The Data Reuse Vault will help to document these as provenance and allow the research process to be understood and repeated. In addition, the Data Reuse Vault will develop and promote guidelines and technology to improve the sustainability and preservation of academic software components (one of the challenging questions in Nationale Wetenschapsagenda[‡‡]). These models will support the changing roles between software and data, and support increased pace and quality of scientific discoveries.
- *Secure VRE (Virtual Research Environment).* This will be a collaboration with the Data Fusion Workbench to ensure that *privacy preserving data fusion* can be facilitated by the different instances (or agents) of the Data Reuse Vault.

[††] PRELIDA project (Preserving Linked Data). Deliverables: http://www.prelida.eu/results/deliverables

[‡‡] https://vragen.wetenschapsagenda.nl/cluster/hoe-bouwen-en-onderhouden-we-software-die-morgen-ook-nog-werkt

## References

1 Brown, R. R. *et al.* Interdisciplinarity: How to catalyse collaboration. *Nature* **525**, 315-317 (2015).

2 Obama, B. Remarks by the president on economic mobility. *The White House, Office of the Press Secretary, http://www.whitehouse.gov/the-press-office/2013/12/04/remarks-president-economic-mobility* (2013).

3 Monique Kremer *et al.* Hoe ongelijk is nederland? *Wetenschappelijke Raad voor het Regeringsbeleid* (2014).

4 Kawachi, I. *et al.* Income inequality. *Social epidemiology*, 126 (2014).

5 Walker, S. P. *et al.* Inequality in early childhood: risk and protective factors for early child development. *The Lancet* **378**, 1325-1338 (2011).

6 Zwillenberg, P. *et al.* The Connected World. Greasing the Wheels of the Internet Economy (The Boston Consulting Group, 2014).

7 CBS. *Centraal Bureau voor Statistiek: http://www.cbs.nl/en-GB/menu/home/default.htm*, (2015).

8 CBS. *Stelsel van Sociaal-statistische Bestanden (SSB): http://www.cbs.nl/nl-NL/menu/methoden/dataverzameling/ssb-onderzoeksbeschrijving-art.htm*, (2015).

9 Bakker, B. *et al.* The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Journal of the International Association for Official Statistics* **30**, 411-424 (2014).

10 HSN. Historical Sample of the Netherlands: http://www.iisg.nl/hsn/. (2015).

11 Jongenburger, W. *et al.* Collectieplan Meertens Instituut 2013-2018. (2013).

12 Van der Eijk, C. Design issues in electoral research: taking care of (core) business. *Electoral Studies* **21**, 189-206 (2002).

13 Todosijević, B. Transfer of variables between different data sets, or taking "previous research" seriously. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* **113**, 20-39 (2012).

14 Couper, M. P. The future of modes of data collection. *Public Opinion Quarterly* **75**, 889-908 (2011).

15 Venter, C. *et al.* The century of biology. *New Perspectives Quarterly* **21**, 73-77 (2004).

16 Boomsma, D. *et al.* Classical twin studies and beyond. *Nature reviews genetics* **3**, 872-882 (2002).

17 Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics* (2015).

18 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* (2015).

19 Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* **95**, 535-552 (2014).

20 Wild, C. P. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention* **14**, 1847-1850 (2005).

21 Visscher, P. M. *et al.* Five years of GWAS discovery. *The American Journal of Human Genetics* **90**, 7-24 (2012).

22 Keller, M. C. *et al.* Genetic variation links creativity to psychiatric disorders. *Nature neuroscience* **18**, 928-929 (2015).

23 Peyrot, W. J. *et al.* Effect of polygenic risk scores on depression in childhood trauma. *The British Journal of Psychiatry* **205**, 113-119 (2014).

24 Thanassoulis, G. *et al.* Mendelian randomization: nature's randomized trial in the post–genome era. *Jama* **301**, 2386-2388 (2009).

25 Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet* **380**, 572-580 (2012).

26 Kolata, G. Doubt cast on the "good" in "good cholesterol". *New York Times*, 5-16 (2012).

27 Mokry, L. E. *et al.* Mendelian randomisation applied to drug development in cardiovascular disease: a review. *Journal of medical genetics* **52**, 71-79 (2015).

28 Weisman, R. in *The Boston Globe* (2015).

29 Barrett, J. C. *et al.* Using human genetics to make new medicines. *Nature Reviews Genetics* (2015).

30 Abdellaoui, A. *et al.* Association between autozygosity and major depression: Stratification due to religious assortment. *Behavior genetics* **43**, 455-467 (2013).

31 Abdellaoui, A. *et al.* Educational attainment influences levels of homozygosity through migration and assortative mating. *PloS one* **10**, e0118935 (2015).

32 Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *European journal of human genetics* **21**, 1277-1285 (2013).

33 Abdellaoui, A. Behavior ↔ Genetics. (2014).

34 Price, A. L. *et al.* New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459-463 (2010).

35 Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).

36 Stoneking, M. *et al.* Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics* **12**, 603-614 (2011).

37 Veeramah, K. R. *et al.* The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics* **15**, 149-162 (2014).

38 Van Ham, M. *et al. Neighbourhood effects research: New perspectives*. (Springer, 2012).

39 Brandsma, M. *et al.* How to kickstart a national biobanking infrastructure–experiences and prospects of BBMRI-NL. *Norsk epidemiologi* **21** (2012).

40 Das, M. *et al. Social and behavioral research and the internet: Advances in applied methods and research strategies*. (Routledge, 2010).

41 Gerner-Haan, M. Mode Matters. Effects of survey modes on participation and answering behavior. *University of Groningen - Graduate School for the Humanities* (2015).

42 Schalekamp, J. *Bataven en buitenlanders: 20 eeuwen immigratie in Nederland*. (Wind Publishers, 2009).

43 Jennissen, R. Een algemeen beeld van internationale migratie in Nederland. *WODC & Maastricht University (red.), Migratie naar en vanuit Nederland: Een eerste proeve van de Migratiekaart*, 3-41 (2009).

44 Bos, V. *et al.* Ethnic inequalities in age-and cause-specific mortality in The Netherlands. *International Journal of Epidemiology* **33**, 1112-1119 (2004).

45 Selten, J.-P. *et al.* Incidence of psychotic disorders in immigrant groups to The Netherlands. *The British Journal of Psychiatry* **178**, 367-372 (2001).

46 Cantor-Graae, E. *et al.* Schizophrenia and migration: a meta-analysis and review. *American Journal of Psychiatry* (2014).

47 Kirkbride, J. *et al.* Psychoses, ethnicity and socio-economic status. *The British Journal of Psychiatry* **193**, 18-24 (2008).

48 Bolt, G. *et al.* Minority ethnic groups in the Dutch housing market: Spatial segregation, relocation dynamics and housing policy. *Urban Studies* **45**, 1359-1384 (2008).

49 Abdellaoui, A. *et al.* No evidence for genetic assortative mating beyond that due to population stratification. *Proceedings of the National Academy of Sciences* **111**, E4137-E4137 (2014).

50 Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nature Reviews Genetics* **11**, 356-366 (2010).

51 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733-739 (2010).

52 Dallas, C. The post-repository era: scholarly practice, information and systems in the digital continuum. *Springer-Verlag, Berlin and Heidelberg (abstract on https://www.academia.edu/14516809/)* (2015).

53 Hey, A. J. *et al. The fourth paradigm: data-intensive scientific discovery*. Vol. 1 (Microsoft Research Redmond, WA, 2009).

54 Mons, B. *et al.* in *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009).*

## B. Embedding of the M·3 infrastructure

### *2.1. Access to the M·3 infrastructure*

M·3 acknowledges the wishes and needs to practice open science in the public domain and provide open access to research data and outcomes. However, the nature of the data and infrastructure of M·3 poses challenges and risks with regards to privacy and sustainability that need to be taken into consideration. Privacy may not only be threatened by individual data collections, but also by the combination of multiple (anonymous) data sources. M·3 will actively participate in the (inter)national Open Access discussion and strive towards open access where feasible. Provided that privacy and data security are guaranteed, the M·3 infrastructure will be open to all national and international bona-fide researchers (be it academic, governmental, or commercial) for research with sufficient societal and/or academic impact, and aims to have results available in the public domain. Some data, e.g. those that pose potential risks for the privacy of the participants, can be made only temporarily accessible through well-secured remote access or only available at an aggregate level for meta-analysis approaches. Depending on the cost modes of the infrastructure, (international) researchers can be charged on a cost-recovery basis for access to M·3 resources, similar to the policies of for example the *UK Biobank* and the *Swedish Twin Register*.

### *2.2. Connection with existing Dutch facilities*

Dutch scientists and universities rank relatively high internationally for both social sciences and life sciences. The Netherlands houses several high quality research infrastructures that perform well on the world stage and have expressed interest in the M·3 infrastructure. In order to achieve our ambitions, the M·3 infrastructure can collaborate closely with well-established existing facilities in social sciences, such as *Statistics Netherlands* and the affiliated *Social Statistical Databases* (SSB), *Data Archiving and Network Services* (DANS), *Historical Sample of the Netherlands* (HSN), the *Netherlands Kinship Panel Study* (NKPS), and the *Meertens Institute*, and existing facilities in the life sciences, such as the Dutch academic hospitals, the *Sanquin Blood Banks*, the *Biobanking and Biomolecular resources Research Infrastructure NL* (BBMRI-NL), which includes samples from cohorts that have independently set up their own biobanks, such as the *Netherlands Twin Register* (NTR), the *Leiden Longevity Study*, the *Groningen Lifelines Study*, as well as additional parties interested in strengthening and widening their research through intensive interdisciplinary and large-scale collaborations. Besides the opportunity to combine the existing data to form high-dimensional multi-level datasets that lead to novel interdisciplinary research questions and designs, other strong incentives for existing facilities to participate is the exchange of and collaborations between top scientists from various research infrastructures and support in enriching their datasets with additional new information beyond their (sub)disciplines and the support in storing and analyzing the data while ensuring data security and the protection of the privacy of their subjects.

### *2.3. International infrastructures in social and life sciences*

Internationally, this is a unique infrastructure in the sense that social and life sciences are being integrated on such a large scale in a single and well-characterized population. For the social sciences, many governmental institutions around the world make statistical information about their population accessible for research, such as *Statistics Netherlands* in the Netherlands or the *Economic and Social Research Council* in the UK. Attempts are being made to unify European scientists from the humanities and social scientists on a larger scale in for example the *European Research Infrastructures for the Humanities and Social Sciences*[1], the *European Historical Population Sample* network (www.ehps-net.eu), the *Synergies for European Research Infrastructures in the Social Sciences* Horizon 2020 cluster project (www.seriss.eu), and through the continued development of the Consortium of European Social Science Data Archives (www.cessda.net). In the life sciences, there are several international large-scale infrastructures that we can learn from and collaborate with, such as the *UK Biobank*, *deCODE Genetics* in Iceland, the *Danish National Biobank*, the *Swedish Twin*

*Registry*, the *Biobank Japan Project*, *23andMe* in the US, and two future infrastructures in the US: the *Million Veterans Program* and the *Precision Medicine Initiative*. These infrastructures are similar in the sense that genotyping is being done on a population-wide scale with the measurement of a variety of phenotypic measurements. However, these life science infrastructures primarily focus on clinical and health related research and do not focus on including high resolution geo-coded data on social and environmental constructs from the social sciences and humanities domains. *23andMe*, which is a commercial endeavor, does offer detailed individual feedback on ancestral background in addition to health related genetic risks, which would also be feasible for the M·3 infrastructure, and likely with a higher resolution, since the relatively small ancestry differences *within* the relatively homogeneous "native" Dutch population are already being mapped[2,3], and the M·3 infrastructure offers the perspective of mapping these in more detail, and including ancestries of more recently migrated participants and their descendants.

In addition to the novel research opportunities that are made possible by the inclusion of high multi-level and geo-coded data in a genetically and phenotypically well-characterized population, the M·3 infrastructure also opens up the possibility of collaborations with existing international infrastructures in order to reach unprecedented sample sizes for current and future study designs in social and life sciences. Low-powered studies, because of small sample sizes, lead to overestimation of effect sizes and low reproducibility of results in social sciences[4], neurosciences[5], and the pre-GWAS era in molecular genetics[6,7]. A crucial part of the solution is a substantial increase in sample size, which can be achieved by combining data across centers, which lead to rapid progresses in the field of molecular genetics[8], and has successfully been applied for neuroimaging data as well[9]. At this moment, the Netherlands has potential resources for large scale studies in the humanities and social sciences, but lacks the resources to compete with future large scale life science endeavors. The M·3 infrastructure would keep the Netherlands relevant in future life sciences, while making the Netherlands a unique player on the world stage with its multi-disciplinary character and the contributions it can offer in the understanding of the human race on every level.

### *2.4. Feedback to Dutch Society*

An important responsibility of science is to inform the public about its findings. A research infrastructure that is dependent on information that effectively encompasses the entire population has the responsibility to give back to the public and to help transform and improve lives and society as a whole. In order for science to have an influence on policy through government, the public needs to be engaged first and given a sufficient understanding of science outcomes and implications[10]. Raising scientific literacy and awareness within our society also benefits the people directly by making them more empowered to make important choices in their lives. We therefore need to invest in an online and offline presence in everyday life through active campaigns that inform the general public about the broad range of achieved and expected benefits of our infrastructure. One of the aims of these campaigns should be to lead the people to a central website and/or our (social) media outlets. We should aim to have our scientists give regular talks and lectures that will be freely available on our online platforms, accessible to a broad public, and cover the broad range of subjects the M·3 infrastructure entails.

A public with a sufficient understanding and appreciation of the scientific endeavor of M·3 will make policymakers more prone to consider the evidence-informed policy options that our scientists can elucidate. Based on the robust findings that we expect from a population-size cohort, advice can be given to the government for a broad range of complex issues, such as social inequality, public health, overall well-being, economic issues, and (public) education. Therefore, research proposals should be required to have sufficient societal impact (i.e., be able to contribute to the improvement of the quality of life in the general population), sufficient academic impact (i.e., contribute to advances in methodology, theory, and basic understanding of the human subject), or both. A better understanding of children's developmental trajectories and how to incorporate this into improving children's environments and the educational system should hold a privileged place among the areas of research and feedback to society. Properly developing our children's potential is vital in maximizing the future wellbeing and economic and social success of the population[11].

Besides feedback to society and policymakers, M·3 should start the debate on whether feedback to the participants on an individual level could be desirable. Direct-to-consumer personal genomic testing was introduced in the past decade and allows for personalized genetic risk information without going through a health care provider[12]. Participants are generally able to interpret the health implications of personal genomic testing results, and generally do not experience increased anxiety, regardless of the conveyed genetic risk[13,14]. However, estimating someone's risk based only on one's genetic make-up, while ignoring family history, ethnicity, lifestyle choices, and environmental factors can be misleading[15]. The multi-level and longitudinal information included in our research creates potential for better individual feedback on health, well-being, and social and educational success. Personalized feedback that can benefit the health, wellbeing, and social success of the participants and their children, and thereby society, which would create a powerful incentive to participate in our research (it seems to work exceptionally well for the commercial research facility *23andMe*), which would in turn would improve the dataset, the research, and thereby the feedback itself. However, before such a venture can be undertaken, there are a number of risk factors[15] that should be discussed and resolved, preferably in the preparation phase of this infrastructure.

**References**

1 Duşa, A. *et al. Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. (Scivero, 2014).

2 Abdellaoui, A. *et al.* Population structure, migration, and diversifying selection in the Netherlands. *European journal of human genetics* **21**, 1277-1285 (2013).

3 Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).

4 Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

5 Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365-376 (2013).

6 Munafo, M. Candidate gene studies in the 21st century: meta-analysis, mediation, moderation. *Genes, Brain and Behavior* **5**, 3-8 (2006).

7 Siontis, K. C. *et al.* Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *European Journal of Human Genetics* **18**, 832-837 (2010).

8 Visscher, P. M. *et al.* Five years of GWAS discovery. *The American Journal of Human Genetics* **90**, 7-24 (2012).

9 Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107**, 4734-4739 (2010).

10 Gluckman, P. Policy: The art of science advice to government. *Nature* **507**, 163-165 (2014).

11 Knudsen, E. I. *et al.* Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences* **103**, 10155-10162 (2006).

12 Caulfield, T. *et al.* Direct-to-consumer genetic testing: perceptions, problems, and policy responses. *Annual review of medicine* **63**, 23-33 (2012).

13 Bloss, C. S. *et al.* Effect of direct-to-consumer genomewide profiling to assess disease risk. *New England Journal of Medicine* **364**, 524-534 (2011).

14 Ostergren, J. E. *et al.* How well do customers of direct-to-consumer personal genomic testing services comprehend genetic test results? Findings from the Impact of Personal Genomics Study. *Public health genomics* **18**, 216-224 (2015).

15 Frueh, F. W. *et al.* The future of direct-to-consumer clinical genetic tests. *Nature Reviews Genetics* **12**, 511-515 (2011).

## C. Organization & Finances

### *3.1. Organization*

The M·3 infrastructure will be set up with the support of experienced and established large-scale infrastructures from the social sciences, life sciences, and ICT domain. *Statistics Netherlands* will contribute by sharing its statistical expertise, which comprises expertise on data linkage and on data protection, including the ICT tooling as well as its expertise in data collection, statistical methodology and statistical analysis. DANS has expertise to offer with respect to data archiving, networking services, and ensuring high quality data storage and accessibility. The National Roadmap proposal *Nationale Data-infrastructuur voor de Sociale Wetenschappen* (NDSW; submitted in December 2015) lists several affiliated and well-established facilities from the social sciences and humanities involved in longitudinal mass surveys that are likely to join in our endeavours and become part of and support the M·3 Infrastructure. Possible partners for biobanking and generating biological measures are BBMRI, and commercial companies with experience in high-throughput biological measurements for a wide range of omics.

We envision the M·3 Research Infrastructure to include at least the following organizational components:

- M·3 Management Board, which includes a Scientific Director, the heads of all other M·3 components, legal and ethical advisors and international Scientific and Ethics Advisory Boards. The M·3 Management Board oversees the overall management and operation of the M·3 infrastructure and has the responsibility to ensure careful budgetary and corporate governance. The Management Board meets on a regular basis to organize and discuss the daily supervision of the different departments, and the annual plans, budget, and reports.
- An M·3 Data Access and Approval Panel; a committee that evaluates research proposals. The panel consists of expert scientists from the relevant (sub)disciplines and one or more member(s) with a legal and/or ethics background. The panel reviews each research proposal with the help of external (international) peer reviewers with the appropriate scientific expertise that are selected based on the subject(s) of each research proposal. Research questions should have sufficient societal impact (i.e., contribute to the improvement of the quality of life in the general population), sufficient academic impact (i.e., advance methodology, theory, and basic understanding of humans), or both. Requests for the extraction of biomarkers from biological samples that are limited and depletable will undergo a more scrutinized evaluation and will be carefully controlled and coordinated with the appropriate collaborating research infrastructures.
- The M·3 Legal and Ethics Committee, which is an independent committee similar to the Ethics & Governance Council (EGC) of the UK Biobank (http://egcukbiobank.org.uk/). This committee acts as a guardian of the interests of the participants and the general public. The committee is kept up to date about the activities by all other M·3 departments and independently monitors, advices, and reports on the conformity of the M·3 infrastructure with legal, ethical, and moral guidelines. The M·3 Legal and Ethics Committee reports back to all M·3 departments as well as the participants and the general public on subjects such as the benefits for society, the standards of data security, and the protection of privacy.

### *3.2. Financing*

We envision that at least the following cost components pertaining to infrastructure need to be distinguished, where the financial model for each component requires substantial funding:

1 *Guaranteeing sustainability of existing (longitudinal) datasets and cohorts*; including motivating participants to continue to take part in research efforts, to take an active role in recruiting their family members (and e.g. non-biological relatives, friends, neighbours, colleagues), and harmonizing and bringing together existing data and enrich cohort data at the individual level.

2 *M·3 development and management structure:* the infrastructure will require housing and material costs, a central facility (IT-specialists, methodologists, legal and ethical specialists, communication, administration, facility management).

3 *Setting up the M·3 Facility and Repository*: state of the art high-throughput equipment for a broad range of omics, facilities for storing samples, archives, etc. and a range of approaches for contacting and physically reaching all segments of the Dutch population and setting up labs with equipment and personal for phenotyping and genomics, transcriptomics, and e.g. metabolomics for large-scale efforts as outlined in the Science and Technical case.

4. *Setting up the ICT infrastructure* as described in the Technical Case is foreseen in three phases: experimental exploration, prototype, and implementation/management.

## D. Further Development

In the Netherlands, the current state of the various components on which the M·3 infrastructure will be built is excellent. Dutch social science stands out internationally for its expertise and innovative power in data collection and analysis. Dutch researchers have always played, and continue to play leading roles in practically all major internationally coordinated and longitudinal survey projects, including those that are on the current ESFRI and national roadmaps. Statistics Netherlands has a long tradition in operating at the international forefront of collecting, linking, and protecting data on individuals, households, firms and other organizations. Computer science in the Netherlands is internationally of very high quality, and its collaboration with social scientists is rapidly extending in the wake of the current data revolution. Finally, since 2006 DANS has been among the leading partners of data archiving and networking services worldwide and provided unique input to the quality of data storage and accessibility through its Data Seal of Approval.

M·3 is the common dream of all these involved parties. This project will open up completely new avenues for understanding human behavior – individually, in groups and in society at large – by discarding disciplinary and paradigmatic boundaries. We envision a gradual, piecemeal construction of the M·3 infrastructure over the next 10-15 years. This construction will take place along *four lines*: preparation and construction of the actual infrastructure, data protection (including the legal and ethical framework), sustainability of the infrastructure, and population coverage.

*Preparation and construction of the actual infrastructure:* The M·3 infrastructure will consist of a main physical facility that will house its central services and central laboratories (see above). The infrastructure connects a variety of data collections, including biobanks, register data at Statistics Netherlands, longitudinal sample surveys, organizational, and historical data. In the sections on the Technical Case, the requirements of the envisaged M·3 infrastructure have been addressed. These include the development of new models of data management and knowledge graph management. Issues to be addressed in the coming years include the technical and organizational aspects of the infrastructure. A guiding principle in this respect is that data will be located in various repositories, whereas the analytics (procedures for linking data and analyzing linked data) will be submitted to a central portal.

Important first steps towards the construction of the infrastructure are currently already being taken. The leaders of most of the important recurring, longitudinal mass surveys and cohorts in Dutch social science are joining forces in order to arrive at a national data infrastructure for the social science NDSW. A pre-proposal for the new National Roadmap for large research infrastructures has already been submitted in December, 2015, and additional steps are being planned. It should be emphasized, however, that this NDSW will only be a building block of the much more ambitious M·3 infrastructure.

*Data protection:* Statistics Netherlands has extensive experience with protection of sensitive (individual) data. Issues of privacy protection are expected to become more important over the next decade. Building the M·3 infrastructure implies that problems of data protection will become intrinsically more complex, as the infrastructure is based on linked information. These problems therefore need to be addressed. The legal, ethical and informational dimensions of these problems need to be considered together, and legal and ethical specialists need to work together with computer scientists and social scientists.

*Sustainability:* In the coming years, a sustainable model for the M·3 infrastructure will be developed. This model will firstly position M·3 vis-a-vis related initiatives and infrastructures. A characteristic of M·3 is that, in contrast with biobanks, it covers the population of the Netherlands (through population data or through probability samples). Taking the individual as its core unit, M·3 will connect data at the individual level with data at the sub- and supra-individual level. Secondly, sustainability also implies that a suitable financial model is developed. Such a model can be obtained by transferring a part of the costs of the infrastructure to its users, who will need to budget their intended use of M·3 in project proposals. It is furthermore envisaged that in the coming years relevant commercial partners will be invited to join the project, in order to create economic value in addition to scientific value.

*Population coverage:* As stated above, a distinguishing characteristic of M·3 is its coverage of the population of the Netherlands. It is however well-known that voluntary participation in social research may lead to biased representations of this population. Several properties of individuals, groups, and environments may enhance or impede participation in social research, e.g. surveys. Since population coverage is usually very important for arriving at statements about our society, a specific challenge for M·3 is to find new ways of involving all individuals, also those who have a diminished likelihood of participating in research. We will work on innovative ways of accomplishing this goal, building on the knowledge that has already been gained in a variety of specific research projects. We aim to increase individual involvement by, among other things, giving individuals a more active role in providing relevant information.

By systematically addressing these four lines of development over the next years, we will prepare the ground for a fruitful further development of M·3.

**Applicants**

prof. dr. D.I. Boomsma, Vrije Universiteit, Amsterdam
prof. dr. C.W.A.M. Aarts, Rijksuniversiteit Groningen/Universiteit Twente
prof. dr. F. van Harmelen, Vrije Universiteit, Amsterdam
dr. P.K. Doorn, Data Archiving and Networked Services (DANS) - KNAW
dr. K. Zeelenberg, Organization: Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS)
dr. A. Abdellaoui, Vrije Universiteit, Amsterdam
prof. dr. Pearl Dykstra, Erasmus University Rotterdam
dr. Ruurd Schoonhoven, Statistics Netherlands
prof. dr. Eline Slagboom, Leiden University Medical Centre
prof. dr. Maarten van Ham, Delft University of Technology
prof. dr. Kees Mandemakers, International Institute of Social History
prof. dr. Arnold Bregt, Wageningen University and Research
dr. Maarten Hoogerwerf, Data Archiving and Networked Services
prof. dr. Herbert van de Sompel, Los Alamos National Laboratory
prof. dr. Aat Liefbroer, Vrije Universiteit Amsterdam
dr. Ruben Kok, Dutch Techcentre for Lifesciences (DTL)
prof. dr. Gert-Jan van Ommen, BBMRI-NL
Prof. dr. Cisca Wijmenga, Universitair Medisch Centrum Groningen / BBMRI-NL
prof. dr. Hans Bennis, Meertens Instituut
prof. dr. Pieter Hooimeijer, Utrecht University
prof. dr. Wouter van der Brug, UvA
prof. dr. Cornelis (Kees) Kluft, Good Biomarker Science, Leiden
prof. dr. Marcel Das, Tilburg Univ /Centerdata
prof. dr. Hans Schmeets, Maastricht Univ / CBS
prof. dr. Roel Boschker, Groningen
Nationale Data-infrastructuur voor de Sociale Wetenschappen (NDSW)