Research Council of Norway – Research Infrastructure 2012 – Resubmitted project # 208760

National Historical Population Register for Norway 1800-2022 (HPR) The University of Tromsø, Faculty of Humanities, Social Sciences and Education: professor Gunnar Thorvaldsen, Norwegian Historical Data Centre (NHDC). Phone +4777644179, gunnar.thorvaldsen@uit.no

Category

	Advanced scientific equipment/facilities (2-200 mill NOK)
x	Scientific databases and collections (2-200 mill NOK)

1. Relevance

- **In relation to call for proposals**
  The Historical Population Register (HPR) will give rise to unique new opportunities for research on a large array of topics spanning medicine, social sciences and humanities. The most relevant scholarly fields are epidemiology, genetics, public health, demography, economics, history, sociology, onomatology and dialectology. For medical research the register will be extremely useful in epidemiological studies of causes of disease and more general public health issues. In particular, the genealogical data will be crucial to learn more about hereditary causes of complex diseases and fully exploit the potential of the rich biobanks, health registries and cohort studies already found in Norway. The construction of the register will involve important challenges within the field of informatics. The HPR will utilize supercomputers to link data files with fuzzy, nominative information. It will develop advanced procedures for efficient extraction of information from hand written sources, building large and complex information systems such as a wiki-based site for cooperation among thousands of genealogists via the Internet. In the more contemporary setting it will give different user groups access to data about individuals, in each case based on specific ethical vetting procedures.
- **In relation to societal challenges**
  The diversity of possible applications of the HPR limits this discussion to only address some main examples where this research can help resolve major societal challenges. To the extent that the data can answer questions about heritable and non-heritable causes of disease it is obviously of crucial importance to public health. Having access to rich genealogical information can in itself also be used in prevention of illness by better targeting sub-populations at high risk for specific diseases. By facilitating research that provides a better understanding of major processes driving societal and economic development, the HPR can also give rise to important implications for a broad set of future policy recommendations. Furthermore, the HPR will contribute substantially to the understanding of Norwegian history. The HPR will also bring important added value to society as a cultural public good for genealogists and others interested in their ancestries. In particular, the HPR can give rise to substantial cost reductions in the production of community histories, and it can give interesting new approaches to the teaching of local history. As is said in the Research Report 2009 from The Ministry of Education and Research: “Globalization and the development of a multi-cultural society also demands more knowledge about languages, culture, history, religion and identity” (Introduction 1.2).

2. Vision and scientific goals

- **Long and short-term scientific goals**
  The long-term goal is a nationwide Norwegian population register for the time period since 1800, obtained by merging and linking information in the nominative censuses with church books and other protocol data. The HPR will be a national pillar for empirical research in a broad array of research fields, and hence benefit numerous national research institutions. The

short-term goal is an expanding population register covering an increasing number of representative regions during this period by linking information in currently transcribed and additional computerized sources. During the pilot project period financed by the Research Council we started to install systems that make future addition to the register more cost efficient and facilitate and speed up the ongoing inflow of contributions from volunteers to the National Archives. We (the project group) focus on how the HPR can best facilitate the needs of medical research, in particular in connection with the Biobank Norway cooperation, and by working on extending the separate Causes of Death Registry backwards in time.

- **Expected impact on Norwegian science, technology and innovation in an international perspective**
  The HPR will become a pillar for all types of research on the Norwegian population performed by researchers all over the country. Together with a few other countries, Norway has access to rich population data of high quality for the contemporary period. A major possibility will be to merge information from extensive registries, surveys and biobanks by use of personal id numbers. Norway has one of the world's best coverage of nominative protocol data in censuses and church registers from the last two to three centuries. So far, researchers have not utilized these sources much because there only exits a modest number of modern transcriptions which are also quite heterogeneous and organized separately. The establishment of the HPR will enable Norwegian researchers to take full advantage of our rich contemporary data *and* historical population data in order to bring them into an internationally unique position. Importantly, information from the various sources can be studied jointly, so there is a huge potential for the modern registries, biobanks and the HPR mutually benefitting each other. Only Iceland presently holds similar data with long-term coverage, but their private enterprise deCode database has other limitations that reduce its value for external users. Thus, the HPR will be an international spearhead for many types of population-oriented research. Many international research groups will be interested in collaborating with Norwegian researchers with legal access to and experience in using the HPR. In particular this will apply to epidemiological research.

3. Scientific and technological status

Given the wide spectre of disciplines likely to employ the HPR we will in the three first bullet points limit the discussion to the research status in production of historical population data.

- **Status of current research and technological development, main scientific challenges.**
  As already mentioned, Norway has rich population related microdata for the period after 1960. For the earlier period, relevant datasets with more than five million records have been transcribed (primarily censuses, but also approximately 10 percent of the church books). There is a large potential for making these available in more rational formats for researchers through "academic secondary refining" as recommended in the evaluation of Norwegian historical research (Forskningsrådet 2008). Much less work has been done in Norway on constructing long-term longitudinal databases for historic periods that are also suitable for scientific research. Currently two local and long-term longitudinal databases are available for researchers: Rendalen covering 1730-1900 (Bull 2005) and Asker covering 1800-1878 (Fure 2000). In the NFR # 224 pre-project these have been extended and made compatible. A similar database for Etne parish resides in the Regional Archive in Bergen and will be added. Interesting research has been carried out with these, but the populations covered are not nationally representative and they are usually too small to allow powerful use of statistical methods. Several local longitudinal databases are in the pipeline. Most promising are the *Busetnadssoge (Community History)* BSS-databases developed by partners for this application. These data sets are primarily developed for the printing of local genealogies, but we have developed a prototype for exporting from its proprietary format to a format more suitable for research. Nationally, linked data with more than two hundred thousand persons are available in such databases. A sample population is already available, and we plan to use BSS as a platform for converting the available community history databases to the HPR. Internationally, researchers have access to several local samples of historical

longitudinal population data, and there are countless examples of scientific studies employing these. Notable is the research connected to the Demographic Database in Umeå, and the EurAsia collaboration (Bengtsson et al, 2004). All known examples of such data are restricted to historical periods only. The national scope and the ability also to link to modern register data will be a crucial and unique feature of the HPR.

The biggest challenge is to link individuals across different sources, such as transcribed censuses and church books, electronic community histories, and the many electronic genealogies. Partial automation is feasible today due to development of sophisticated record linkage software (e.g. Christen 2008), and access to a lexicon of standardized names in Norwegian censuses (NFR project 164186). Better organization of the cooperation between researchers, community historians and the network of genealogists organized in DIS-Norway will complement the work done by the HPR partners, employing the more rational transcription techniques which we are developing. The National Archives are in negotiations with commercial, international genealogy companies which by exchanging data files will reduce our dependency on voluntary transcribers and make the completion of the project realistic even if the NFR contribution is reduced to 25 million kroner – an assessment based on operating contracts with similar projects in the USA and Great Britain.

Another challenge is defining relationships between indviduals, especially between children and parents, which is essential for many research projects that will be based on data from the HPR. To do this for one family may often require data from *several* censuses as well as parish records. We have access to advanced software for this purpose developed at the Minnesota Population Center.

- **National and international "state-of-the-art" in the relevant disciplines, technologies and research themes, requirements for frontier research.**
  Our work and plans meet the recommendations from the evaluation committee (cf above) to be more internationally oriented. In NAPP (North Atlantic Population Project) our transcribed and encoded census manuscripts have been integrated into a database together with sources from Sweden, Great Britain, the USA, Canada Germany and other countries for comparative research across time and space, cf nappdata.org. The record structures and coding schemes have been harmonized and variables that are common across national borders have been encoded in the same way. Nationally unique information is included in additional variables. Norwegian censuses are the only non US ones to be linked in the NAPP database. Marianne Erikstad of the Norwegian Historical Data Centre (NHDC) participated intensively in the development of the HISCO standard for the international encoding of occupations in historical sources which is implemented in NAPP. The longitudinal database we are developing will initially be based on the same relational principles as those applied at the Demographic Database, Umeå University (Miller and Thorvaldsen, 1997). However, an Intermediate Data Structure (IDS) for the dissemination of longitudinal databases is being developed by the Historical Sample of the Netherlands and the Interuniversity Consortium for Political and Social Research (ICPSR) in Michigan (Alter et al 2009). Especially in the international context the IDS standard is becoming the future vehicle for disseminating longitudinal data to researchers. Support from the ESF and NFR already allows the University of Tromsø (UiT) to take part in this development and an IDS prototype of linked Norwegian census data is available from the HPR pre-project.
- **The scientific environment and development of Norwegian research activities in the field, relevance with respect to the priorities given in the call for proposals.**
  The infrastructure called for in this application will expand and strengthen networks for comparative medical, information science, demographic, historic and other types of population related research in Norway and internationally. The institutions behind these plans have cooperated to assemble, transcribe, refine and research population data sources from numerous perspectives for decades and have gained extensive experience making such source material available for researchers. Transforming this material into a joint database requires us to organize our efforts more tightly, creating a networked expertise in the field of source-oriented information

science also involving supercomputers and wiki techniques. Of special importance in this network is our co-operation with the Demographic Database at Umeå University, a world leader in the computerized creation and employment of longitudinal historical records in research. Some of us (Nygaard and Thorvaldsen) have gained first-hand experience from working with their databases.

- **The level of existing research infrastructure, the need for the new investment.**
Since the HPR will benefit a multitude of researchers in a broad specter of academic fields, we can only outline the importance of this investment. For the contemporary period Norwegian researchers in the social sciences and medicine are today blessed with access to rich information at the individual level on a large array of topics from sources such as administrative registers, health registers, biobanks, and a large variety of extensive surveys. For much of the research using these data a critical limitation is the ability to follow central processes over long enough periods to assess seminal changes. A fundamental obstacle is the lack of a population register and personal id numbers before 1960. Moving this limit backwards would be a crucial addition to the analysis of the existing data. As an example, the Causes of Death Register today extends back to 1950, but the first decade of data has limited value because it cannot be linked to a population register and the census of 1950 when the need arises. For the investigation of genetic relationships it is highly desirable to know more about family relations and pedigrees than what can be gleaned from existing registers. The same obstacle applies to studies of inter-generational processes generally, e.g. in studies of social mobility. Obtaining such data at a significant scale in a small country demands the national population register described in this application.

  For research in the humanities and the social sciences it is critical to build the HPR to properly assess a multitude of research questions about social and economic development also for the period before 1960. Obtaining detailed knowledge of social and economic processes covering Norway's transition from a poor undeveloped society in 1800 to the present rich welfare state may give crucial insights of relevance for developing countries. Given the keen interest in~~for~~ the Scandinavian welfare model internationally in fields such as economics and public health, HPR-based research will help assess the model's development and functioning.

  The need for this investment is also of a more practical and economic character. At present much of the information to be included in the HPR is already in place, but is fragmented and cannot be used efficiently by a wide audience of researchers. In addition, a substantial amount of voluntary efforts already complement professional activities enhancing this type of information (production of genealogies, transcription of source material, etc). The facilities planned under HPR for promoting these activities will give substantial benefits through exploiting economies of scale.

- **Scientific ambitions and relevance with respect to the existing national and international research agenda/strategies.**
By providing genealogies extending over several generations the HPR can become a valuable complement to the information found in biobanks for genetic and epidemiological research. It can also make the future collection of biological samples more cost-efficient by concentrating on selected pedigrees. The HPR is thus likely to complement the scientific value of the biobanks funded by the large-scale infrastructure grant to Biobank Norway. Similar added value is relevant for research in bioinformatics and population genetics under the SEQ project, also this a major grant from the Research Council.

  In economics there is keen interest in the data contained in HPR from ESOP at the University of Oslo, a Center of Excellence funded by the Research Council. In Statistics Norway the FDTrygd is a database on demography, employment, social insurance and income for all inhabitants in Norway since 1992, which is used extensively by Norwegian researchers in social sciences and medicine. This is also the case for the Population and Housing Censuses which have been merged into one file for 1960 – 2001.

  The future Internet is central in both the VERDIKT programme in the Research Council and EUs framework programs. The HPR may become larger than Wikipedia in the number of

pages, have geotags for all farms in Norway, become one of the most used Internet sites in Norway and open for mass contributions in an advanced semantic structure. Thus, the HPR will contribute to forming the future Internet. Our US partner utilizes supercomputers in the linking, and the application of similar national resources and grid technology will be considered.

4. Description of the Research Infrastructure

- **Scientific, technological and physical description of the new development and its relation to existing national and international research infrastructure.**
  The national HPR is a joint venture planned by a number of Norwegian partners listed in paragraph 7. In addition the project will cooperate with many other institutions in Norway and abroad, cf the letters of support. The population register will be a database where information from censuses, church books and other relevant sources are linked together so that all residents can be followed across time and place, and for most individuals also have detailed knowledge about family ties. The information will be compatible with modern registers through personal identification numbers, but for protection reasons these numbers will not be the same as the currebt numbers used in the CPR. The register will primarily include information about identifiers (name, date of birth, parents, spouse, place of birth and of residence, etc), which will allow linking with other data sources with information on education, occupation, health, etc., depending on the research to be conducted.. Currently only local parts are implemented, but the aim is to cover the whole country for the period from 1800 to 1964 when the Central Population Register (CPR) takes over. Where source material and local projects allow it, the HPR shall also cover the period before 1800. During the decade 2014 to 2024 we shall continue implementing an advanced database structure allowing the coverage of data about virtually all Norwegian citizens and their ancestors. Norway's population grew from 880,000 in 1800 to 2.1 million in 1900 and by now to 5 million inhabitants. The register will include 9.7 million people, based on 37.5 million entries from censuses and church records which have virtually complete coverage. Record linkage rates will vary by period and region from two-thirds to over 90 percent of the records, and with national coverage overall linkage rates should come close to or exceed 90 percent. An important advantage of covering the *entire* population is that this will in itself improve HPR quality because researchers currently have better long term grip on emigration than internal migration. Therefore, a national register will have fewer incomplete life courses than the current local databases, and linkage reliability will especially increase among internal migrants. Our emigration registers have been computerized and will be integrated into the population register. Given the necessary resources we are confident that in a decade Norway will be the first country to provide an integrated HPR covering a population significantly larger than the Icelandic. The censuses include detailed information also on the place of residence (often a farm). Using the system for identification of properties (Matrikkel) both persons and residence units can be followed in parallel. This will increase the quality and interest in the register significantly. Both institutionally and in terms of joint personnel, the HPR project has close cooperation with an ongoing project headed by the Norwegian Mapping Authority developing a database covering the development of administrative divisions in Norway from 1660 to the present. As a spinoff from this project, a national longitudinal register of properties (Matrikkel) from ca 1838 to the present is well under way (Bævre, 2010). The HPR will cooperate closely with these projects and will be fully integrated with the property register, adding geography to history. For work on urban residences we cooperate with the BergGIS project at the University of Bergen.
- **Description of the needs and relevance with respect to the technological and scientific challenges.**
  In the 2010-2014 strategy plan for the National Archives (NA) scanning of paper originals of archival material is given high priority. Much of this is relevant also for the HPR. However, even after scanning some recent material cannot be published for decades due to legal reasons, but only made available for researchers on application. This applies to the 1920, 1930, 1946 and 1950 censuses, the baptism lists in parish registers from 1930 onwards and parts of the burial

lists. It is difficult to give this non-disclosable material high internal priority in the NA because the many non-statistical users cannot be granted access, so this scanning has to be financed by the HPR project. It is still natural that the NA carries out the scanning of these sources, because they have the material in-house, most of the necessary equipment and the know-how. The scanning of the four censuses mentioned above for the whole country on a batch scanner is estimated to require a total of 13 work-years. The scanning of the main vital or parish registers 1910-1960 is estimated to require a total of 3 work-years. The scanning is expected to be distributed over 5-10 years. The HPR partners will rationalize transcription procedures, fund two five-year transcription experts for the 1930 to 1950 censuses and construct software for controlling the quality of the database.

We intend to make the BSS software system available to interested municipalities in order to stimulate local activity, including the transcription of sources using other HPR modules before importing data via BSS. The UiT and BSS intend to establish a service to help such users, based upon full cost coverage principle presupposing their final BSS database will be made available to HPR. Further development of the BSS software is undertaken to do this work.

- **Specify whether the proposal concerns a new research infrastructure or an upgrade of existing research infrastructure, whether the research infrastructure is a national facility or part of an international network of connected and interoperable research facilities, and whether the research infrastructure is localised, distributed and/or electronically accessible.**
  The HPR is a new electronically accessible infrastructure building upon electronic versions of source material and connected to the existing Central Population Register 1964-2012. It will extend present registers for local areas to national coverage. The HPR will be distributed between a public access database for the period up to 1919 and a bona fide researchers only database for the period 1920 to 1964 and beyond. (When the HPR has been completed in 2022, the open part will extend until 1929.) The former will be built at the UiT, NA and NR, the latter in the National Archives in cooperation with Statistics Norway. The NA already builds the HPR for the period 1801 to 1815 with assistance from the UiT and NR. Through the ESF-supported European Historical Population Samples Network (EHPS-Net) the register's open part will be shared with researchers internationally by means of the Intermediate Data Structure (IDS) which already accommodates Swedish and Dutch longitudinal data. The modern part of the HPR with limited access may also in the future be distributed through a remote access system such as the STAR network at Statistics Sweden or the data enclaves run by Statistics Canada. Statistics Norway (SN) and the Norwegian Social Science Data Services (NSD) are developing the RAIRD system for this, funded by the NRC Infrastructure Program.
- **Describe the national character of the new research infrastructure.**
  The scope of the HPR is national because it will cover the whole nation, because all relevant institutions are involved in its construction, and because it will become a pillar in many types of research on the Norwegian population performed by researchers all over the country. The HPR will cover virtually the whole population present in Norway since 1800. Only a minority (less than 10 %) is expected to have no coverage or be mentioned in only one source.
- **The main user modes and research disciplines of the intended users.**
  Historians and name researchers will use the open part of the population register online in order to identify predefined groups of individuals. For simple statistical use they can generate aggregates online, while advanced statistics will necessitate the downloading of data with subsets of variables and population segments. For the modern part of the registers this will necessitate thorough vetting by ethical and judicial criteria. The same goes for researchers in medicine, economics and social sciences, who will usually need to access the closed parts of the HPR. Genealogists will access the open part via the HPRwiki or the existing user interfaces of the Digital Archive and the NHDC. Internationally, users will access the open statistical part through the Minnesota Population Center or IDS files containing subsets of periods, variables and popula-

tion groups. Due to legal restrictions the closed part will normally be available for international research only in cooperation with Norwegian research institutions.

- **Specify possible needs and requirements for the use and development of e-Infrastructure, such as resources for data storage, tools for data handling, electronic services and communication.**

Building and operating the HPR will not be very demanding or expensive when it comes to purchasing hardware, net access or commercially available software. Full scale running of the public HPR-wiki, which is expected to be very popular and heavily used, may require one or two new servers costing about NOK 50,000 each. The only expensive piece of equipment is a fast scanner costing about 1 mill NOK. All the central partners are already connected to Uninett, the fastest fibre optical data network in Norway. When it comes to software, the National Archives has good experiences with, and rely on open source products like the Red Hat Linux operating system, the MySQL database system, and the Perl and PHP programming languages. The NR has developed the wiki -program directly on MediaWiki using php and MySQL instead of building on the American WeRelate. This gives larger flexibility, better control over the software and better adaptation to our project. A prototype of the program is already developed and available at slekt.nr.no. The prototype was developed in the pre-project and is now extended for the NA as part of the 2014 celebrations. -Also the HPR's limited access research database (1910-1964) will probably be operated in MySQL.

The creation of a longitudinal population register is feasible today with freely available record linkage software such as FRIL or FEBRL (Christen 2008). Once the static source information (date and parish of birth and death, gender, nationality and names) has been encoded, this software uses these variables to identify record matches. So far the pilot project has primarily developed methods for linking the censuses. At the NHDC more than half of the persons in the 1865 and 1875 censuses for Troms were linked with the FEBRL software. In order to test the reliability of the results, special routines were written in PL-SQL to test intrafamily links. These linkage routines have been extended at the University of Tromsø, making record linkage between the censuses and the church records possible. The Minnesota Population Center has linked several census samples for the US and Norway which have been made available to researchers as part of the NAPP project. The NHDC has also tested the FRIL record linkage program and found it well suited for linking smaller data sets such as the Rendalen and Asker databases. These efforts and the linking of the census samples from 1910 through 1950 with new SQL techniques are described in the report from the HPR pre-project.

- **Critical factors of the project.**

For the last five decades Norway has maintained a longitudinal population database called the Central Population Register, but for earlier decades the sources are either in manuscript form or exist as computerized stand-alone transcriptions. After linking historical material in electronic longitudinal formats for the last two or three centuries we can conclude that it is realistic to build a national population register extending backwards in time, at least for the period from 1800 to the present. This national register will be based on the new integrated data model specified in the pre-project, building on the rather diverse longitudinal and cross-sectional data sets which are progressively made available. The fundamental entity in the register will be the *individual* (with an ID and the in principle stable information of names, date and place of birth, date and place of death, gender and nationality), source-documented *events* in their life course (type, time and place, plus time dependent individual information such as event role, residence, marital status, family position and occupation), and *relationships* (especially parenthood, which may be used to derive other relationships between the individuals).

Graphics routines are to be developed for more efficient handling of handwritten protocol data. Several procedures are investigated in an online report from the Norwegian Computing Center (Eikvil et al 2010). While ordinary optical character recognition (OCR) is generally not feasible for handwritten material, most digits can be interpreted automatically (cf zip codes). Also, software exists which divides protocol data into separate rows, columns, fields and words.

The latter can then be grouped by priest or census taker, so that similar images can be transcribed en block. Our pre-project implemented and tested two such techniques successfully and this has now been extended to numerical and text fields in the 1891 census. We expect these techniques to at least double our current transcription rates, after their implementation has been fully funded.

Major parts of the existing databases have been exchanged internationally, and the new database will be made available through cross-national formats which are being developed together with our international partners. The oldest parts of the database (up to 1919) can be made available with documentation in Norwegian and English on-line, but the closed part will only be made available to eligible researchers in Norway upon ethical and legal approval. Only in special cases will identifiers such as names be made available. Extracts from the current Central Population Register will be anonymized according to Statistics Norway's (SN) norms after linking has been made with the pre-1960 data set and before distribution to researchers. Thus, the record linking will be done partly at SN and the NA, and partly at the University of Tromsø and the Norwegian Computing Center for open data sets. Development copies of the longitudinal data sets will be held both by the NHDC in Tromsø, the National Archives in Oslo and Statistics Norway. Long term storage of the database will be in the NA and SN.

The major challenge is to link the heterogeneous information found in various sources about individuals and families over time, even as people migrate and information may be inconsistent. While this is done via personal ID numbers in modern registers, special record linkage software must be employed for the period before 1960. Nearly half of the men and couples in the 1865, 1875 and 1900 censuses for Norway were successfully linked by the Minnesota Population Center, a better result than in the corresponding US censuses, and we expect to improve the results due to our more detailed knowledge of the sources. The material has already been used by economists to substantiate that Norwegians who immigrated to the US improved their income compared to siblings who remained in Norway (Abramitzky 2009, Vick 2010). Our own record linkage involving church records show even better results. Thus, our experience with such procedures, including name standardization, shows that automation is realistic for the majority of the individuals, but additional manual linkage must be done for the remainder. All efforts made by community historians and genealogists can be put to good use in this regard through our wiki-based application on the Internet, cf slekt.nr.no. The pre-project has investigated how the family oriented results of record linkage can be applied in order to identify and remove false links and duplicates. Future extensions of the data set is continuously done by the register owner, the Directorate of Taxes, and by Statistics Norway.

Data sets with information about persons who are still alive, and *sensitive* information about deceased persons, must be handled according to strict guidelines to prevent accidental disclosure. These guidelines are based on the Law of Public Management, the Personal Data Act and Regulations, the Statistics Act, and indirectly on the Law of Archives.

The HPR-wiki (1800-1920) is open to the public, and may be edited by authorized users by means of interactive record linking software. All changes are logged as an integral part of the wiki based software. Safety against deliberate manipulation and accidental or technical destruction of the database will be taken care of by routines for traditional, daily backup and regular security exports.

Since the long-term goal is to store data on the entire population of Norway during last centuries in the HPR, the statistical *representativity* of the linked part of the individuals and families represented there, will in the end cease to be a significant issue. While the population register is being built, however, it will necessarily contain only parts of the population. For reasons of research economy and in order to test the database in ongoing research there are methods to ensure that the biases of any skewed sample will not seriously distort research results. For this purpose, the Minnesota Population Center and others have developed measures to estimate to what degree the linked persons are statistically different from the whole population as represented in the decennial censuses, whether nominative or statistical. These measures build on the

most central variables such as gender, age, marital status, occupation and birth place, and make it possible to correct statistical bias based on a partially linked population register.
For non-statistical research, e.g. genetic studies following families over time, a complementary method must be used, employing only the parts of the population register where the links have been substantiated beyond reasonable doubt, in order to make sure that false genealogical links do not invalidate the tracing of e.g. inherited deceases in the pedigrees. For this purpose the quality and origin of all links will be flagged in the database.

5. Impact on science, technology and innovation

- **Analysis of the new research opportunities for the Norwegian and international scientific community.**

Comparative research in all the scholarly fields mentioned above will profit from integrating the continuous information in the Norwegian church books with the cross-sectional information from the censuses. This will make all kinds of population related research more efficient. For a host of research questions in the humanities, social sciences and medicine, a longitudinal structure opens up possibilities in three new ways: 1) The period before 1960 will be opened up for longitudinal research on a national scale. 2) The analysis of recent social and other population phenomena can be based on data with a longer time horizon. 3) It is of crucial importance to be able to follow individuals, kinship and genetic networks over extended time periods.

As is seen from the use of the census series 1850 to 2000 available at the Minnesota Population Center, extending the current databases throughout the 20th century will open up for the use of historical data by researchers primarily interested in contemporary issues, and the use of more modern data by historically oriented researchers, promoting cooperation across scholarly fields. This is because these data series are compatible across time and space, so that extending the scope of the investigations can be done with small extra methodological efforts.

Access to detailed and extensive information on family ties and pedigrees is crucial for many applications of population data in the social sciences and medicine. On a large scale this information can only be derived from a population register extending several generations back in time. Many social phenomena "run in the family": education, occupation, income and wealth, social status, company ownership etc. For instance, Jåstad (2010) has researched ethnically different heritage patterns. With longitudinal data sets researchers can develop both generational and gender perspectives on long-term social change. In studies of the heritability of different diseases this type of information is crucial. For rare diseases where one seeks to identify specific mutations, information on families can be of considerable value when conducting segregation analysis of biological samples from biobanks. For studies of complex diseases in genetic epidemiology, knowledge of pedigrees allows for construction of more efficient case-control designs when seeking candidate genes from sequencing of the genome. Using statistical methods genealogies from the HPR in combination with modern health registers can give valuable insight about heritability also in cases not covered by biological material. A major obstacle in this type of research is sample size, and the HPR will be a unique addition in terms of its large size. Occurrence of some diseases, such as mental disorders, will be documented far back in time in burial protocols and censuses. The HPR opens up exciting possibilities for tracing such diseases through generations (Andersen and Hynnekleiv 2007). Such sensitive data will of course only be given to researchers after thorough vetting following all relevant ethical and legal procedures. Both for social and medical studies of inter-generational relationships it is valuable that the HPR includes extensive socio-economic information, particularly from the censuses. This opens up many possibilities for investigations of relationships between genes and the environment and the relative importance of these factors.

- **Impact on the Norwegian research in the field, contribution to the development of the research discipline, and on future possibilities.**

Most population and health related fields discussed in this application are relevant both in a Norwegian and international setting. The HPR will give Norwegian researchers access to inter-

nationally unique data, thus opening up new possibilities for conducting innovative research at the international research frontier. Access to such data will most likely result in more international researchers seeking cooperation with Norwegian researchers, which will provide valuable stimulation of national research groups. The unique position of Norwegian data will probably also generate substantially more international research on the Norwegian population and society. Finally, the HPR leans on Norway's strong tradition in community and regional studies. There should be powerful synergy effects inherent in bringing so many disciplines and place-related experts together to work on a national population register.

- **Impact on the recruitment to science.**
  The wide scope of the research based on the HPR means that we cannot discuss at large the potential recruitment to work on this unique resource both from domestic and international rosters of candidates. But it ought to be mentioned that the generation of researchers who started their work on the source material and population oriented issues in the 1970s approaches retirement. The HPR project will allow thorough documentation of the technical sides of their work and continued cooperation with younger colleagues who have, e.g., been trained in demography at the University of Oslo, and who cannot wait forever for demography-relevant work. Among researchers who have expressed interest in working to create or employ the register in research are Eli Fure of the National Archives, Hanne Marie Johansen at the University of Bergen, Hilde Jåstad, and IT consultant Karen Bjørndalen, University of Tromsø.
- **Impact on internationalization of Norwegian science.**
  Longitudinal population registers are gaining ground in many European countries, making *questionnaire based* censuses a thing of the past. The historical registers available for selected regions in Sweden have attracted many international researchers, cf for instance the web pages at www.ddb.umu.se. An open population register with national coverage will be attractive because it will allow the study even of marginal groups without running out of numbers and allow the tracing of migrants both domestically and in the fully transcribed emigration lists. In order to utilize the integrated HPR rationally, foreign researchers will want to work with Norwegian scholars, both in order to grasp the Norwegian context and get access to the protected parts of the database. Articles about the HPR have already been published in US and Russian journals and an article in the Spanish *Revista de Demografía Histórica* is forthcoming.
- **Impact on future innovation, value creation and national competitiveness.**
  As far as we are aware, the graphics routines the NR develops for more efficient handling of handwritten protocol data are unique. They build on the automatic recognition of digits present in dates, id numbers and codes, and are extended into semi-automatic handling of frequently occurring words such as names written with the same hand style in the sources with clustering techniques. Although these are not yet handling handwritten *free* texts, automatic recognition of handwritten words in protocol data is becoming an important field within information science during the next couple of decades. In addition, the potential linking of the Norwegian biobanks with ancestries can become a world class repository for studying the heredity of diseases.
- **Examples of potential results, other possible use of expected results, contribution from and benefits for all partners.**
  A longitudinal database encompassing the Norwegian population in the nineteenth and twentieth century will open up new terrain in disciplines such as history, economics, demography, social medicine and sociology. The censuses, parish books and other nominative sources include extensive information on demography and social structure that can only be fully utilized through the creation of an integrated longitudinal HPR. The nineteenth and twentieth century form a critical period in the study of medical advances, fertility decline, urbanization, international migration, household composition, occupational structure, and gender equality. The database will allow statistical modelling on a wide range of topics that have not been covered by census publications or have been incompletely tabulated. Even more important is the potential for longitudinal and multilevel multivariate analyses opened up by the availability of the database. A longitudinal database will constitute an invaluable resource in its own right by enhancing the value

of both previous and current historical microdata samples. Used in combination these microdata will constitute an invaluable resource for studying the development that led to our contemporary society. The paragraphs that follow sketch only a few of the most obvious research applications of the HPR.

Public Health. For many common diseases it is thought that the causes reside in interplay between genetic factors and environmental effects. This implies that persons who develop the disease, for instance cancer, heart disease or chronic rheumatic diseases carry certain predisposing genes and that disease occurs when one is additionally exposed to certain environmental stimuli, such as an infection or a component of the nutritional intake. Recent genetic studies suggest that many chronic diseases are genetically heterogeneous. Many different genes may give rise to the same phenotype. Some of these variants may be regarded as rare mutations only occurring in one or a few families. In order to disentangle the vast genetic variation in today's available DNA data and tomorrow's more detailed sequencing DNA data, the study population must be reconstituted into families, as can be done with the new HPR. The familial disentangling of the genome together with information from linked cohorts and health registries will promote the understanding of the genetic basis of diseases, and subsequently understand the biological mechanisms that will lead to better prevention and treatment.

Social sciences. The time window from 1960 to the present, for which there is rich access to data on population related issues, cover only the latter phases of several processes involving major changes to society. By expanding this window researchers can learn considerably more about the reasons for and consequences of, e.g., increased female labour participation and higher levels of education. When it comes to aging, it is evident that this process cannot be satisfactorily studied without data covering several decades. An important feature is the ability to follow families and households over several generations. This allows for a multitude of new approaches to the study of family and household organization, social mobility and other inter-generational processes such as transmission of education investments. Even when the research interest lies exclusively in present day phenomena, the access to family and generational data from the HPR can be extremely useful. For example it can shed light on the role of family background in connection with studies of labour market outcomes, or social differences in mortality and health outcomes. Also in the social sciences there is a substantial interest in genetics. The HPR will give rich possibilities for studying the interplay between environmental and genetic factors in a host of outcomes of relevance to the social sciences (e.g. cognitive abilities, political preferences, crime).

Fertility transition. During the late 19th and early 20th centuries Norwegian women and men started deliberate fertility limitation. A historical population register will allow the study of differential fertility patterns in this critical period of demographic transition, to assess the importance of such factors as occupational class, region, literacy, local economy, size of locality and family structure. Studying these shifts in population structure has the potential to enhance understanding of ongoing demographic change in the contemporary developing world. Aggregates do not allow controlling for individual-level socioeconomic characteristics, and the fertility pattern of families cannot be followed through the reproductive period of the mother only. With a longitudinal database demographers will be able to study the starting, spacing and stopping behaviour of families directly. Thus, the database will allow a new and more sophisticated generation of comparative studies of the first demographic transition.

Industrialization and economic developments. By the nineteenth century most of Norway was affected by industrialization. The HPR will allow unprecedented opportunities to explore economic structures within Norway and with other nations during this and later transitional periods. For the first time, it will be possible to follow people at the individual level over generations throughout the country, with consistently coded occupational and other variables. This will al-

low comparative analysis of the careers of persons and families, and investigation of the geographic organization of economic activity. For example, the database will allow a comparative national and possibly international investigation of maritime communities.

Household and family composition. Political theorists, sociologists and historians have been debating the relationship between industrialization and the family. Some studies argued that the harsh economic conditions of early industrial capitalism strengthened the interdependence of family members and led to a high frequency of complex households (Anderson 1971; Hareven 1978). In recent years, numerous national and regional studies of family composition in the late nineteenth century were based on population samples, but few incorporated community-level economic measures (Sogner 1990; Gunnlaugsson and Garðarsdóttir 1996; Ruggles 2000). Thus, there is presently little agreement about national similarities and differences in family and household composition in the late nineteenth century. In order to understand the transition between family forms, such investigations need data sets where households can be followed over time. This type of analysis will be particularly conducive if longitudinal databases from several countries can be combined and the context of changing family forms can be related to their setting through multi-level analysis. An example can clarify the need for longitudinally organized source material: In a family and aging oriented study the main research question is to find factors that motivate parents and their grown-up children to live in the same household. With access to a cross-sectional source we can study the differential characteristics of aged people who live with or without their children, for instance the extent to which this depends on gender and age. But differential characteristics of the relevant grown-up children cannot be analyzed, because the children who are not living with parents are not linked to them. As soon as we have links to persons in at least one earlier census, we can study differences between children who do or do not co-reside with their aged parents. Thus, only research based on longitudinal data will be able to contribute more than half-way answers to one of the longest-lasting debates in social history research, the question of the extended family structure (Jåstad 2009). This paragraph is but one example of why *family networks* became so topical for understanding social and demographic history among researchers internationally. In addition, kinship is the pivotal point in many fields of human interaction, from the family firm via the welfare of children and the elderly to chain migration.

Name studies. The culture of changing name traditions for men and women over two centuries can be related to different family types, occupations and geographic areas in longitudinal data. In a longitudinal database names can be studied more source-critically and dynamically because onomatologists get access to information about name forms for the same person at different points in time. (Cf NFR project 164186 which standardized all names in the censuses 1801-1900.) During the pre-project we extended the standardization of names from the censuses to the 1910 census and the church registers. Linguists see a potential for using the HPR in studies of dialects. The acceptance of the project leader's article "Religion in the Census" by *Social Science History* demonstrates that studying additional cultural aspects is feasible.

National and international migration. The late nineteenth and early twentieth century saw geographic population movements on an unprecedented scale. The massive emigration to America profoundly shaped both the receiving and contributing countries. Many of the emigrants remained only a few years before returning to their homelands, often bringing home money and always bringing new ideas and experiences (Gjerde 1992; Thorvaldsen 1998; Eide and Thorvaldsen 2011). The HPR will be a rich resource for the study of migration history, and will open a new window on the implications of national and international population flows.

6. User groups

- **Justification of the level of national interest and participation.**

A historical population register covering the whole country during the last two centuries will be used in many types of research, which explains the size and variety of partners in the consortium behind the project. The potential user groups are, however, even wider as expressed in the letters of support accompanying this application..

The needs and interests of national research teams and user groups.

The intensive migration inside and out of Norway especially after 1850 is a specific reason why we need a national register; otherwise migrants will be underrepresented in the database together with regions having unique characteristics. An example of a sub-database which is already national is the one containing normalized names. All names in the 1801, 1865, 1875 and 1900 censuses—nearly six million person entries—have been standardized with the support of a grant from the Research Council of Norway. The standardization was conducted with the assistance of Gulbrand Alhaug, Professor of Nordic Languages, and was enhanced during the current HPR pre-project. The initial list contained 71,396 different first names, 125,631 different last names and 87,116 different place names. This renders the population register usable in name studies where the onomatologists can study name variants in the longitudinal database even for the same person, and how names were inherited over the generations.

- **How the new research infrastructure may create new international research collaboration, its attractiveness for high quality international research groups, the added value of such collaboration, Nordic and European dimension.**

  The project partners are already cooperating internationally to build and employ population research infrastructures. The NHDC has made censuses internationally compatible through the NAPP project and link Norwegian censuses in cooperation with the Minnesota Population Center. This longitudinal material has already been used by US economists to assess the relative economic benefits to families from emigration (Abramitzky et al 2009, 2012). Through the Boreas project supported by the EU and NFR the UiT standardized ethnicity variables in cooperation with the Centre for Sami Studies at Umeå University and Canadian colleagues. Such links to international partners will be strengthened when implementing and documenting the population database for research. This is especially the case with the well-funded Demographic Database and Centre for Population Studies at Umeå University where they work with comparable data sets built on excellent Nordic source material. In the Minnesota Population Center record linkage of Norwegian sources is given high priority. Our work with international colleagues to develop a standard for the dissemination of longitudinal data has already been mentioned and is now expanded to many European countries with ESF sponsorship. We have to some extent succeeded in piggybacking onto the ongoing strong international research at the Demographic Database in Umeå and the Minnesota Population Centre. The HPR will render Norwegian source material significantly more attractive by making the population data longitudinal and extending the time series through the 20th century – thus removing our two biggest handicaps compared with the datasets available from these two institutions.

- **Users from industry and/or the public sector, nature of the collaboration and use of research results for innovation.**

  In the public sector the HPR will make the relevant source material easier to use for solving contested heritage cases and land right issues. Norwegian municipalities spend approximately 40 million NOK annually funding community history monographies. Among the various genres in the field, the so called "farm and family history" is by far the most financially demanding, as well as the most heterogeneous with respect to professional quality. The HPR can potentially reduce public spending on this activity at the same time as the services to the significant part of the population who are interested in genealogy and local history will be greatly improved. The HPR will magnify the empirical basis, improve the methods of research and fundamentally alter the premises for conveying the results of farm and family history to the public. Altogether, the HPR will make possible a sophistication of social and demographic analyses in community histories and facilitate comparative research, thus transcending the somewhat or sometimes parochial character of publicly commissioned and financed local history.

In medicine and public health pedigrees can be used to facilitate prevention and treatment of certain heritable diseases, though legal issues at present limit how far one can go in targeting specific individuals. A telling example of the potential gains is, however, the discovery of a genetic malfunction behind the loss of three siblings by doctors at Haukeland hospital which required much use of unprocessed source material and needed a budget of three million kroner – to help one single family (*Bergens Tidende* 25/1-07).

Our biggest user groups are among the general public: Genealogists and community historians. They already use the cross-sectional versions of the sources heavily, but will be able to extend and detail their ancestries and local histories by searching the HPR via the Internet. Through the HPR-wiki they will also complement data entry and record linkage to the HPR. For judicial reasons, their participation must be limited to the open part of the database, before 1920.

- **Nature and goals of relevant ongoing and future projects. List examples of relevant funded current national and international research projects.**
  - Biobank Norway, a large infrastructure grant from the Research Council of Norway
  - The Nord-Trøndelag health study (HUNT), health survey
  - The Tromsø Study, health survey
  - Norwegian Mother and Child Cohort Study
  - The autism study – Autism Birth Cohort (ABC)
  - ESOP, Center of Excellence, Dep. Economics, Univ. of Oslo
  - Extending the Population and Housing Census 1960 – 2001 file back to 1950
  - 1814: the ancestry of the Constitutional Assembly and the population
  - 1913: identifying new voters due to the law about general female suffrage
  - Ethnic differentials in changing family structures
  - Source critical assessment of international census data (UiT)
  - European Historical Population Samples Network (EHPS-Net)
  - North Atlantic Population Project
  - International name standardization and record linkage
  - Using supercomputers for record linkage (UiT)
- **Estimation of the total use, annual number of Norwegian and international scientific users, industrial/public users.**

Traditionally, historians have concentrated their population and social history research on the period with open sources up to 1900, while economists, demographers and other social scientists have worked mostly with material from recent decades. Relatively little has been done on the early 20th century in Norway, a period when the basis was laid for crucial changes in Norwegian society, including the welfare state and fighting diseases such as TBC. A longitudinal database which bridges the entire period from 1800 to now will inspire cooperation between these groups of researchers because many questions can only be answered with integrated source material that needs to be utilized with a common pool of research methods.

Most of this longitudinal research will be significantly enhanced by international comparative collaboration. This is particularly the case with respect to research on economic development and migration. Since the USA has relatively weak longitudinal source material, migration researchers there are easily drawn to nations where the life histories of potential emigrants can be studied in detail.

The storage and exchange structure for international cooperation on longitudinal source material is being designed right now, and it is important to participate with relevant Norwegian data. Otherwise, there is a risk that a standard is established which may not be well suited to the HPR. Hence, the HPR will be essential for historical and social science research on this period and open up new research fields due to the accessibility of the data. Also for medical research on genetic diseases where it is necessary to identify several generations, the HPR will be crucial.

Educational applications of the database. We anticipate that the major publicly available parts of the new database will make important contributions to teaching in history and the social sci-

ences, helping to bring the excitement of discovery into the classroom. The close focus on individuals' and groups' life spans made possible by the new database makes it a suitable vehicle for introducing both a micro and a quantitative dimension into secondary, undergraduate and graduate courses focusing on community history. Once the HPR is established, we plan to collaborate in the development of web-based instructional materials that capitalize on the fine detail available for local areas and small population subgroups. Especially for educational use, but also for research it will be important to interface the dissemination of longitudinal data with a GIS solution.

Uses in genealogical studies. Tracing ancestors is one of the most popular activities on the Internet. Norwegian genealogists form the biggest user group of the transcribed sources and contribute constantly towards transcribing additional source material. Many genealogists using our material live abroad, and there are more people of Norwegian descent in the USA than in Norway. The travel industry can use a database where the tracing of ancestors is easy to launch tourism campaigns towards overseas markets, invite people to learn genealogy on location and visit the old family farm or search for Norwegian relatives still alive. The NR investigates to what extent it is realistic to let genealogists contribute towards linking of records over the Internet without jeopardizing the quality of the result. The wiki-technology employed in HPRwiki is designed for mass cooperation, each change is identified to facilitate the possible removal of changes done by a specific user. We apply stronger security rules than are used in Wikipedia, based on the stricter regulations in lokalhistoriewiki.no. There are several levels of interaction; identify possible errors and new links, or alternatively mark as a possible link/duplicate between two observations in the database.

7. Partners

- **Description of the scientific consortium, the research groups, institutions, possible partners from industry or public institutions, international partners.**
  The HPR consortium comprises all major relevant institutions in the field of preparing nominative population data in Norway and several of the major user groups. The National Association for Local History with ca 100 000 members and the Association for Computers in Genealogy (DIS-Norge) representing 10 000 members are main partners in the public sphere. Through the NAPP project and the EHPS-Net we interact with virtually all potential international partners.
- **Documentation of the scientific and technological competence of all partners, their roles, expertise, responsibilities and commitments (as partner, host, contributor), priorities with respect to their individual strategies.**
- **Describe specifically the institutions involved and expertise needed in the relevant phases of design, construction and/or operation of the new or upgraded research infrastructure.**

  The National and Regional Archives of Norway (Arkivverket) is responsible for preserving and giving access to records from the governmental administration, both paper-based and digitally born records. Lots of paper-based records have been digitized (transcribed or scanned) and made available to the public through the website digitalarkivet.no (The Digital Archive). Main website: arkivverket.no. *Relevant contributions* to the current project is transcription work (7 man-years per year) and (long term) storage, maintenance and preparation for HBR (the latter task in cooperation with the NHDC) of transcribed versions of all the censuses 1801-1910, church books 1800-1930 and emigrant records 1865-1930. *Special responsibility* in the current project: Technical operation and maintenance of the open and closed HPR. Is building the HPR for the period 1801-1815. Scanning of relevant source material for transcription.

  The Norwegian Historical Data Centre (NHDC) at the University of Tromsø transcribes, encodes and disseminates nominative records according to international standards necessary to render the material fit for record linkage and statistical use. Encoded versions of the national 1865, 1875, 1900 and 1910 censuses are available both locally and via the NAPP project at the University of Minnesota. *Special responsibility:* Overall coordination, normalizing and

extending the cross-sectional and ministerial sources into longitudinal databases in compatible formats as key building blocks in the HPR. Website: www.rhd.uit.no

Statistics Norway (SN) was responsible for collecting the information in the census manuscripts, holds computerized versions of the censuses 1960 through 2001, and established the Central Population Register in 1964. SN has extensive experience in linking and using data from different administrative registers, and distribution of microdata from censuses, sample surveys and registers to researchers. SN will also be an important user of the HPR for statistics and research. *Special responsibility:* Competence in CPR and censuses, linking the CPR backwards in time, quality control, distribution of data, legal issues. Website: www.ssb.no

The Demography unit at the University of **Oslo** has supervised a number of doctoral students writing demography theses, both in history, economics, sociology and geography. The unit will especially work to develop and carry out demographic analyses, applying more advanced event-history techniques, such as multi-process and multilevel models to longitudinal demographic data. *Special responsibility: Assessment of representativity.*

NR, Norwegian Computing Center combines outstanding IT expertise with an interest in genealogical research. *Special responsibility:* Public web interfaces, Norwegian HPRwiki, transcription methods and software. Web site: nr.no/

The Norwegian Institute of Local History is the focal point for local historians' contributions to building a population register. *Special responsibility:* Coordinate input and data requests from local historians, contribute to HPRwiki and co-ordination with lokalhistoriewiki.no. Website: lokalhistorie.no

Snøhetta forlag / Volda University College has developed cost-efficient software **(BSS-database, Busetnadssoge)** for the production of local history volumes used in several community studies. *Special responsibility:* Extend, reformat and document their longitudinal databases. Support and complement the HPR-wiki e.g. with extended articles on biographies and places through inter-wiki cooperation. Website: tilsett.hivolda.no/ak/BSS/Busetnadssoge.html

The Norwegian Institute of Public Health. The Norwegian Institute of Public Health is a national centre of excellence in the areas of epidemiology, mental health, control of infectious diseases, environmental medicine, forensic toxicology and drug abuse. *Special responsibility:* implementation and methodological analysis of optimal integration of information from the HPR with biobanks and health registers. Encoding of causes of death, establishing a complementary Causes of Death Registry 1925-1950 Web site: www.fhi.no/eway/?pid=238

The Migration Research Group **at the University of Stavanger**. Its members hold phds in migration research and teach a masters course in migration studies. *Special responsibility:* Handle the out-migration from the HPR on the individual level, viz. the emigrations protocols.

University of Bergen, The Institute of Archeology, History, Cultural Studies and Religion holds special expertise in the areas of urban GIS and the study of household structure. *Special responsibility*: to relate the nominative data to the grid of urban domiciles and detail the intra-household family relations and pointers as they change over time.

8. Project management

- **Project management, administration of roles, competence and contribution of the partners in the relevant phases of design, construction and/or operation.**

  The overall project administration will remain at the University of Tromsø including budget surveillance. The project group will continue to have at least one member from each consortium partner, and is strengthened with specific expertise as need arises. The board will still be headed by the National Archivist and will discuss how to supplement itself at the start of the main project period. Practical implementation of measures to safeguard ethical and legal non-disclosure standards will be handled by the National Archives with Statistics Norway in an advisory role. Main surveillance of IT standards will be with the Norwegian Computing Center.

The project's steering group: National Archivist Ivar Fonnes; director, Elisabeth Nørgaard (?), Statistics Norway; lecturer Kåre Andersen, University of Oslo; director Lars Holden, Norwegian Computing Center; director Per Magnus, Norwegian Institute for Public Health; director Knut Sprauten, The Norwegian Institute of Local History; professor Gunnar Thorvaldsen, University of Tromsø

The project's working group: Senior advisor Lars Nygaard, the National Archives; senior researcher Helge Brunborg, Statistics Norway; director Lars Holden, Norwegian Computing Center; researcher Kåre Bævre, Norwegian Institute for Public Health; consultant Hans Hosar, Norwegian Institute of Local History; professor Gunnar Thorvaldsen, University of Tromsø. The groups will be supplemented with new partners.

- **Justification of suggested localisation, host institution(s), alternatives that have been considered, opportunities and risks.**
  The project group and the board has discussed whether the responsibility for storing the modern (post 1920) part of the database should lie with Statistics Norway or the National Archives, and has decided that admission to data will be given to researchers by the rules and procedures in the two institutions. Since the time split between the two registers will move forward in time and in order to integrate the two parts in a flexible and safe manner for research over long periods, it has been decided that both the National Archives and Statistics Norway should store the modern part of the database. For safety and flexible use the open part will be stored both at the National Archives and the University of Tromsø.
- **How the research infrastructure will be managed subsequent to the project period, responsibilities for construction, operation and upgrade, how the new research infrastructure fits into the host institutions long term planning and research strategy.**
  The main long term responsibility for the management of the HPR will lie with the National Archives as part of their general responsibility for storing and making available official public databases and registers. The Norwegian Historical Data Centre will as a permanent part of the University of Tromsø cater for the compatibility of the HPR with similar repositories in other countries and provide help both for national and international groups of researchers utilizing the population register. The Strategy Document for the UiT 2009-13 explicitly promotes digital databases among its infrastructure components also for the humanities and social sciences, and with ethnicity as an important variable. The HPR represents an instrument for regionally oriented teaching and the dissemination of research results. The NHDC is already one of the most used web sites in Northern Norway and its use by genealogists contributes significantly to the cultural exchange with the general public. Linking the HPR to national and regional health registers may reveal to what extent health differentials correlate with ancestry.

9. Plan for access and use, data and knowledge management

- **How access to the new research infrastructure will be effectively managed, principles of project selection, capacity issues.**
  The IDS standard together with commonly available software (open source) will be the major interface for academic researchers. Some specialized data extraction for specific purposes will still be necessary. During the first five years this will be handled as a free service for users who can document that their projects will enhance the HPR or result in reusable software components. Afterwards, specialized data extracts will be organized as service to be paid for by the projects that are not able to make their own data extracts.
- **Knowledge management, transfer and utilization in innovative environments, how data and knowledge are generated, managed and made available for research and innovation, the transfer of knowledge and results to academic, industrial and public partners.**
  Bona fide researchers will have read access to parts of or the whole HPR as a relational database upon a vetting of their competence and data need. The public will utilize the pre 1920 HPR via the HPRwiki, the Digital Archive or the NHDC web pages (by 2022 the pre 1930 HPR). Standardized individual level extracts will be made available for international researchers via the

NAPP project, for instance cross-sectional data interpolated between census years. (A new virtual 1810 and/or 1815 census is planned as part of the NA's 1814-project.) Longitudinal extracts will follow the IDS standard.

- **Policies for publications, patents and the dissemination of research results.**
  Users of the modern part of the HPR must follow rules set up by Statistics Norway for distribution of micro data to researchers. The National Archives can demand that all or selected copies of the files transferred to or generated inside a research project are transferred to them for storage and deleted at local equipment after a period of time.
- **Use of electronic services and portals, databases, sample and publication archives, etc.**
  Since the HPR will be a more rational tool for utilizing the population records than what is presently available, much of the new data traffic it will generate will be compensated by a reduction in the heavy transmission of digital copies of original source material now going on. Thus, it is feasible that the traffic can be handled by existing or already planned extensions of administrative and research data networks. Some limited investment in database servers will be necessary, together with adaptation of database, web and record linkage software.

  The HPR's limited access research database (1920-1964) will be stored on one or two servers which will be connected only to a separate, protected, internal data network at the National Archives (NA) and Statistics Norway, and maintained only by authorized personnel. From about 2012 the NA will have a "trustworthy digital repository" (according to the TRAC criteria) for long-term storage of digitally born archival material, digitized material and also databases. Authorized researchers will in both situations be given access to copies or extracts from the database through a dedicated data network ("outer zone") within the archival buildings and other specially approved institutions, alternatively de-identified files through a download service for researchers based on encrypted file transfer (SSH/SCP). Extracts will be tailored and produced by authorized and competent personnel by specification from the researcher.

10. Time-schedule and deliverables

- **Detailed time-schedule for the main activities and progress of the project.**
- **Milestones and project deliverables in relation to the relevant phases of the design, construction, installation, test periods, operation and/or possible upgrades and decommissioning.**
- **Emphasise critical dates with respect to funding and commitments, the technical and scientific development.**

❖ Building an open wiki database (mainly data until 1919), called "HPR-wiki". First open version including 1910 census after 12 months by NR
  ➢ Fully operable after 36 months

❖ Managing the open database with many volunteer contributors
  ➢ Central management operable after 12 months at UiT/NA in cooperation with DIS-Norway
  ➢ Establish local groups in 10 provinces (fylker) after 24 months
  ➢ Establish cooperation with local groups in all provinces after 36 months
  ➢ Launch the HPR for the period 1801-1815 as part of the national celebrations in 2014

❖ Building a closed database (data from 1919 and after) at Statistics Norway (SN) / NA
  ➢ First version including parts of the 1920 to 1950 censuses (and provisionally the whole 1910 census) after 18 months
  ➢ Second version including significant parts of the ministerial records after 30 months
  ➢ Complete will all censuses and vital records 1920 to 1960 by 2021

❖ Scanning of primary sources at the National Archives
  ➢ Scanned censuses 1920 to 1950 within 60 months

❖ Scanned parish registers within 108 months

❖ Making software that improves the efficiency of transcribing from primary sources by combining graphical and interactive techniques at NR
  ➢ First open version after 12 months

- Fully operable after 36 months including new functions every 6 months from start.

❖ Transcribing from open and closed sources by volunteers and professionals at DA and NHDC
- First version operable after 12 months of project and fully operable at end of project.
- Transcribed contiguous municipalities in all regions with 1/3 of all records after 24 month
- Transcribed 2/3rds of all censuses and vital records at end of NFR-support to project
- Transcribed all censuses and vital records by 2021.
- Importing one BSS municipality database every half year, complete by 2015.

❖ Coding and linking persons and places in both open and closed databases aiming at
- Transcribed censuses: 90% of persons linked to places when registered in both open and closed database by NHDC and NA.
- Open database: 2/3 of population 1800-1920 with at least two sources and 90% in regions covering 1/3 of total population at end of project by UiT and MPC
- Open database: 85% of population 1800-1920 with at least two sources 5 years after end of NFR-financed project by UiT and MPC
- Closed database: link 50% of transcribed persons at end of NFR-project by NA and SN. Link 90% of transcribed persons 5 years after end of NFR-project by NA and SN

❖ Preparing for research based on database
- Document retrieval techniques with wiki and relational methods by NR within 6 months.
- Develop electronical search possibilities and cross-sectional excerpts from the database that are suited for selected research problems by NHDC and MPC
- Update Intermediate Data Structure (IDS) prototype to current version within 6 months by NHDC and EHPS-network (for quality controls)
- Extend IDS prototype to complete longitudinal coverage within 12 months by NHDC
- Export all complete BSS and other longitudinal datasets to IDS format within 24 months by NHDC and EHPS-network (for quality controls)
- Full IDS versions of the HPR datasets as a continuous service after 60 months by NHDC
- Establishing harmonized coding schemes for causes of death 1900-1950, validation of quality of the historical information on causes of death after 24 months. Extend the Causes of Death Register for 1925-1950 at pace with inclusion of recorded deaths in HPR by FHI.
- A sub-project for the period since 1950 has been singled out and budgeted in the attached spreadsheet upon agreement with NFR administrators. It is described in the attached 4-page update, and will use two years for data entry, two for record linkage and one for quality assessment.

11. Budget and funding plan

- **The total financial plan must cover all partners for the realisation of the research infrastructure.**
- **For applications concerning amounts beyond/above 75 MNOK, the applicant should consider specifying a down-scaled version of the project and indicate a minimum financial requirement for this.**

  A reduced contribution from the NFR makes it possible to carry on work with the HPR, but the progress will be much slower. The project will then depend more on voluntary contributions from genealogists, so in particular the closed database from 1920-1964 would suffer from a reduced budget. Since central data for the post-1950 sub-project are already machine-readable, this can be singled for rational separate processing, however.
- **Contributions from national and international partners, the applied contribution from the Research Council, own funding from the host institution, other partners, industry, public institutions and user groups.**

  The Minnesota Population Center has contributed constructed variables and record linkage of Norwegian censuses and transfers expertise on the use of supercomputers to achieve this. They will continue to make available enhanced versions of extracts from the HPR. The Demographic Database contributes expertise on the building of longitudinal databases. The National Archives

and the NHDC contributes 15 work years of transcribed data each year in addition to the 12 million records which have already been transcribed. The former also contributes scanned images from original sources, while sources from after the 1920s must be scanned at the expense of the HPR project. Statistics Norway contributes 12 work months yearly of legal and statistical advice. The NHDC, the National Archives and the NR will contribute 2, 1 and ¼ work year respectively of IT expertise funded outside the NFR each year during the project period.

- **The plan must include all necessary costs related to investment and operation through the project period and the subsequent period , such as equipment, services, installation and other costs, dedicated personnel and costs to secure long term operation of the research infrastructure, the provision of laboratory space and services, etc. Please use and attach the economy excel-form that can be downloaded from our website.**
  Budget with overall use of NFR's infrastructure contribution in 1000 NOK:

Scanner						1000
Salaries etc for development						
NFR Funding of partners (modest scale) 0,102 mill x 5 years x 4 partners + 0,625						2665
NFR Funding of major partners	NR	NHDC	FHI	NA	SN	
	4995	6565	2500	6500	1775	21335
					Sum	25000

Cf NFR's and our attached spreadsheets for further details.

- **Special emphasis must be put on describing how various costs will be covered to secure operation of the infrastructure after the funding from the Research Council is terminated. Please use and attach the economy excel-form that can be downloaded from our website.**
  The permanent storage of the HPR will be in the National Archives as part of its responsibility for providing access to official historical databases. The extension of the Central Population Register is a continuous activity at the Directorate of Taxes and Statistics Norway. The international documentation and access to the HPR is a natural extension of ongoing activities in the NHDC.
- **List relevant national or international funding of research infrastructure, design studies, networks etc. if existing.**
  UiT's participation in the NAPP project is financed by the US National Science Foundation. The ESF-supported EHPS-Net is funded by national research agencies. The NHDC is financed as a permanent part of the University of Tromsø, and the Digital Archive website as an activity within the National and Regional Archives. Project partners are applying for supercomputer funding for strengthening work on source transcription and record linkage respectively.

12. Environmental and ethical perspectives

- **Potential consequences on the natural environmental of the infrastructure and research activities.**

Making this source material available in rational formats online saves travel to the archival deposits.

- **Potential ethical issues related to the research infrastructure, the research projects, or the results from the research.**

All data management and research carried out within the project will conform to the Helsinki Declaration and to national guidelines including the Law of Statistics, the Personal Data Act and Regulations, the Law of Archives and the Law of Public Management. Access must be regulated by period and type of basic source material. For the period up to 1920 the data can be used publicly with few restrictions. For the period from 1920 approval must be obtained from the Data Inspector and relevant ethics committee before studies can begin. In projects where only deceased persons need to be included, not requiring sensitive information, data from before 1920 can be used publicly.