



DISC

Magnus Stenbeck



- What is DISC?
- DISC Goals and activities 2006-2008
- Research data infrastructures in the long term
- MONA (registry data access system)
- SND (dataservice for research)
- Funding for large databases
- Registries, statistical authorities, and legal problems
- Database technology and organisation
- DISC future

Contents



- Part of the national effort to create research infrastructure
- Expert committee for databases under the Committee for Research Infrastructure, Swedish Research Council
- Active researchers: humanities, social science, medicine, climate and environment, data base technology
- Director and directorate placed at MEB
- 2008 budget: 48 MSEK

What is DISC?



- To make Swedish research world leading through the systematic use of common large databases and registries in research
- Advocate a collaborative attitude among researchers
- Coordinate the national development of research database infrastructure
- Support to all scientific fields where data analysis and database management are important

DISC Goals 1



- Promote rational long term use of research funds (cost efficiency, synergies)
- Promote international collaboration and exchange of data
- Promote development of efficient tools for storage, searching, documentation and analysis of data
- Promote data quality and adapt national non-research registry data to research needs

DISC Goals 2



- Promote development of secure data handling and increased awareness of integrity legislation and ethics among researchers
- Promote the development of rational legislation and praxis to increase access and use of research data
- Promote the adaption of commercial, copyright and ownership rights to the needs of research

DISC Goals 3



- Large databases
- MONA
- SND
- Language technology
- Digitalization in the Humanities
- Legal problems
- Database technology

2006 - 2008



- Infrastructures must be preserved longer than the life of research projects
- Preservation is a technical, legal and budgetary problem that needs attention
- Needs to be influenced by:
 - Strategic long term considerations
 - Current top research interests
 - Long term agreement in the research community

Research data infrastructures in the long run



- Life gene (genetic and personal information on 500 000 Swedes)
- A Swedish national corpus (a representative sample of the Swedish language)
- The Olive tree (national cohort study of a birth cohort across 100 years)
- Problematic in relation to current methods of financing and preserving the

infrastructure of large scale efforts in terms of funding



- Typically moderately sized data holdings – perhaps 10 million rows and a few hundred columns
- A fairly complex data structure
- Very dependent on extensive documentation (example: ICD coding of diseases)
- Very high security demands (e.g. person identified data)
- Moderately dynamic data

Research data in social science and epidemiology



- Constitutional rights of access to official information (offentlighet)
- Constitutional rights of privacy protection (sekretess)
- Law on ethics trials for research
- Copyright and other ownership laws
- Archival legislation (long term storage and retrieval)
- The Biobank law
- The legal system sometimes creates problems for research access to data
- Researchers awareness of legislation and ethics can be improved

Legislation



- Complex data analysis
- A combination of central authority data and researchers generated data
- Typical data:
 - Administrative registries (e.g. income, health care events)
 - Surveys (attitudes, self reported health and social conditions)
 - Biobanks (large datasets, combination of physical storage of blood, tissue with analysis data)

Research data in social science and medicine



- Statistics Sweden (social and economic data on persons, households, and organizations)
- National Board of Health and Welfare (health data)
- National Council for Crime Prevention (crime data)
- The Swedish Social Insurance Agency
- etc.

Registry data in social science and medicine



- Several hundred data sources which can be combined using person IDs
- Data reside with different authorities which cannot easily share data among each other or with the research community
- Extensive legislation to protect individual rights and proprietary interests
- The registries have been called a national treasure for research
- An efficient system for secure access is needed

National Registries



- Statistics Sweden is developing an online system
- Centralized storage
- Common documentation system
- Offers other authorities participation
- The system is sufficient if user data are asparse and simple
- Insufficient for large databases
- Insufficient for dynamic user data
- Insufficient for sensitive user data

Microdata ONline Access MONA



- Formerly Swedish Social Science Dataservice SSD
- Since 1 Jan, 2008 a national dataservice at Gothenburg university
- The purpose is to provide know-how, technical, legal, and other resources for empirical research
- Should serve the humanities, social science, and medicine
- Data should be made available to the research community through the SND portal

Swedish National Dataservice SND



- Data supported by the Swedish Research Council should be made available through SND
- Data availability is contingent on the legislative framework
- Data can be stored in Gothenburg, but this is not necessary
- Data should be documented according to a common format (DDI)
- SND should have a help desk for data management and perhaps other methodological advice

SND 2



- Constitutional rights of access to official information (offentlighet)
- Constitutional rights of privacy protection (sekretess)
- Law on ethics trials for research
- Copyright and other ownership laws
- Archival legislation (long term storage and retrieval)
- The Biobank law
- The legal system sometimes creates problems for research access to data
- Researchers awareness of legislation and ethics can be improved

Legislation



- Creating a permanent infrastructure of research data is not legal
- Ethics trial and Secrecy trial should be coordinated, and is sometimes contradictory
- Central statistical authorities instructions from the government do not include supporting research
- Knowledge among researchers is limited
- Data handling is in many cases

questionable
Registries, authorities, and legal problems



- We need an official investigation into these problems:
 - The structure of the legal system: principles for copyright, person integrity protection, and adaptation of the biobank law to research needs
 - The role and financing of central statistical authorities in relation to research
 - The development of new technology in database management and its legal implications
 - Can universities create and keep general purpose research registries?

Legal problems 2



- Many problems cannot be solved in a centralized online system
- The old system does not promote efficient collaboration and accessibility
- Federation: an association of independent participants who have agreed to collaborate on a long term basis
- Federated data structures can provide an easier technical solution to sharing data?
- Federated data structures can provide an easier legal solution?

Database technology and organisation



- THE MONA system is particularly problematic for epidemiological research with large dynamic external databases
- The current trial at MEB encompasses gynecological testing data from a number of screening centers, data from health registries, and social/demographic data from Statistics Sweden
- The purpose is to search data at the host and combine and analyse data in a federation server
- The data cannot be stored in large datasets
- Data must be made restorable for archival and future research purposes

Database technology 2



- Climate and environment data
 - collaboration with SNIC and its partners
- Biobank data
 - national coordination and
 - combination with health and social registries

**Future database areas for
DISC**



