DESCRIPTION OF THE INFRASTRUCTURE AND ITS ACTIVITIES

DEFINITIONS

Data Access Unit (DAU): a function at a university that assists researchers to provide access to data at the end of a project or at publication. Their exact duties and how they are organised will depend on their institutional context. May require resources from library, archive, grants and innovations office, and researchers. DAUs are not part of, but work in close contact with, SND 2.0.

Domain: area defined by the particular skills, methods, and/or knowledge required of researchers and data professionals working with data from this area. These domain-specific requirements can be related to for instance scientific, analytical, ethical/legal, or process-related factors.

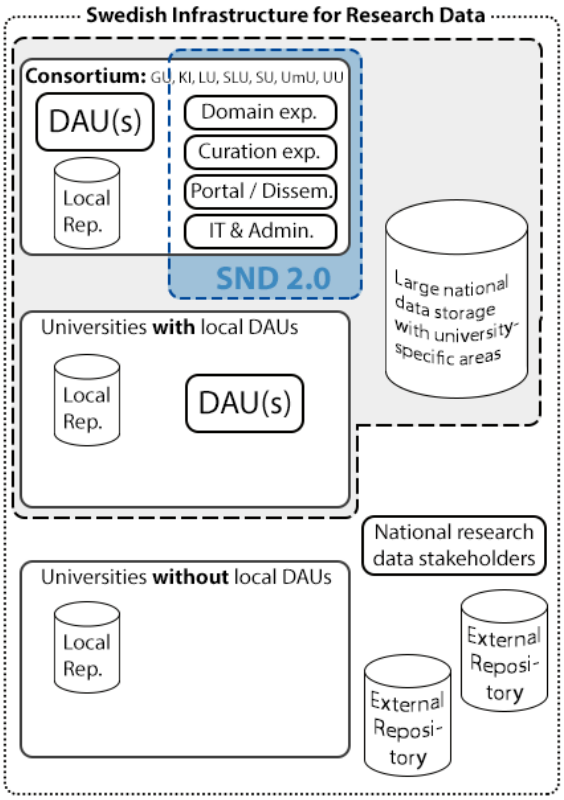
Metadata: structured information about (research) data that describes their various properties with the main purpose of making the data findable and reusable.

Repository: "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution" (Lynch, 2003: 2). In this sense, a repository does not include the actual storage infrastructure itself, nor is it an archive as defined by the Swedish Archive Act (1990:782).

ABBREVIATIONS

ARIADNE	Advanced Research Infrastructure for Archaeological Dataset Networking in Europe
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
DAU	Data Access Unit
DCU	Data Curation Unit
DDB	Demographic Data Base
DDI	Data Documentation Initiative
DI	Datainspektionen/The Swedish Data Protection Authority
DiVA	Digitala Vetenskapliga Arkivet
DOI	Digital Object Identifier
EHPS-Net	European Historical Population Samples Network
EISCAT	European Incoherent SCATter Scientific Association
EMA	Environmental Monitoring and Assessment
ERA	European Research Area
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum on Research Infrastructures
ESS	European Social Survey
FAIR	Findable, Accessible, Interoperable, Reusable
GU	University of Gothenburg
HRM	Human Resource Management
ICOS	Integrated Carbon Observation System
ICPSR	Inter-university Consortium of Political and Social Research
IDF	International DOI Foundation
KI	Karolinska Institutet
LU	Lund University
LULC	Land Use/Land Cover
MONICA	Multinational Monitoring of Trends and Determinants in Cardiovascular Disease
NeIC	Nordic e-Infrastructure Collaboration
NordiCom	Nordic Information Centre for Media and Communication Research
NSB	Northern Sweden Biobank
PID	Persistent Identifier
QA	Quality Assurance
RDA	Research Data Alliance
RDM	Research Data Management
RECODE	Policy Recommendations for Open Access to Research Data in Europe
SCB	Statistiska Centralbyrån/Statistics Sweden
SciLifeLab	Science for Life Laboratory
SEAD	The Strategic Environmental Archaeology Database
SHARE	Survey of Health, Ageing and Retirement in Europe
SHFA	Svenskt hällristningsforskningsarkiv/Swedish Rock Art Research Archives
SIME	Swedish Institute for the Marine Environment
SLU	The Swedish University of Agricultural Sciences
SND	Svensk Nationell Datatjänst/Swedish National Data Service
SNIC	Swedish National Infrastructure for Computing
S-NICE	Swedish National Infrastructure for Climate and Earth System Research Data
SOM	Samhälle Opinion Medier/Society Opinion Media
SP	Service Provider
SSHRI	Social Science and Humanities Research Infrastructures
SU	Stockholm University
SUHF	Sveriges universitets- och högskoleförbund/The Association of Swedish Higher Education
SUNET	Swedish University computer Network
Tilda	Tillgängliggörande och arkivering av forskningsdata vid Sveriges lantbruksuniversitet/Disseminating and archiving research data from the Swedish University of Agricultural Sciences
UmU	Umeå University
UU	Uppsala University
VIP	Västerbotten Intervention Program
VR	Vetenskapsrådet/The Swedish Research Council

Figure 1: SND 2.0's Location in the Research Data Infrastructure



SND 2.0 will be a key actor in a future Swedish infrastructure for research data. The SND 2.0 consortium interacts with the university DAUs (including those of the consortium) but can also rent out DAU capacity to universities that have not (yet) established a DAU. (The international infrastructure for data, while important to Swedish stakeholders, has been left out of the illustration.)

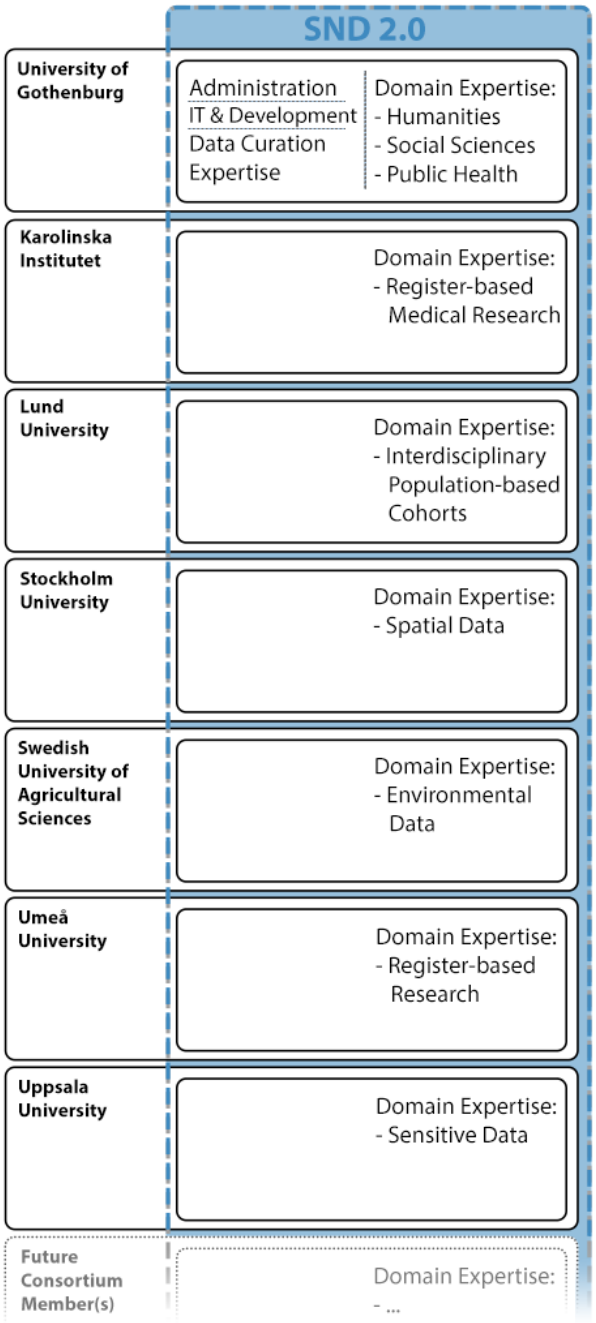
This application, the future challenges it identifies, and the infrastructure and activities developed to overcome these have emerged from numerous discussions with and a close collaboration between Swedish universities within and outside of the consortium. Through presentations, discussion forums, pilot projects, workshops, and an open writing process, the SND 2.0 of tomorrow has developed in order to fulfil a central role in how access to well-documented, high-quality research data can be provided.

SND 2.0 will be committed to providing FAIR research data to the scientific community but will not be able to do this alone. Future requirements for and interest in providing access to research data will necessitate the development of local support functions at Swedish universities. These functions, referred to in this application as Data Access Units (DAUs), will offer training and advice to their respective university's researchers in data management and documentation. The DAU functions will be organised differently depending on their institutional context, into a single unit, into a number of more specialised units, or as a distributed function. Experience and expertise from libraries, archives, grants and innovations offices, and active research may all be valuable in building a DAU. Together, the university DAUs and the SND 2.0 consortium will constitute an essential data access service to the Swedish research community.

1 Organisation and Management of the Infrastructure

This section addresses the construction and organisation of the SND 2.0 consortium; SND 2.0's strategic, scientific, and operational management; current and potential users, SND 2.0's role visavi them and how they interrelate with SND 2.0; and SND 2.0's memberships in and collaborations with international research infrastructures.

Figure 2: Distribution of SND Over the Consortium



As indicated in the boxes above, SND will have domain expertise distributed over the consortium universities.

1.1 The SND 2.0 Consortium

The primary consortium partners (University of Gothenburg, Karolinska Institutet, Lund University, Stockholm University, Swedish University of Agricultural Sciences, Umeå University, and Uppsala University) have agreed to provide local offices that will develop and maintain expertise available to the infrastructure as a whole. These domain specialists will widen the scope of SND 2.0's expertise and address a key overarching issue in mobilising open access to research data identified by the RECODE project (2014:3): "a lack of attention to the specificity of research practice, processes and data collection." Domain specialists will initially be supplied for the humanities, social sciences, public health & epidemiology, register-based medical research, environmental data, sensitive data, spatial data in the humanities and social sciences, register-based research, and interdisciplinary population-based cohorts linked with registers and biobanks. **SND 2.0 will thus be distributed over the consortium** (see figure 2), with domain specialists having close access to cutting-edge knowledge within their domains, through conferences, extensive networks, and their own research activities. They will be well versed in data management, documentation and metadata, and repository issues and thus well situated to provide advice on these matters to researchers within their domain. Their activities will be divided over all modules but with a concentration on modules 3 and 4 (modules described below). The number of domain specialists will increase over time, widening the scope of SND 2.0's services to more research disciplines. Current discussions concern the inclusion of engineering and artistic research data. The distributed organisation will afford close ties to the respective domains and to scientific communities of practice, thus establishing broad expertise in domain-related forms of research and data types.

For the sake of optimisation and efficiency, administration, IT and technical development, and expertise in data curation will be concentrated at the SND 2.0 office at the University of Gothenburg. The Gothenburg office will also manage SND 2.0's role as service provider for CESSDA.

1.1.1 Previous Experience and Focus of Consortium Partners

The consortium partners have demonstrable experience in research data management and dissemination across diverse research domains, as well as extensive networks both within these domains and to the global data management community.

Today's SND at the University of Gothenburg has provided Findable, Accessible, Interoperable, and Reusable research data (FAIR; see Wilkinson et al 2016) to social sciences researchers for more than three decades; humanities and health sciences domains were added to SND operations in 2008. As a Trusted Digital Repository, SND has reliable workflows, systems, and solutions in place from data ingestion to dissemination.

> Try out SND's online form to describe research data: snd.gu.se/en/beskriv-och-lamna-in-data/form. (When you upload, please use the Study Title "Test".)

SND has an extensive metadata register and an advanced search portal, including options to search for specific survey questions in a Question Bank. SND's participation in external projects includes the development of a web portal for the ARIADNE project. In 2010, SND became national allocation agent for DOIs and currently provides them to several Swedish research institutions, as well as to all data deposited at SND.

As part of a national network of stakeholders, SND regularly organises national conferences, workshops, seminars, and training courses. SND is an active part of the global data infrastructure, being a national service provider for CESSDA (see submodule 4a) and a CLARIN centre, and managing Sweden's membership in ICPSR. Through its membership in the DDI Alliance, SND participates in the development of the metadata standard DDI. SND also collaborates within the Research Data Alliance (RDA), whose global working and interest groups develop infrastructure to promote data sharing and data-driven research; and is a member of DataCite, which provides PIDs to make data citable. Moreover, exchanges with other EU-financed infrastructure projects have provided invaluable on-the-job-training, international contacts, and data infrastructure expertise.

> Run a few searches in SND's search portal for research data: snd.gu.se/en/catalogue
> For example, search for "working conditions" and see what results you get.

Karolinska Institutet has extensive experience in building, maintaining, and regularly updating large population-based cohorts for medical research, with detailed phenotypic data, biological samples, as well as exposure and covariate information from national health registers, questionnaires, examinations and interviews. Examples are the Swedish Twin register, the largest twin registry in the world, the Stockholm Public Health cohort, the Swedish Mammography cohort, and the cohort of Swedish Men. These are unique infrastructures of considerable national and international value for epidemiological, genomic, epigenomic, proteomic, metabolomic research on a range of diseases and health problems.

Lund University has for decades built up extensive expertise in establishing, maintaining and utilising population-based cohorts in the social and medical sciences linked with data from registers and biobanks. A recent initiative at Lund University to strengthen the infrastructure for interdisciplinary register-based research builds on these experiences.

Stockholm University (SU), especially within the geographical departments, has been working with digital spatial data since the 1970s. Today, GIS and spatial data are used in several academic disciplines across the faculties. Remote sensing has been employed extensively to create Land Use/Land Cover (LULC) datasets. There is also a focus on reconstructing past environments and land use to do time series analyses. New spatial and statistical methods and approaches have been developed and implemented to study geomorphology, marine geophysics, LULC, vegetation, climate, hydrology but also social and historical phenomena. At SU, spatial data are also used to investigate for instance languages, place-names, ancient monuments, abandoned settlement, and historical statistics. In the social sciences a number of methodically oriented research projects, bridging demography and human geography, have focused on population distribution and composition. By using longitudinal georeferenced data, researchers have studied people's behavior and spatial movement, uncovering new spatial configurations.

The Swedish University of the Agricultural Sciences (SLU) has knowledge and expertise in the environmental data domain, as well as in curating, archiving, and publishing data. SLU has a

government mandate to conduct environmental monitoring and assessment, and has developed extensive databases and science-based decision support tools. SLU has a Data Management Guidance and Development Unit, which coordinates quality assurance strategies to ensure the long-term availability of SLU's environmental data. In addition, a new university-wide Data Curation Unit (DCU; corresponding to a DAU) will support researchers within digital archiving and publication of research data. The DCU will operate as part of the university library, combining competence from the library, the Unit of Documentation and Legal Affairs, and the Data Management Guidance to support researchers. In 2017, SLU launched the *Tilda* system for archiving and publishing research and environmental data. Tilda facilitates input and linking of metadata, datasets, and publications, and will interoperate with SND 2.0 systems.

Umeå University has leading national and international expertise in register-based research, database modelling, and data curation and retrieval. Unique conditions for population-based biomedical research exists through a deliberate and long-term development of high-quality biobanks and longitudinal research registers. The Northern Sweden Biobank (NSB) is the largest population-based biobank in Sweden, characterised by very long follow-ups and high test quality. Key registries and databases, such as the Demographic Data Base (DDB), the MONICA study, and the Västerbotten Intervention Program (VIP), form unique research infrastructures with leading global positions in their respective fields. Umeå University is part of two ERICS, as Scientific Partner Institution for the Survey of Health, Ageing and Retirement in Europe (SHARE) and as national coordinator for the European Social Survey (ESS). Researchers at the university are co-founders of the European Historical Population Samples Network (EHPS-Net) with the mission to model and harmonise longitudinal population databases across Europe, and foster the next generation of register-based researchers. Umeå University spearheads the development of database federation techniques. Efforts include highly relevant theory and methodology for the development of a federated database infrastructure, with research on privacy issues and security. Around these resources strong research is conducted together with the establishment of advanced skills and competences.

Uppsala University has extensive and broad expertise in the area of life science and is a main actor in the major national initiative SciLifeLab. Within the university, there is significant experience on research based on collecting and managing life science data including the specific requirements attached to this. The university also hosts the national infrastructure for storage of sensitive data, SNIC Sens, and the national infrastructure for bioinformatics, NBIS.

1.2 Strategic, Scientific and Operative Management (Module 5)

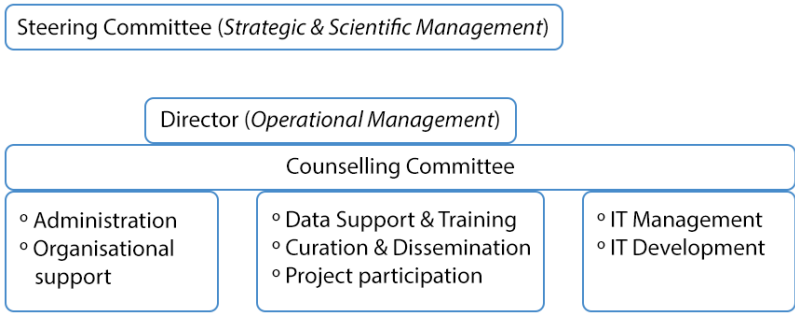
- ***Strategic and scientific management:***

SND 2.0's strategic and scientific management will be the purview of the Steering Committee. Its membership will be constituted by representatives of the consortium partners and the director of SND 2.0. The committee will be chaired by one of the consortium partner representatives. A Scientific Community Advisory Board to advise on the data requirements of the researcher community will be established. The possibility to include stakeholder representation in the steering committee will ensure that functions, tools, and standards developed within SND 2.0 are apposite to the Swedish research community, universities, and funding agencies. To further ensure that SND 2.0 will move in the right direction, input on the strategic development of operations will be solicited from DAU universities even though they are not formally a part of SND 2.0's management.

- ***Operative management:***

The operative management of SND 2.0 is arranged into a line and staff organisation, organised into three lines under the executive management of a director. The consortium agreement will establish the procedure for how the director is appointed. SND 2.0 is divided into three lines, each with a staff supervisor: Administration and Organisational Support; Repository, Data Support and Training, and Projects; IT and Development. (See figure 3.) The director appoints the staff supervisors, who will supervise the daily work in the operation lines and provide specialist advice to the director as members of the counselling committee.

Figure 3: Operative Organisational Structure for SND 2.0



The steering committee with representatives from the consortium partners and the users will advise the director in strategic and scientific management. The director is ultimately responsible for the infrastructure and reports to the host university. The staff is organised in three operative lines and the line supervisors constitute a counselling committee to the director.

- ***Administrative support:***

Most administrative support in HRM, economy, communication, legal issues, and project management will be provided within the SND 2.0 organisation.

Required Equipment and Other Resources

A total of 100 per cent Director, 250 per cent Administrator, 100 per cent Communication Officer, and 50 per cent Lawyer is allocated for the administration. SND 2.0 is fairly independent within the university and responsible for most of its administration. A lower overhead is charged by the University of Gothenburg for the central administration. For externally funded projects, e.g. Horizon 2020, the project administration is the responsibility of this module even if the activities are in the other modules. Travel costs for general management and the costs of the steering committee is covered within this module.

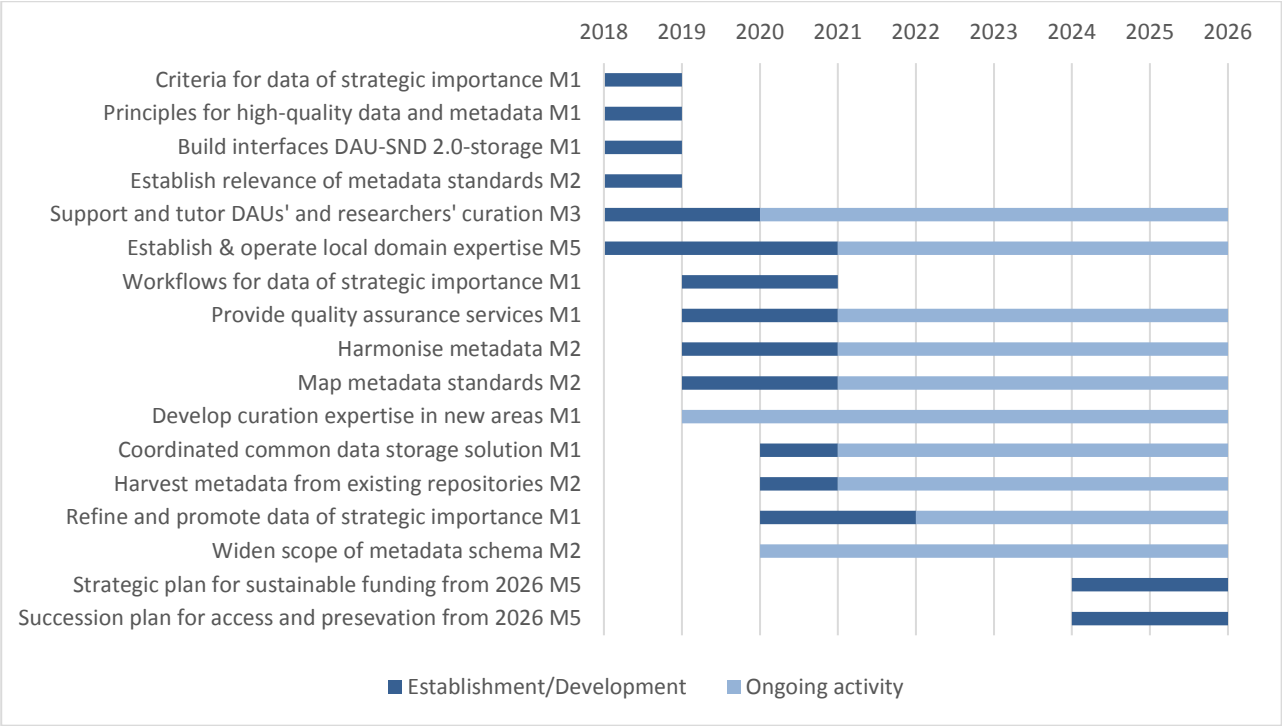
2 OPERATIONS

2.1 TIMETABLE

2018: Establish the components necessary for basic management, curation, preservation, and access within a restricted number of subject domains. Includes technical systems, organisational units, strategies, and training programmes.

2019–21: Widening the scope of the system in terms of domains, services, and technical solutions. Integration of operations through common workflows, automated processes, and standards.

2022–25: Further widening of the scope, with expertise in more domains added and solutions to complex problems (in terms of technically, legally, or ethically challenging data management and dissemination). High-level interdisciplinary interoperability and a Discovery Service that serves researchers from a wide variety of disciplines. Adaptation to emerging e-archive solutions.



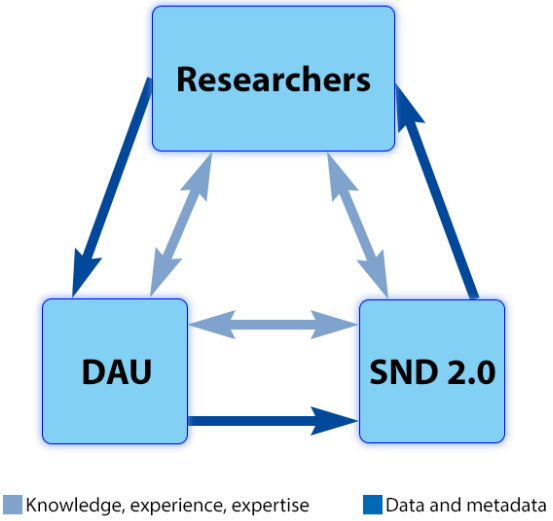
M1 through M5 are the various modules.

2.2 Current and Future Users

Researchers in Sweden are facing a growing demand for data publication, both from funders and publishers. Yet high-quality solutions to this new demand are scarce and researchers are often forced to use data-publication solutions that fail to meet requirements for long-term stability, adequate searchability, and legal compliance.

Over the next eight years, SND 2.0 will see a gradual shift in its user base, from separate researchers and research groups who avail themselves of SND 2.0's services as needed, to DAUs, which take care of rising volumes of datasets requiring curation. Whether they use local, external, or SND 2.0 repository solutions, DAUs will thus be a central user group. Researchers will still be welcome to consult SND 2.0, but will, at the end of the funding period, constitute a much smaller user group than today.

Figure 4: Users of SND 2.0 and Their Relations



Documented research data moves in one direction from researchers to DAUs to SND 2.0 to researchers while knowledge, in the form of domain literacy, curation expertise, data management training etc. moves in both directions between all parties.

Outreach to the two user groups are different but overlapping. DAUs are already being established with support from SND. Both open and DAU-specific workshops will be dedicated to common issues, joint solutions, and the establishment of national quality requirements. To reach the scientific community, other relevant organisations, and the general public SND 2.0 will communicate through several channels, such as websites, targeted information material, and visits to universities, academic conferences and other forums.

SND 2.0 will also arrange a wide range of outreach and training activities for researchers, DAU personnel, university librarians, and university archive staff to facilitate knowledge-sharing and development of best practices. (See the description of module 3 for how such activities are planned for the funding period 2018–2025.)

Non-consortium universities with DAUs will be required to contribute to the co-funding of SND 2.0. More information on this is provided in the budget. Commercial users will pay the full cost of SND 2.0 services.

2.3 Interaction with National and International Infrastructures

From its inception as national domain repository for research data from the social sciences, humanities, and health sciences, SND has been actively involved in pan-European endeavours to make research data accessible. Sweden is currently part of four Social Science and Humanities Research Infrastructures (SSHRIs), all of which have data sharing as one of their core goals: CESSDA, CLARIN, ESS, and SHARE.

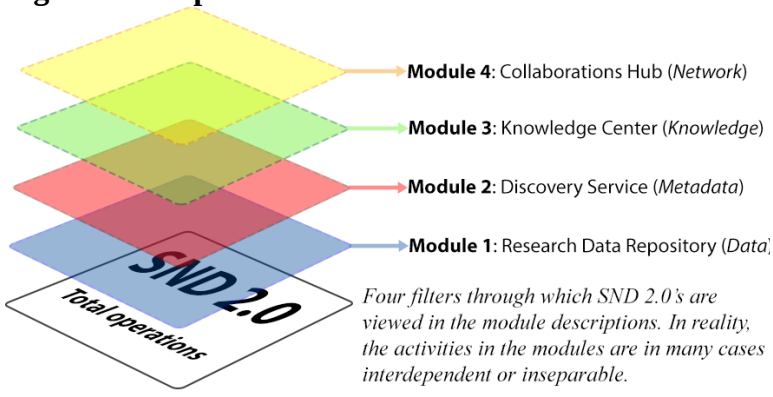
SND's collaborations in the Nordic arena have so far been largely concentrated in NordForsk and the Nordic data services in CESSDA. SND currently participates in the *Nordic e-Infrastructure Collaboration* (NeIC), an organisation for development and operation of high-quality e-Infrastructure in areas of joint Nordic interest; and *Making Nordic Health Data Visible*, a project aiming to build a common portal for Nordic data within the health sciences.

On the national level, SND is leading promising discussions with several infrastructures, aiming at cooperating in the system described in this application. Among these are SNIC, SUNET, DiVA, Språkbanken, SHFA, and the infrastructure for the integration and accessibility of data for climate-related research also applying in this call (S-NICE). Cooperation with these infrastructures will include various activities such as mapping or harmonising metadata, defining common standards, data management and dissemination support and possibly partly common management. (For further information on cooperation, see module 4.)

2.4 Operation, Structure, and Management of the Modules

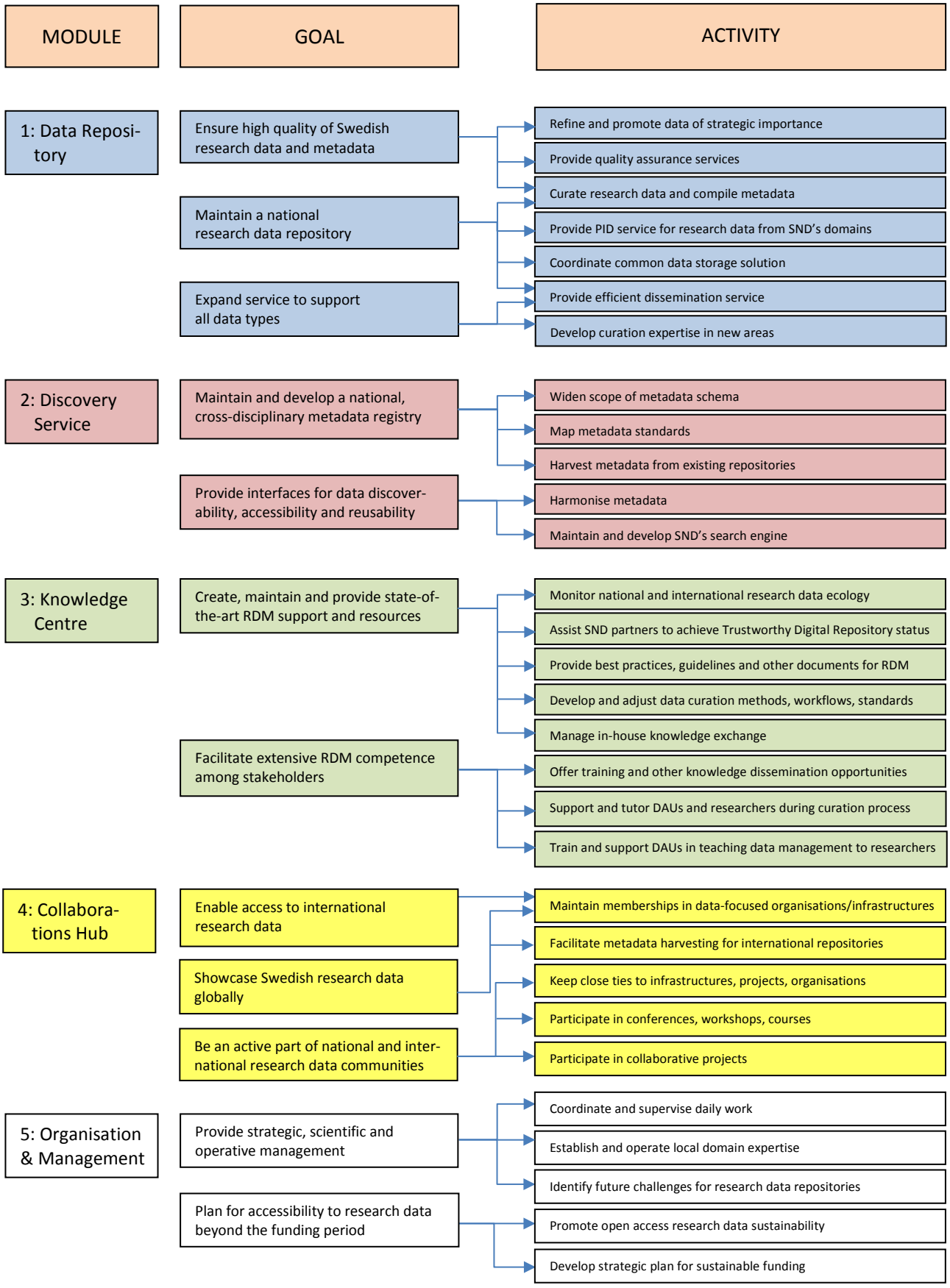
This section describes SND 2.0's operations as four modules, each of which details a particular view of the system, looking at it through a certain filter, as it were. **The modules are not self-contained work-packages or organisational units within SND 2.0.** The modules highlight, respectively, the handling, storing, and dissemination of *research data* (module 1); the ingesting, storing, and presentation of *metadata* (module 2); the acquisition, management, and dissemination of *knowledge and information* (module 3); and the establishing and maintaining of national and international *collaborations networks* (module 4).

Figure 5: Graphic Overview of Module Relations



Note that quality assurance, a key activity, has been divided over modules 1 through 3 because it contains components connected to data, metadata, and knowledge. In actuality, it will be carried out as a comprehensive whole.

Figure 6: Graphic Overview of Modules, Goals and Activities



2.4.1 Module 1. The Swedish Research Data Repository

Our repository concept defines the services necessary to create FAIR research data, including management, organisation, access, and dissemination. The concept *does not* include the actual physical data store, which will be hosted by a well-established infrastructure for large-scale data storage and (for legal reasons) separated into university-specific storage areas. This is described in section 3, Data Management and Supporting e-Infrastructure.

The Swedish Research Data Repository consists of repository services with data ingest and curation services, a dissemination service to provide researchers with access to data, and quality assurance.

Repository Services

With SND 2.0, DAUs at each university will assume responsibility for data curation and storage operations. Repository Services will provide the DAUs at each university with a common technical solution for uploading data and metadata to the large-scale storage area[1], thus saving time and resources in training and the provision of technology. A common solution will provide for consistent versioning protocols, ensuring that old versions remain unchanged, and managing authorisation and roles. SND 2.0 will compile national guidelines for versioning data, including rules for when and how new versions of data should be made accessible. The solution will also automatically provide persistent identifiers (PIDs) for uploaded datasets. Use of PIDs is a key element of good data management. PIDs make it possible to keep track of datasets, parts of datasets, and dataset versions, facilitating research to verify and reproduce a particular study. A major player in the field of PIDs is the International DOI Foundation (IDF). Data citation is made easier through a DOI, and cited data can be indexed like other types of publications, e.g., in Thomson Reuter's "Web of Science" Data Citation Index.

Whilst some universities will have DAUs in place to handle most curation from 2018, SND 2.0 must continue to provide assistance to researchers at other universities. Over time, the number of universities with experienced DAUs trained to curate and upload data will increase, although smaller universities will probably find it more cost-effective to enter agreements with SND 2.0 or a larger university in order to acquire access to DAU capacity. Most of SND 2.0's data curation will concern datasets that are complicated to curate – they may use old or unusual file formats, for example. Curation of such datasets will be managed in collaboration with the DAU in question, and solutions documented and disseminated to other DAUs.

Repository Services will also focus on data of strategic importance with a higher potential for secondary research. Such data will be identified by SND 2.0 in collaboration with the consortium partners, DAUs and university research groups, and will be further curated to be especially well described and easy to use for a large number of researchers. Examples of such data include datasets from major survey programmes such as SOM, Swedish National Election Studies, and the Swedish Level of Living Survey. These datasets will also be actively promoted.

Dissemination Service

Providing open access to research data is more complicated than open access to publications (VR 2015a:18). The dissemination service must differentiate between data that can be directly downloaded and data that require a login procedure or other checks, for instance indicated by a secrecy mark. Whether a dataset is subject to secrecy must be reviewed by the university that owns those data. A small number of datasets may, for various reasons, not be disseminated at all, indefinitely or for a limited amount of time.

Because technical systems for all situations will not be in place from the beginning of the funding period, the Dissemination Service will provide interim processes for dissemination, drawing on workflows and solutions already in use by consortium partners. The data dissemination service will also collaborate closely with the register data research initiative led by the Swedish Research Council to reach a broad range of researchers, and create good solutions for access to register data.

[1] For more details, see section 3.

The implementation of one universal solution is, however, improbable, and it is clear that research data will also be made accessible from a variety of other repository solutions. This includes for example the case of super-big datasets which will be hosted at their point of production rather than transferred to another repository. SND 2.0 will continue to support other well-established and well-managed domain-specific repositories in which considerable capital has already been invested.

Quality Assurance – Data

To make sure that the accessible research data reach maximum levels of reusability, QA-Data will guarantee that accessible data are stored and disseminated in a usable format with complete provenance and versioning information. Through distributed workflows and digital tools, as well as systematic spot checks, data files will be checked for viruses, readability, appropriate filenames, and complete datasets. When required, anonymised datasets will be examined to ensure that research subjects cannot be identified.

In practice, these quality goals will be reached through a control function that will inspect all data to be published and disseminated via SND 2.0. This control function will have access to the university-specific storage areas to be able to inspect data which the DAU wishes to publish.

Required Equipment and Other Resources

For module 1 a total staffing of 400 per cent Data Curation Specialist, 100 per cent Research Domain Specialist, 100 per cent Data Delivery Coordinator and 200 per cent IT is budgeted. No budget is allocated for the actual storage of data since SND will continue to use the free storage at EUDAT during the first years of the funding period. There is a long term solution proposed by SUNET, SNIC and SND[2]. Cost for travel to the DAUs at the universities and for participation in national and international data repository meetings are included in this module.

2.4.2 Module 2. The Swedish Research Data Discovery Service

Through the Swedish Research Data Discovery Service, SND 2.0 will provide national and international searchability, discoverability, and visibility for Swedish research data; support metadata standards and mapping that allow interdisciplinary searches; and offer a single portal for researchers in Sweden to discover international data they may need.

The Data Discovery Service will provide a way to search, ideally, virtually all Swedish datasets produced within an expanding range of disciplines. It will collect metadata on Swedish research data in one register, providing the technical solutions for all kinds of research data to become searchable and accessible regardless of location and discipline.

This module consists of: 1) construction of the national registry through ingestion of metadata from the university DAUs, and through harvesting of metadata from repositories other than SND 2.0's; 2) the Swedish Research Data Portal, a search portal for research data, based on the registry; 3) quality assurance for metadata.

Metadata Ingest and Harvesting

The extensive metadata registers already existing within the consortium will form the basis for a national registry. As university DAUs begin to deposit datasets produced by their researchers, they will also send metadata to the national registry. Once a metadata record has been created, the DAU is responsible for ensuring that it meets the requirements of well-established metadata standards. SND 2.0 will work with the DAUs to set up an interface between the local system, the metadata registry, and the data storage solution. Integration of consortium partners' research data systems will be prioritised, and over time, the systems of other actors will also be added.

For already existing domain repositories, and for producers of big datasets that will remain stored at the point-of-production, metadata will be harvested in order to provide as comprehensive a data catalogue as possible. Through mapping and harmonisation, it will be possible to add support for new

[2] See section 3 Appendix 1: *A Swedish national research data store*.

disciplines and types of data over time. Close collaboration with such existing solutions as Tilda (at the Swedish University of Agricultural Sciences) and the data module of the DiVA system, will be given high priority.

Swedish Research Data Portal
At the heart of the Swedish Research Data Discovery System will be the Swedish Research Data Portal, a portal that allows users to carry out advanced searches to find datasets they need. During the previous funding period, SND has already developed an advanced search portal for metadata,[3] and this portal will be developed for integration, optimised searches, and the facilitation of interdisciplinary research by providing discipline-neutral search tools.

The metadata records will include either a link for immediate download (for data with open access) or information about how to apply for data. The systems for dissemination will be developed and managed by the Dissemination Service (module 1).

Quality Assurance – Metadata
Through this module, the metadata quality required for data reuse will be assured. Metadata guidelines and checks of metadata stored with data will be developed. There will also be automated inspection of metadata before they are ingested from DAUs to the national metadata registry. All metadata must meet or exceed required standards for the particular disciplines, and they must be sufficient to live up to the requirements for a PID as well as for publication in international registries.

Required Equipment and Other Resources
For module 2 a total staffing of 200 per cent Data Curation Specialist, 200 per cent Research Domain Specialist and 200 per cent IT is allocated. Travel and meeting costs, primarily for meetings within the DAU network to share knowledge and experiences around data dissemination are allocated to this budget.

2.4.3 Module 3. The Swedish Knowledge Centre for Research Data

The Swedish Knowledge Centre for Research Data has as its central task to collect, compile, create, and disseminate knowledge concerning research data management, preservation, access, and re-use. Its particular focus will be on the SND 2.0 consortium's domains.

Through this module, DAU personnel will be offered training, support, and information, and support will also be offered to domain-specific repositories should they need it. An integral part of the knowledge centre is to actively contribute to the development of new workflows, best practices, and technical solutions within the data management and curation areas, both nationally and internationally. The knowledge centre covers the areas of data management and data management plans (including checklists, training material, and web guides) both from a preservation/access and a research-process perspective; metadata and metadata standards (including controlled vocabularies and thesauri); various data formats suitable for preservation and access, including (to the extent relevant) how particular formats need to be managed in different disciplines; data-documentation and versioning tools; and legal issues concerning research data and networks (e.g. universities, SUHF, VR, DI, SCB, the National Board of Health and Welfare).

Compared to the previous funding period, SND 2.0 will see considerable expansion of domain expertise in order to offer the best possible support and advice to its users. The period of 2018–2025 will see the integration of distributed domain expertise at the consortium partners. This will ensure close ties to the research communities covered by SND 2.0, guaranteeing relevance of the activities in this module.

[3] https://snd.gu.se/en/catalogue.

Acquiring Knowledge
To be able to offer solutions to the many challenges of data accessibility, SND 2.0 will maintain a high level of data management and curation know-how. SND 2.0 staff will participate in courses, seminars, and conferences, and conduct literature reviews of administrative methods of data management and curation and discipline-specific data management methodology. Collaborations with other organisations in SND 2.0's area of expertise will provide hands-on experience, through short-term visits and collaborations as well as long-term projects. They will also give new perspectives and ideas on how to improve processes and workflows. In-house development, not least of methods in which to manage data types previously not encountered, will be another source of invaluable experience. The application and development of metadata standards will support the efficient documentation of data types in order to maximise their usefulness for secondary research.

Gathering and analysing information about the multitude of factors and actors in the research data ecology will allow SND 2.0 to anticipate and deal with potential problems. The discussion forums established during the previous funding period will remain important venues for stakeholders to bring pressing questions and problems within research data management to SND 2.0's attention.

Managing Knowledge
Among the consortium partners, processes for knowledge management have already been established. With a new, distributed SND 2.0 organisation, managing knowledge will be of even greater importance. SND 2.0's knowledge centre will dedicate resources to systematically maintaining organisational knowledge. Chief methods of knowledge management include evaluations, repetitions, and discussions after courses, projects, etc.; in-house knowledge exchange; publications from in-house projects (also an important channel of dissemination); development of information and training material.

Disseminating Knowledge
The knowledge centre will supply the specific curation and data management expertise to a course under development at the Swedish School of Library and Information Science (University of Borås). This joint course is aimed at training new DAU staff and provide further education to already established DAUs during the first funding period. The centre will offer advice and curation support for DAUs and training activities for researchers and doctoral students. For universities without DAUs, the SND 2.0 knowledge centre will offer advice and support to individual researchers on contract.

One particular part of providing support is the development of tools for various data management tasks. An example is the Data Management Plan online tool, which SND is in the early stages of developing together with SNIC and a number of other national stakeholders. This tool will support researchers in writing plans that can be adapted to fit a variety of different requirements.

Quality Assurance – Knowledge
Workflows, checks, and guidelines will ensure that knowledge used and disseminated by SND 2.0 is kept up to date. Together with quality assurance for data and metadata, DAUs will gain assistance to live up to SND 2.0's quality requirements on data, metadata, and knowledge. SND 2.0 will provide training, guides, tools, and specifications that will allow the DAUs to perform at a level that makes Swedish research data trustworthy.

Required Equipment and Other Resources
For module 3 a total staffing of 50 per sent Senior Advisor, 250 per cent Data Curation Specialist, 400 per cent Research Domain Specialist and 100 per cent IT is allocated. Travel and meeting costs are allocated to this module to cover the active exchange of knowledge and experiences.

2.4.4 Module 4. The Swedish Research Data Collaborations Hub
Through its consortium partners, SND 2.0 is already part of a global, interoperable and accessible infrastructure, and this will remain an important part of SND 2.0 operations in the future. The aim of

the Swedish Research Data Collaborations Hub is to manage memberships in and collaborations with relevant national and international organisations, authorities, and infrastructures, maintaining and strengthening the networks that are essential for a research data system. The Collaborations Hub will ensure that researchers in Sweden have access to national and international datasets for secondary analysis and other reuse, and their research data will gain international exposure, leading to a broader awareness of Swedish research. This module will also ensure that Swedish research data practices are developed in a wide national framework and in an international context.

One particular type of collaboration will be with national and international funding agencies and scientific journals. Centrally located in the data accessibility ecosystem, SND 2.0 will be able to offer funders and journals the service of following up on whether their data-publication requirements have been met.

National Collaborations

A central task on the national level will be the consolidation of the consortium and its cooperation, and the additions of new consortium partners and DAUs. This also entails keeping close contacts with other data infrastructures, such as Tilda, the DiVA consortium, SNIC and SEAD (the Strategic Environmental Archaeology Database).

Getting data producers in Sweden together to cooperate on solving common problems is another major national task. While the Scientific Community Advisory Board will advise consortium management on matters of strategic importance within their domains, close cooperation is called for when practical problems need to be solved.

An obvious other reason for bringing data producers together is the goal of making their data FAIR. SND 2.0 must be sensitive to its users' needs and the users should have the opportunity to exchange experiences and information with each other. Data producers to be included in these networks include Swedish universities and other data-producing authorities, but also other large data producers such as national units[4]; and the five research groups that received the Swedish Research Council two-year coordination phase funding 2015 for infrastructures with databases within medicine and social sciences with focus on individual data: COHORTS.SE (The Swedish Cohort Consortium), SWEEP (The Swedish Survey program), REWHARD (Co-ordination to establish a national infrastructure for research about relations, work and health across the life course), NEAR (Towards a National E-Infrastructure for Aging Research in Sweden), and SwedPop (Swedish population databases for research). These will receive support similar to a DAU, including storage in the repository system, searchability through the portal, and data management training.

International Collaborations

SND 2.0 will continue ensuring that international digital research data are easily accessible for Swedish research and facilitating the use and citation of Swedish research data internationally. Both these tasks rely on SND 2.0 being part of a strong international network that enables SND 2.0 to maintain best practices in data management, curation, and dissemination. This network can be roughly divided into two categories of partners: research data infrastructures and other stakeholder organisations.

Close collaboration with other research data infrastructures, including other data repositories and ESFRI infrastructures such as CESSDA, CLARIN, ESS, and SHARE is important. In particular, joint projects with our Nordic neighbours will offer ways to share experience with sister organisations in countries whose legislation and research cultures are similar to Sweden's.

Other stakeholder organisations do not provide data, but are responsible for other necessities in the data publication field, such as developing metadata standards or providing forums for knowledge

[4] E.g. Swedish Polar Research Secretariat, the Nordic Information Centre for Media and Communication Research (Nordicom), ICOS Sweden (Integrated Carbon Observation System), and Swedish Institute for the Marine Environment (SIME).

exchange. Developers of metadata standards, international promotors of data-sharing, and contributors to findable and citable data will be essential parts of SND 2.0's international network, providing a global context for Swedish research data.

SND has also cooperated internationally through taking on roles in several EU-funded projects. It is likely that similar and new opportunities for EU collaboration will emerge in the future.

Submodule 4a: the Consortium of European Social Science Data Archives (CESSDA)

The objective of CESSDA is to provide a comprehensive, distributed, and integrated social-science data research-infrastructure, which will facilitate and support research, teaching and learning of the highest quality throughout the social sciences in the European Research Area (ERA) and increase the impact of the activities of its members.

CESSDA is organised as a distributed infrastructure where each member, represented by a national research authority, has a designated service provider (SP) that meets specific demands and requirements specified in the statutes. SND is the SP for Sweden and as such has to adhere to the CESSDA statutes in several different areas. The areas in which CESSDA is active are constantly evolving and the SPs have to adapt to new demands from both political and scientific entities.

As a mature SP, SND is invited to take part in several activities financed by CESSDA. These contribute, both practically and financially, to develop SND 2.0 as an SP and provide further services to the research community.

Required Equipment and Other Resources

For module 4 a total of 150 per cent Senior Advisor, 200 per cent Research Domain Specialist and 100 per cent IT is allocated. Of those, submodule 4a (CESSDA) has an allocation of 100 per cent Senior Advisor, 50 per cent Research Domain Specialist and 50 per cent IT. Participation in and organising meetings are key activities and included in the budget for this module.

2.5 Risk Analysis

The overall risk level in SND 2.0 is estimated as low. The consortium partners have substantial scientific and economic interests in a long-term commitment to a national multidisciplinary research data infrastructure such as SND 2.0. The possibilities for a financially sustainable expansion during the funding period and beyond are also considered favourable, thanks to the dedication of more than twenty Swedish universities supporting this application and to increasing activities among the research funders. SND 2.0 will systematically analyse risk areas during the funding period in order to detect possible problems early on. The following potential risks are identified:

Table 1: Risk Analysis

Risk description	Likelihood (1–5)	Impact (1–5)	Risk* (1–25)	Risk management actions
Difficulties in recruiting personnel with relevant expertise	4	4	16	Strategic human resource management; development of training programmes; active recruitment through established contacts
Loss of key personnel	3	4	12	Strategic human resource management; well-defined and documented workflows not dependent on single staff members; good career opportunities within the organisation
Difficulties to ensure access to data in case of terminated funding from 2026	2	4	8	Strategic plan for sustainable funding; succession plan including actions for access and preservation of research data
Problems related to the implementation of distributed organisation (SND 2.0)	2	3	6	Well-defined and documented workflows and structures for collaboration; short and well-defined decision-making processes at all levels of the organisation
Changes in the legal framework	2	3	6	Monitoring of legal framework development; cooperation with other stakeholders to adapt to changes
Underestimation of necessary resources	2	3	6	Options for additional funding will be explored; resource allocation will be reviewed
Delays of deliverables and milestones	2	3	6	Regular monitoring of time plan; rescheduling and revision of time plan if necessary
Management failure (unfortunate decisions)	1	5	5	Well-defined organisational structures; reporting lines and responsibilities will be reconsidered; problems will be referred upwards in the management chain

** Risk is estimated by multiplying likelihood by impact. Low risk = 1–8; Medium risk = 9–15; High risk = 16–25.*

3 DATA MANAGEMENT AND SUPPORTING E-INFRASTRUCTURE

Long-term storage of data is essential for SND but is not covered in this application since VR does not fund long-term storage (table E1 and E2 are thus not included). Currently SND is using storage at EUDAT which is sufficient for the coming couple of years and free of charge. To solve the increasing needs of large long-term storage the Swedish large data-producing infrastructures such as MAX IV, SciLifeLab and EISCAT, along with SUNET, SNIC and SND, have formed a working group to develop and pilot a common national long-term storage solution for research data. Swedish research infrastructures in several research domains do currently, or will in the near future, produce large and rapidly accelerating amounts of research data. This has created gaps in the research data life cycle that prevents valuable data from being properly managed and shared, and as a consequence limits the scientific impact and value of the investments made in these infrastructures. The working group can offer competences to help the users in Sweden of these facilities to better manage their data over its full lifecycle, from collection to publication and archiving. The group plans to work together in the interests of building a unified solution, pooling expertise, sharing competences, rationalising costs and facilitating a sustainable funding model.[5] The aim is to have the large long-term storage piloted, financed and running by 2020.

Because technical systems for all situations will not be in place from the beginning of the funding period, the SND 2.0 dissemination service will provide interim processes for dissemination, drawing on workflows and solutions already in use by consortium partners. The data dissemination service will also collaborate with the new Microdata Online Access (MONA) system at SCB, and the register data research initiative (Register Utiliser Tool, RUT) led by the Swedish Research Council, to reach a broad range of researchers and create good solutions for access to register data.

SND 2.0 will establish bilateral contracts with each university regulating the handling and delivery of data to researchers. SND 2.0 will act on behalf of the owner university (as personal data assistant [personuppgiftsbiträde] in the case of personal data). This arrangement will fit with the current Swedish laws and to the new EU regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

The cost of data management is not negligible but through coordination can be cost effective. The role of consortium partners will be to contribute domain expertise and access to networks and research communities in which particular kinds of data are used.

[5] See Appendix 1: A Swedish national research data store.

4 REFERENCES

Lynch, C.A. (2003) “Institutional repositories: essential infrastructure for scholarship in the digital age.” *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI and SPARC*, 226, 1–7. http://www.arl.org/storage/documents/publications/arl-br-226.pdf

RECODE project consortium (2014). *Policy Recommendations for Open Access to Research Data*. http://www.recodeproject.eu

Vetenskapsrådet (2015a). *Förslag till nationella riktlinjer för öppen tillgång till vetenskaplig information*.

Wilkinson, M. D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship.” *Sci. Data* 3:160018. http://dx.doi.org/10.1038/sdata.2016.18

A Swedish national research data store

Concept and Requirements

INTRODUCTION AND MOTIVATION

The Swedish large data producing infrastructures along with SUNET, SNIC and SND have formed a working group in order to propose a common national long-term storage solution for research data. Swedish research infrastructures in several research domains do currently, or will in the near future, produce large and rapidly accelerating amounts of research data. This has created gaps in the research data life cycle, that prevent valuable data from being properly managed and shared, and as a consequence limit the scientific impact and value of the investments made in these infrastructures. This group feels it can offer competences to help the Swedish users of these facilities to better manage their data over its full life cycle from collection to publication and archiving. We propose to work together in the interests of building a unified solution, pooling expertise, sharing competences, rationalising costs and facilitating a sustainable funding model. This paper outlines the functionality along with proposals for operation and governance.

INFRASTRUCTURE AND THE RESEARCH WORKFLOW

Research activities often rely on shared infrastructure (e.g. scientific instruments) to produce data which is analyzed, transferred, stored, retrieved, and in some cases published. The cost of operating cutting edge scientific instrumentation and the need to maximize use of such costly equipment often means that there is a need to streamline the flow of data through the various steps in a *research workflow*. Here, both the universities (where most researchers are employed) and the research infrastructure owner/operator play important parts. By spreading cost over many organizations it is possible to minimize total cost of ownership for the instruments, networks, computation and storage resources that are essential for research.

In Sweden, networks and computation resources have long been the remit of two separate organizations: SUNET and SNIC. Long-term storage – in particular archiving – of research data, while legally the responsibility of the data owner has never enjoyed the same attention and focus as networks and HPC resources. Recent focus on open data and open access means that new initiatives are needed to provide the additional capacity necessary to fulfil these new goals. This paper outlines the requirements for a common national research data store to partially address this issue.

Storage shares several important properties with large-scale HPC and research networks:

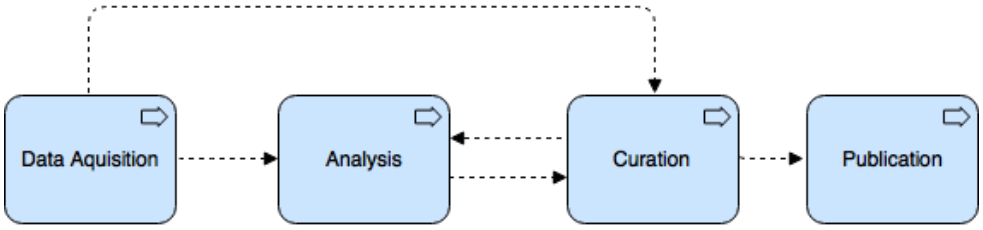
- Petabyte and Exabyte scale storage requires highly specialized knowledge
- Enterprise or Consumer products and services do not match scalability requirements
- Relative cost decreases dramatically with increased scale

Storing, transferring and retrieving datasets is not the primary focus of any scientist, and the advent of ubiquitous and “free” cloud storage (Dropbox, Google Drive, etc) has probably created an expectation among many that “the cloud” has boundless capacity and is always easy to use.
Modern cloud services are often used to drive research collaboration precisely because they are easy to use. The “easy” part of modern cloud services is sometimes achieved by optimizing for use-cases such as sharing documents and pictures and it is not always easy to carry the “easy” over to large scale research workflows.

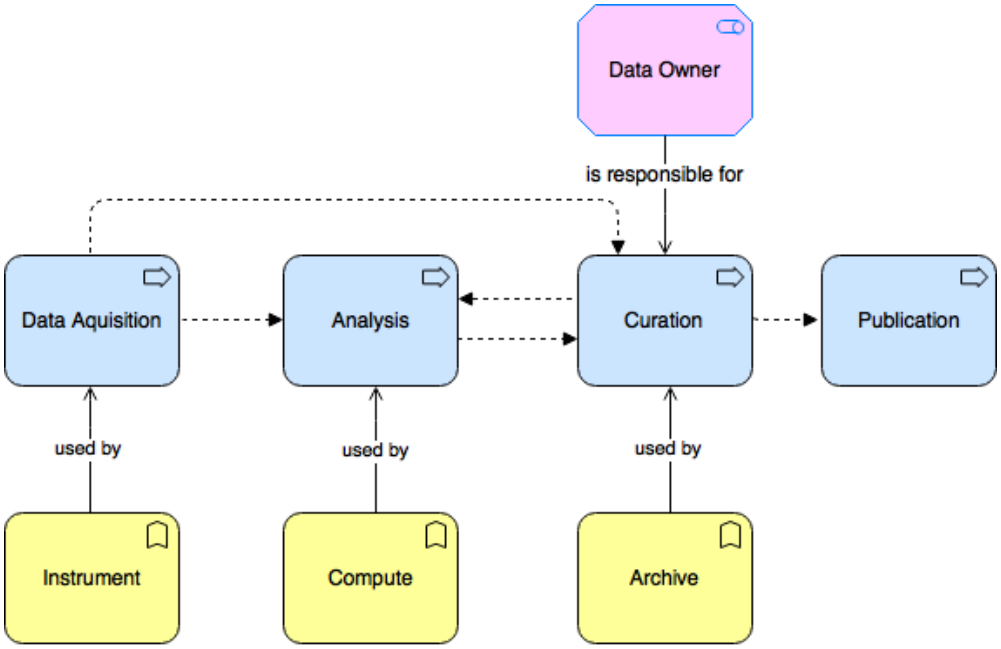
To support research relying on datasets at the scale of multiple 100s or 1000s of TBytes objects, scientists will most likely expect “cloud-like” services which can support user requirements while (as much as possible) exhibiting the simplicity of public consumer-style cloud services.

A Reference Model

A very simple concept diagram of a typical research workflow as associated with scientific instrumentation or other infrastructure might look something like this:



Data acquisition is followed by analysis, curation (archiving) and publication. Some data is retrieved from archives for further analysis etc. Adding the infrastructure functions that each part depends on and calling out the data owner, the model now looks like this:



In our model, an archive is a *service* that implements the *curation process* on behalf of a data owner. Often the term "archiving" is used synonymously with "storing" but there are important differences: specifically, the main difference between the processes of curating data and storing data is that the former focuses on metadata, access control, ownership and fulfilment of legal requirements while the latter is the business of keeping bits safe on electronic storage media.

For the purpose of this paper the distinction between Archive as a function and the physical storage of digital objects is very important. The Archive as a function is the legal responsibility of the data owner and as long as that responsibility is fulfilled, the location of the bits is a secondary consideration.

Both Compute, Instrumentation and Archive services rely on the safe and reliable storage of data objects regardless of which process these data objects are associated with. This does not mean that all processes can, or even should, share a common data store but it is probably true that several archives and data owners could share a common infrastructure if provisions are in place for keeping data logically and administratively separate.

In terms of this model, this paper proposes forming a collaborative process with the goal to establish a shared infrastructure for long-term storing of research data which could be used by data owners to establish cost-efficient archives and curation processes for research data.

Below we describe the functional requirements that a shared national storage infrastructure needs to fulfil along with some strawman design for how to realize the service.

FUNCTIONAL REQUIREMENTS

A national research storage service is a long-term engagement. To earn trust the service must appear to be a problem-free and limitless service to most researchers. In particular, the service must:

- Keep data safe for as long as the data is of value to the data owner
- Maintain persistent identifiers based on best practice (e.g. DOIs) for all data objects
- Enable data owners to fulfil the requirements of Swedish archive law
- Expose the underlying storage technology via a set of user-friendly technical interfaces
- Support flexible quota management and multiple per-group/area quota policy
- Support public access for select data objects
- Have a transparent and flexible cost-recovery model
- Have support for storing and managing sensitive data
- Connect and interoperate with other storage-infrastructure providers (e.g. EGI, EUDAT)
- Support multiple layered application with independent governance structure

STRAWMAN

The fundamental design principle can be summarized like this:

> *Make the common simple and the uncommon possible*

The design choices outlined in this section are meant as a starting point for discussion and are based on discussions and experiences from the scientific infrastructure and research networking community.

OPERATIONAL DESIGN

The national storage service (NSS) could be operated jointly by SNIC and SUNET on a cost-recovery basis. Cost would be shared equitably between data owners making use of the service. Cost sharing model, operational oversight and other governance issues would be the remit of the normal governance structure of SUNET and SNIC.

Establishing a service like NSS by a series of procurements is possible, if complex. However, as the service grows the requirement that the service appear “limitless” means that the growing financial risk may become too big for a commercial service provider to bear, or the service will become too expensive. In addition, the requirement to re-tender the service adds the risk of having to make sudden changes in ownership etc.

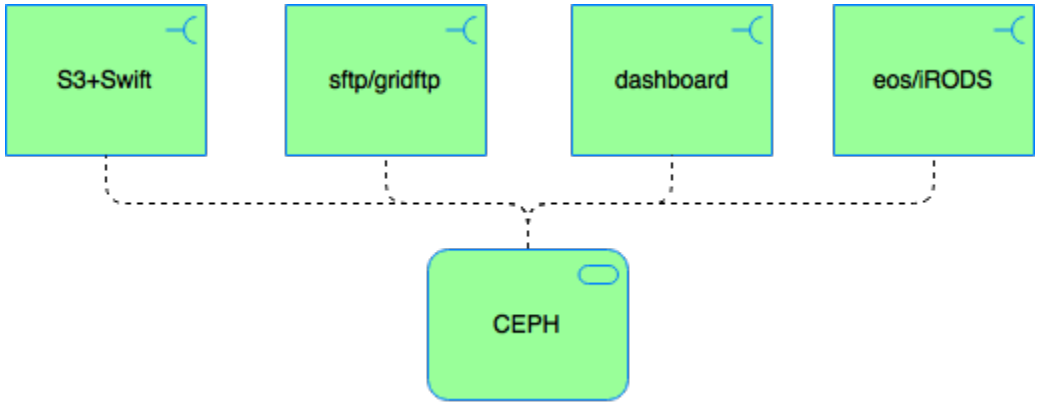
An alternative approach is for SUNET and SNIC to own most equipment and rely on service providers to operate the service. This approach would reduce the financial risk of the service provider (to zero) but at the cost of making the service entirely reliant on the budget process of SUNET and SNIC. This has several drawbacks, notably that sudden spikes in demand are almost impossible to manage without budget overruns.

A middle ground is a *public private partnership* which combines both investment by SUNET and SNIC and to a lesser degree by commercial service providers. The service would plan and execute build-out of storage

capacity based on the best available knowledge about upcoming research initiatives while the commercial service provider would be allowed to sell excess capacity in exchange for providing flexibility and the ability to quickly grow and change the service to allow it to adapt to sudden spikes in demand.

TECHNICAL DESIGN

The technical architecture should be based on an underlying object store technology – CEPH is a strong candidate – with built-in support for replication and scalability to 100s of PB together with a number of access interfaces designed to support multiple use-cases and user communities' needs. Here is a diagram illustrating this design:



CEPH is a mature and scalable object data store used by research organizations around the world to provide similar services. The service should provide multiple frontends to allow research communities to make their own decisions about what works best for them. Common choices will probably include industry standard interfaces like

- Amazon S3
- gridftp/sftp
- CERN eos
- iRODS

Other attempts to build similar services have typically focused on the needs of a single (albeit often large) user community which often has the effect of raising obstacles for other user communities that have made other technology choices. While there are no real universal standards in the field of storage access, the requirement to provide multiple overlapping interfaces will remain a very real one and is probably necessary for the service to be easy to use for all research communities in Sweden.

For the same reason, authentication and authorization must adopt the intrinsic requirements of each interface, which implies an architecture based on a token/authz exchange point (STS) where basic authentication based on SWAMID/eduGAIN credentials can be "exchanged" for the appropriate credentials for each interface. There are no silver bullets for authentication and authorization and the goal should be to make life for the researcher as simple as possible.

In addition, and perhaps as important as many of the other interfaces, the service should provide an easy-to-use end-user web dashboard that affords individual users the ability to perform simple operations on datasets, e.g. ingress, copy, rename, publish etc. This dashboard will of course be tied to federated authentication and could also be used as a one-stop shop for getting access to all the other interfaces, documentation etc.

The dashboard should integrate functions for annotating the stored datasets with rich metadata via the SND infrastructure. The metadata will make the datasets discoverable in several catalogues and referable via persistent identifiers e.g. DOIs. The creation of metadata will be a part of preparing the datasets for long-term storage.

NEXT STEPS

We will form a working group to come up with a complete proposal along with rough budget estimates for the first 3 years of operations. An initial solution will be piloted during 2018. The group will be formed by representatives from any data owner (infrastructure or university or research community) with more than 1PB storage need yearly, SUNET, SNIC and SND.