Report from the DASISH SSH Workshop, Gothenburg 4-5th October 2013

The workshop was held with the purpose of exchanging information and enhancing coordination between the infrastructures[1] and projects active in Social Science and humanities[2]. A central theme of the workshop was the way forward.

The first day of the workshop contained presentations from the five infrastructures: CESSDA, CLARIN, DARIAH, ESS and SHARE and from the projects present: DASISH, ADRIADNE, CENDARI, CHARISMA, DwB, EHRI, InGRID.

On the second day the presentations were followed up by general discussions about the future challenges facing the projects and the infrastructures. A central purpose of these discussions were to exchange information and views between the participants in the hope that this will help future coordination between them.

The five infrastructures have some longevity as they have completed or are near to completion of the establishment of legal entities, whereas the projects by the very nature of things will end at a known date. The cooperation between the involved project partners may continue and the resources created through the projects will need to be maintained.

Horizon 2020 will be one of the elements that can facilitate the future cooperation between the project partners and between the infrastructures. For this reason the actual implementation of Horizon 2020 is naturally of the highest interest. However it was noted that some cooperation might be beyond the SSH domain and that interdisciplinary cooperation should be encouraged where most salient.

This paper is a result of the discussion at and after the workshop and identifies some of the areas where the workshop participants find that future activities and cooperation is most needed and valuable.

1. Infrastructures are different

There is a lot of diversity in the social sciences and humanities and it is not possible to find one definition as to what a research infrastructure in the social sciences and humanities is, yet there is a spectrum. This needs to be taken into consideration. Basic types of infrastructures are either those that are based on a very large initial investment

[1] The five infrastructures are CESSDA, CLARIN, DARIAH, ESS and SHARE

[2] The invitation for the workshop was extended to the list of Networks of RIs funded under FP7 as Integrating Activities: (http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=ri_projects_fp7)

and modest maintenance cost or those that have a modest initial investment and much larger expenses on maintenance cost. Social science and humanities infrastructure are often of the second type. This has important consequences for how the infrastructure functions and more importantly how sustainability is achieved. Some SSH infrastructures are based around the provision of shared tools, others the provision of shared data and yet others the joint exploitation of data resources. With this range of different infrastructure forms support for SSH infrastructure needs to be flexible and multifaceted to ensure that the SSH infrastructures can be effective in answering the grand societal challenges.

Among the research infrastructures in social science and humanities there are clearly a different emphasis on the research aspect and the infrastructure aspect. All are working on improving accessibility, sharing and interoperability of architectures and solutions, however, some are to a higher extent more data oriented while others are more focusing on providing tools and services for the research community. This has implications for the support needed.

The wide range of project types - and thus of EC funding schemes (Integrated Activity capacity-building projects, RI projects, cluster projects, etc.) - clearly serves the goals set for the next generation of programmes. However, every effort should be made as to allow for a greater integration of projects with each other, which is crucially missing in the current generation. This is of particular relevance to the hardly-existing formal / institutional links between IA (Integrated Activity, former FP6 I3) projects and RI projects. Potential pathways from IA projects to RI projects should also be further explored.

This lack of integration is also applicable to the horizontal links between IA projects: in some cases, they are working on noticeably identical issues (e.g. metadata quality management, accreditation processes, legal issues, access conditions, etc.) and may advocate different solutions to such cross-cutting issues - thus being detrimental to their integrated capacity-building role. Though these projects should naturally coordinate their effort, it cannot be expected such coordination be as efficient as possible considering (1) their limited resources and (2) the continuous multiplication of such projects.

This said, some communality remains.

- There is a move to bigger tools and use cross-discipline to do it.
- Interdisciplinary is a possibility to be creative. For instance cooperation with infrastructures outside social sciences and humanities such as life sciences or the other clusters. At the same time new opportunities for collaboration within the SSH cluster have been identified.
- There is an interest in open data & linked open data & big data
- Social data (e.g. Facebook) creates new possibilities for data collection and research

2. Joint registries for Centre and Services and Social Sciences and Humanities Centres network

The need to establish a federation of trusted centres for research data that store, manage, preserve and give access to the mass of data in a trusted way is widely recognised. Trusted centres must fulfil certain criteria and undergo a regular assessment to ensure that their policies are adequate. This is especially required as a consequence 'policy based archiving' which is another item on the common cluster activities list. They will also need to provide certain standard entry points to support data access, monitoring etc. For example, to support citation and reproducible science a reliable and persistent service is needed. To provide access to confidential data, security has to follow agreed standards;

This topic has been accelerated by RDA-WDS (Research Data Alliance/World Data System) initiative and this is the reason that some people around WDS/ICSU[3] started now an initiative to suggest a common worldwide registry with human and machine readable information with agreed upon structure to allow automating procedures as it is known from computer networking.

- Funders want to see that the data funded by them will be stored in "trusted centres"
- There are some centres in social sciences and humanities that can play a role as trusted centres in the worldwide game:
  - CLARIN has now about 15 certified/almost certified centres
  - CESSDA has a network of 23 trusted centres, certification on going
  - DARIAH has some strong national centres (DANS, ADONIS, etc.)
  - There are strong libraries ready to fulfil the requirements
- However Europe is not evenly covered
  - There are less centres in the south east
  - Many humanities centres have insufficient resources to live up to future requirements.

The Social Science and Humanities should participate and it must be ensured that there are enough resources available to adapt all social science and humanities centres to participate in this world wide registry – it's about visibility, about recognition and so fourth. There is a need for funds to close the gaps and remain competitive. There is a close relation to the next point "policy based" archiving which describes ways to make the centers meet the future challenges of increasingly more data.

3. Policy based archiving

This is one of the essential requirements to establish trust in data providers in the long run. All policies that are applied by a centre should be based on explicit and declarative statements, which are then turned into executable, certified procedures. In the future there will be no other ways to assess the quality of centres. This issue is being addressed

[3] http://www.icsu-wds.org

in the DASISH project and in CESSDA, but more works is needed. There is also an initiative to create a registry of accepted policies for different tasks such as preservation, replication, curation, giving access, etc.

Data centres in social sciences and humanities are presently working to provide explicit policies. A point of start for expressing basic policies is OAIS. However, we need to go far beyond to meet the needs emerging from using concrete architectures and solutions.

This an urgent issue in many respects, for example, it will take a while to get these demo cases ready and it will take a while to get this all into production and to create training material, update procedures etc. so that others can start adapting. There is a need for immediate action for the future and a need for funds in Horizon 2020 to support this major change in how data are dealt with.

4. Sustainability

Sustainability is a very important issue that has various dimensions and the social sciences and humanities needs a long-term commitment in this issue.

Sustainability of resources created by projects is an obvious issue. In some cases the output of a project is something that is well dealt with by existing infrastructures – such as the five SSH ESFRI infrastructures, data archives or libraries. But in many cases the results created represent big challenges. Take for example a web resource presenting the legal and ethical rules presently in force in the social science and humanities domain in Europe. Such a resource will very rapidly be reduced to a historical document of limited interest if it is not maintained.

Moreover, a particular attention should be paid as to not multiplying the number of research infrastructures when the existing ones may not be stabilised yet (notably in terms of funding) for the reasons described in part 1. Though we agree the focus should be on the future prospects toward a competitive and world-class ERA based on strong and innovative research infrastructures, current obstacles and challenges related to the consolidation of the "existing" must not be left aside or overlooked.

Software tools and services represent a specific challenge. Although there is no escaping that maintenance costs money, paying attention to software sustainability can limit the costs. The subject should also be considered important from aspects as

- Verifiability of results: some results need the original tools to be reproducible
- Persistency of knowledge, cost-effective training of students and PhDs.
- Maintenance of created tools and updates of generated resources (e.g. databases)
- Limited funding – most of the existing SSH infrastructures do not have permanent funding, but live on a project like funding scheme with no security that funding will continue. Sustainability must be secure in order to avoid waste of investment
- In the social sciences a key challenge is the regular collection of data across as much of the ERA as possible. Ensuring the sustainability of data collection for

infrastructures like SHARE and the ESS is critical and the transition to ERIC statutes is causing considerable challenges in this respect.

- In the humanities similarly a broad coverage in terms of e.g. languages is necessary in order to live up to the basic idea of the infrastructures.

Next to applying better software development methodologies, there are also organizational strategies that can help. First of all enlarging the user-base of a tool should be considered, which also increases the possible funding base. So, domain specific, software registries with a purpose of sharing knowledge about the existence of specific tools are an important means to further this. Those registries should also support commenting from the user community and sufficient resources should be made available to check on entered tool information. Another organisational strategy to further tools sharing is the use of broker organizations; such an organization has (domain specific) expertise on the available tools and is able to broaden the user-base by searching new user-groups for a tool.

5. VRE: Virtual Research Environment

VREs are software applications that can integrate (existing) tools and services for a community's research workflow. It allows sharing of data and services and is a one-stop shop for (specific) research workflows. They can be considered either specific tools with considerable logic contained in one software or they are themselves flexible research infrastructures that can be modelled for use in a specific workflow.

Different groups and projects have been developing and using such VRE (like) functionality e.g.: TextGrid (DARIAH), or developing infrastructure use-cases that come close as CLARIN's (WebLicht+VLO+Annotation) and off course the NESSTAR software and other CESSDA products have been providing some of this for decades now. The DwB project underlines the importance of multi-level VREs for a European Remote Access Network that will allow the researchers to work together on confidential microdata.

Some important aspects of a VRE:

- Allow collaborations of (also virtual) organizations of users for access to services and resources
- Using virtual storage and remote services
- Use of registries for services and resources via metadata and PIDs: the (internal) administration
- Ensured interoperability between services and data
- Using data from the web (annotation & processing)
- Transparent archiving and publication of end product data
- For sustainability of VRE software one has to rely as much as possible on existing general infrastructure services rather than making specialised services within the VRE.

6. Crowd sourcing

Crowd sourcing is not entirely new, however *massive crowd sourcing* is and it will change how research is undertaken in many areas. Citizens will be able to actively participate as producers and consumers and the role of researchers will change.

The tools and infrastructure to support this increased interest in crowd sourcing are generally not in place yet. Therefore, new projects are creating new tools that are virtually identical to existing tools. What is needed is an infrastructure that provides the framework around the crowd sourcing projects so that the projects can concentrate on the actual subject matter they are dealing with.

Providing such an infrastructure will also create new possibilities but it will result in big challenges. What is needed is:

- Quick action to prevent future data losses due to amateurish not sustainable developments
- To involve small and medium sized enterprises, since they know how to do cross-platform programming for mobile devices and how to design user friendly apps.
- Design studies that address the issue of representativity in crowd sourcing studies and prevent resources being wasted on non-scientific approaches
- A strong European initiative that will ensure that the EU will have a strong hold on infrastructure level.

7. Big Data

There are increasing opportunities to access administrative and social media data and to take advantage of new data collection technologies. The SSH domain is central to these exciting opportunities. What is needed is:

- Support for design studies that facilitate innovation in this area
- Effective support for collaboration with the private sector
- Flexible cluster projects that allow true interdisciplinary working
- The close involvement of NSIs in collaboration with the CESSDA network
- The close involvement of Eurostat

On the other hand, this domain might also be highly relevant as it may create pathways and incentives for a better involvement of commercials - and a greater integration of the private sector with the broader research community - which is partly unsatisfactory in the current generation of programmes.

Prepared by:
Hans Jørgen Marker, (with excellent help from the workshop participants)

Workshop participants (27 people):

DASISH: 8 representatives

ADRIADNE: 2 representatives
CENDARI: 1 representative
CHARISMA: 2 representatives
DwB: 2 representatives
EHRI: 2 representatives
InGRID. 1 representative

SHARE: 2 representatives
ESS: 1 representative
DARIAH: 1 representative
CLARIN: 1 representative
CESSDA: 1 representative

Europeana/The European Library: 1 participant
Organisers from the 'Facing the future Conference' in Berlin, November 2013: 4 participants