Data Service Infrastructure for the Social Sciences and Humanities

EC FP7
Grant Agreement Number: 283646

Deliverable Report

Deliverable: D2.1
Deliverable Name: State of the Architectures Report
Deadline: M18
Nature: R

Responsible: UEssex, CentERData, KCL

Work Package Leader: KNAW-DANS

Contributing Partners and Editors: KNAW-DANS

www.dasish.eu

SEVENTH FRAMEWORK PROGRAMME

Contents

1. Executive Summary

This report examines the technical architectures of the five ESFRIs (European Strategy Forum on Research Infrastructures) which are CESSDA (Council of European Social Science Data Archives), CLARIN (Common Language Resources and Technology Infrastructure), DARIAH (Digital Research Infrastructure for the Arts and Humanities), ESS (European Social Survey) and SHARE (Survey of Health, Ageing and Retirement in Europe). It attempts to find areas of commonality which can feed in to the creation of a reference architecture for a more integrated and interoperable architecture for a Social Science and Humanities data domain.

An initial analysis of the published documentation found little commonality between the technical architectures of the five ESFRIs. The first phase of this approach found that three ESFRIs have a high number of sections describing how the architectural components will be implemented, and a different subset of three have a high number of sections describing what will be implement. The second phase looked again at the sections of documentation identified in phase one and concluded that only the CESSDA documentation set provided detailed information about some relevant topics, with CLARIN, DARIAH and SHARE providing some deep, medium breadth coverage.

This was followed up by conducting interviews with technical experts from each of the ESFRIs, to try to fill in some of the gaps. The conclusions focus on five key areas: Discovery and access; Tools, workbenches and data transformation; Data acquisition; Growing/sharing expertise; Innovation.

In essence, no uniform view emerged across the ESFRIs as the technical approaches taken are too varied. One or two emerging/best practice technologies were referred to (such as HTML5 and CSS3 to support adaptive and/or responsive design, Solr for Enterprise-strength search) which are already well known outside of the ESFRIs. However the documentation review provides a firm platform for the next task (Task 2.2 – Reference Architecture) to build on.

2. Acknowledgements

The authors would like to thank the following current and former members of the five ESFRIs who provided additional information to help enhance the documentation-based analysis, for which we are very grateful:

Name(s)	*ESFRI*
Ken Miller	CESSDA
Daan Broder, Przemyslaw Lenkiewicz	CLARIN
Tobias Blanke	DARIAH
Kirstine Kolsrud, Eric Harrison, Bjarne Øymyr	ESS
Dr. Frederic Malter	SHARE

3. Objectives and Methodology

This task lays the foundations for the rest of the work carried out by Work Package 2.

3.1 Overall WP Objectives

"To acquire and make available knowledge and expertise on both the current data architectures and tools and technologies in use within all five domains and liaise with on-going international initiatives working on data preservation and access issues – in particular the emerging data e-Infrastructure.
• To define & design a reference architecture, and to adapt it based on further insights gained by work packages 3, 4 and 5.
• To acquire an extensive list of requirements underlying the reference architecture, focused on integration and technical interoperability between heterogeneous systems, and determine which of these requirements would benefit the architecture and the domains the most.
• To provide a blueprint of the architecture and an assessment/validation procedure for the realization of an integrated architecture across the five domains based on existing components and foreseen results of work packages 3, 4 and 5."

Source: DASISH Annex I - Description of Work, version date 2011-08-15

3.2 Task Objectives

"The architectural solutions for data management, curation and access in the different SSH infrastructures must meet the requirements of the communities that they serve and are therefore very heterogeneous in nature. To make progress in creating a more integrated and interoperable SSH data domain DASISH needs to understand the current solutions to the community needs in great detail. It is likely that aspects of the research & data lifecycles for all the SSH communities coincide at certain points and for certain requirements, however these points of correlation have yet to be identified. This task will identify points of commonality in the five SSH infrastructures' community research & data lifecycles needs. Furthermore, current state-of-the-art solutions and best practices, in e-infrastructure for each community, will be identified and described. The goal is to describe the current solutions in each SSH community in a consistent manner which shall in turn inform the development of the reference architecture.

The output of this task will be an analysis report of common infrastructure requirements that is the baseline for the integration and interoperability work."

Source: DASISH Annex I - Description of Work, version date 2011-08-15

Methodology

The procedure was as follows:

- Analysis of the current state of technology in use in all five disciplines;
- Identification of architectural commonalities (if any) and state-of-the-art solutions to community needs;
- Description of the solutions in a uniform form;
- Suggestion of potential architectures that could be utilized in the Reference Architecture.

Each of these steps is described in more detail below.

4. Findings

4.1 Analysis of the current state of technology - Documentation Survey

The Analysis of the current state of technology in use in the five ESFRIs (see Appendix A – ESFRI Overviews) began with the publically available documentation (see Appendix B - ESFRI Documentation). The first pass (phase 1) took a grid-based approach, with Policy/Vision, Management, Diagrams, Code, Software, Enterprise on one axis and What, How, Where, Who, When, Why on the other.

The location, and from that the number of occurrences, of references to each topic area (e.g. Software vs. How) was recorded to give an overview of the coverage by each documentation set (see Phase 1 analysis summary on page 22), and hence provide a means of identifying areas of commonality between the ESFRIs. The summary table uses green lettering and shading to highlight ESFRIs that have ten or more references to any of the What, How, Where, Who, When, Why categories.

The second pass (phase 2) looked for areas that had narrow and deep coverage, by making a subjective assessment of the breadth of coverage (Narrow, Medium, Broad) and the depth of coverage (Shallow, Medium, Deep). The results are shown in Phase 2 analysis summary on page 23.

4.2 Analysis of the current state of technology - Follow-up Interviews

In an attempt to identify more common ground (with a view to building evidence of best practice tools and techniques to carry forwards for consideration in the architecture recommendations), phase 3 was conceived. This consisted of an interview with one or more technical experts from each of the ESFRIs, with the questions grouped by Policy/Vision, Management, Diagrams, Code, Software, Enterprise as well as a reflection on how the technical approach taken was proving to be in reality.

Systems Architecture

Questions asked:
Assuming there is a formal architecture:

- name (and briefly describe) the key components
- is there a split between tried and tested and novel approaches? (if so please elaborate)
- who decided on the architecture, and how (what was the process)?
- If no formal architecture, why was it thought unnecessary?

CESSDA

The CESSDA architecture is static, in that it has been in place for some years, without further development. Further development is anticipated once the CESSDA ERIC has been established, subject to the wishes of the partners making up the ERIC.

The CESSDA data portal is based on three elements: DDI, Nesstar and the Multilingual European Language Social Science Thesaurus (ELSST). It uses Shibboleth for access control (restricted to registered users).

DDI was developed by ICPSR with a team made up of IASSIST members including the UK Data Archive. It was first written in SGML but then convert to XML by the Danish Data Archive. It was simple, easy to understand, even readable. It had obvious advantages, consistent records and easy manipulation into other formats - for example WAIS, OAI and Nesstar. Tools were soon available – the UK Data Archive wrote a SPSS2DDI converter which it distributed to CESSDA and IASSIST.

Nesstar was initially developed through some CESSDA EU projects - so CESSDA members got it free. It was however unique in allowing online analysis of data – fairly simple analysis but adequate to allow a researcher, if not to fully carry out their research, at least to determine whether the data set was suitable and then they could download the data to their preferred, more sophisticated PC application. Nesstar converted the data part of the DDI XML to a NSDStat format. Basically this stored the data by variable, i.e. every case's response to a single question rather than SPSS's format which stored every case's response to each question. Nesstar also allowed input of the catalogue record and output of a full DDI record. A lot of CESSDA archives use this method rather than storing directly to an SQL database.

ELSST was developed through several EU CESSDA projects. It is a subset of HASSET. Whereas HASSET is regularly maintained and updated by the UK Data Archive, ELSST has not been maintained since the last project ended. ELSST actually is an English thesaurus translated into various languages. A concept is allocated a code and links are made to other concepts which are either narrower, broader or related. The concept, say E01, is expressed in English by the term Economics, and the same code is allocated terms in the various languages that participated in the projects. Hence there are CESSDA archives whose language is not represented in ELSST. Again, the terms and relationships are held in a SQL database which is converted to an XML format. This XML format is used by the CESSDA portal to display the related terms of search terms entered by a user.

Each time a dataset is published from Nesstar, an automatic indexer checks certain areas of the DDI record (e.g. keywords, question text) for occurrences of terms from ELSST and adds them to a running total in each language of the hits you would find with a search conducted via the CESSDA data portal.

At its inception, the CESSDA data portal was novel insofar as it was the first implementation of Nesstar and because it provided a ‘one stop shop’ discovery for studies held in dispersed locations. The multilingual functions for discovery were also novel. Although not implemented, further novelty has been considered including the addition of a module to enable question construction and variable harmonisation.

The principles of the architecture were developed from meetings between ‘coal face’ staff from CESSDA archives. User requirements meetings were also held and information fed back into the development cycle.

CLARIN

The main components are:

- PID services: minimum requirements, by default those provided by EPIC;
- AAI: SAML based FIM such as provided by the national Identity Federations and eduGAIN;
- CMDI based metadata schemata;
- SOA based on web-services and CMDI web-services services discovery.

The tried and tested approaches cover the implementation of:

- Handle PID system;
- OAI metadata harvesting;
- SAML based FIM AAI.

Whereas the following are novel and under development:

- Complex metadata search with semantic mapping via concept and relation registries.

Stakeholder input was provided by the CLARIN-EU Executive Board, which presented the CLARIN ERIC with input from the national centres. The architecture was only partially specified, as a total top down approach was considered counter-productive.

DARIAH

There is not a formal architecture as such, but some key components of the infrastructure can be isolated, including:

- Authentication and Authorisation infrastructure;
- Persistent Identification Service;
- Software Development Framework.

A trust framework has been developed to enable integration of novel services with the tried and tested core infrastructure.

The architecture was devised by the Virtual Competence Centre 1 (VCC1) - no formal process was followed as the key challenges in the fields DARIAH is concerned with are not tied to architecture, but to integrating the work of existing national infrastructures. In addition, the budget available to support development of a formal architecture was negligible.

ESS

The systems and architecture are being completely redesigned in preparation for transition to an ERIC, planned for June 2013. The architectural changes are expected to be in place at the turn of the year (2012/13).

There will be a number of novel approaches in the new system:

- more emphasis on searching using Solr for harvesting metadata from documents (final decision is yet to be made but the current expectation is that there will be multiple metadata cores);
- functionality to display researcher results;
- new version of the CMS (OpenCMS);

- HTML5 (to build in the potential for offering users more flexible access routes – these are not planned but could, in future, include, for example, access from mobile devices).

The first step was to identify what functionality was required. Then the architecture was built based on similar architectures already in place (use of other projects as models) and with a combination of new technologies where necessary and the application of known and trusted technologies elsewhere. The overall aim is standardisation.

SHARE

There is no formal architecture for SHARE. However, there is a general philosophy based on three principles:

- ex-ante cross-national harmonization;
- longitudinality;
- multi-disciplinarity.

SHARE combines all three principles.

There is a tried and tested development cycle (consisting of programming the questionnaire, holding meetings to discuss the new content of the questionnaire, translations, testing and train-the-trainer trainings for national multipliers). The sequence of development is very similar for every data collection wave, and this cycle is constantly improved.

There is also room for novel approach within this cycle. This is more on the content of the questionnaire. For example a Life History Calendar (LHC) was part of SHARE or more sophisticated innovations like collection of Dried Blood Spots, and Social Networks based on a name-generator approach have been tested in SHARE.

The architecture was decided by the initiators of SHARE and was guided by similar studies that were in the field before SHARE (like HRS and ELSA).

Policy/Vision

Questions asked:

- can the policy/vision be summed up in a single paragraph? (if so, what is it?)
- how does the systems architecture relate to/support the policy/vision (is there any formal traceability between them)
- what plans exist to expand infrastructure?

CESSDA

The vision was to provide a one stop shop from which users could identify, using native language, and locate data held in spatially separated data archives and undertake simple analyses to determine the value of the data for their research.

All the goals were achieved, and the system continues to be used by a number of data archives for publishing their data.

The need for further development was noted and discussed (during the PPP), particularly to make use of the features of future versions of DDI.

CLARIN

The vision is to provide a backbone of strong CLARIN centres that is responsible for creating a unified domain of language resources facilitating easy discovery and access by Social Sciences and Humanities researchers.

This will be expanded by adding centres from new CLARIN countries that implement the basic architecture that is specified by the ERIC.

DARIAH

The vision is integration activities across national infrastructure based on a lightweight EU layer across four strands:

- technical;
- data;
- research and education;
- advocacy.

As there is no formal architecture, no formal traceability exists.

It is expected that over time, services will be migrated from the community in to the DARIAH infrastructure.

ESS

The vision is for free, simple and speedy access to data, metadata and paradata for all users (internal and external) in all countries contributing to the survey; accessibility and usability; direct manipulation of data for those with limited statistical skills using Nesstar; ability to download data for others. Note the architecture supports an intranet for survey co-ordinators and their support teams e.g. for the cleaning and deposit processes. The system has a modular structure to facilitate future development

Several methods in place to link the architecture to the vision:

- An internal scientific team, within the management structure, regularly discusses ESS activities, e.g. access, processing and deposit times which, combined with rigorous reporting, ensure the architecture is supporting timely availability of the data;
- External feedback from monitoring of email queries, newsletters to users, user surveys and informal exchanges at workshops and other meetings;
- The co-ordinating organisation is represented on the development team which sets aims.

Development will be on-going, as the developers undertake technology watch and bring new suggestions and ideas to their regular meetings.

SHARE

The vision is to provide research data about the aging population in Europe free of charge to the global research community.

The system was built to accomplish the vision.

The intention is to broaden the institutional and geographic range of SHARE by adding more countries to the list of participants, and to introduce more cutting-edge novelties in the SHARE questionnaire.

Management

Questions asked:

- how does management know that the built system implements the planned systems architecture?
- is a formal review of the systems architecture planned (or has one taken place) to evaluate its effectiveness/suitability?

CESSDA

The system is now dated but its continued use by data archives suggests that it still meets a need. It was reviewed as part of the PPP project. There are currently no resources available to sustain it (the continued availability and maintenance of most of its components rely on the goodwill of the host archives, but note that Nesstar is a supported product in its own right).

A project is in place to enhance ELSST by unifying its terms with HASSET and building a single management interface. This work will start in 2013.

CLARIN

CLARIN Centre evaluation committee makes an appraisal of the centres' compliance
A formal review of the systems architecture planned.

DARIAH

There is no formal architecture, but in general terms, community feedback provides one avenue to inform management.

The infrastructure is reviewed every second year. In addition, there is an on-going review function performed by VCC1 (composed of three technical experts).

ESS

The current and new websites are being externally reviewed as part of an EU funded project

SHARE

SHARE is constantly evaluated by its scientific monitoring board and the European Commission evaluation committee. Also the scientific monitoring board of Max Planck reviews SHARE.

The availability of comparable data across different participating countries at the end of a fieldwork period indicates to Management that the technical infrastructure is delivering the intended results.

Design

Questions asked:

- have the system designs been updated between planning and implementation of the architecture to reflect reality (i.e. what are the documents showing us, planned or actual)?
- if not, are updates planned?

CESSDA

There have been no recent updates to the documentation, updates are only expected when resources become available through the ERIC.

CLARIN

All available relevant documentation was available for analysis. Updates will be made if relevant based on actual implementation.

DARIAH

As there is no centralised system, there is no homogeneous design.

ESS

Documentation is live and is continually updated.

SHARE

The software tools are always updated according to feedback from the field. There is a constant feedback process during and between waves.

Specification

Questions asked:

- was a formal design methodology adopted (why/why not)?
- if yes, what was it, and was it fit for purpose (why/why not)?
- any published APIs so others can access the services/data?

CESSDA

The design was reached through discussions between the system architects and developers, supported by user requirements exercises.

Nesstar has a public Java API (currently in beta).

CLARIN

Contributors may choose their own methodology. Communal efforts are documented on a WIKI, usually accompanied by UML diagrams. Available APIs are documented on the CLARIN developers WIKI.

DARIAH

A formal design methodology was adopted - Open Group Architecture Framework, which has proved to be fit for purpose.

APIs are not centrally organised but locally. At the moment work on an API integration framework is on-going.

ESS

The design was reached through discussions between the system architects and developers.

There are no technical barriers to making APIs public, but it is not a requirement. API's can be published if needed - it will be an issue for stakeholders to discuss in future.

SHARE

No public APIs are available for SHARE. For the SHARE fieldwork complete software packages are available for fieldwork agencies to use. The Data is only accessible for download, no published API is available for this purpose.

Implementation

Questions asked:

- what influenced the choice of implementation language(s)
- what influenced the choice of third party tools/frameworks (commercial and/or OS)?
- in retrospect, were the right choices made (and why do you say that)?
- any plans to make components available as Open source?

CESSDA

The implementation technology choices were driven in part a desire to standardise and for open access. Nesstar is implemented in Java, runs on Tomcat and uses MySQL database (all Open Source with a wide range of Operating system choices), whereas ELSST uses proprietary technologies (MS ASP.NET, MS SQL Server).

At the time, these were the right technology choices to make. Any future development work would be expected to take advantage of new technologies. Because the ERIC has yet to be formally constituted, it is not known if any components will be made available as Open Source in future, but to date none have been.

CLARIN

The choice of technologies was based on those that were open, proven, available and familiar. Also there is a goal not to become dependent on external providers, whilst considering using what is provided by open EU and national infrastructure initiatives e.g. eduGAIN.

In general the technology choices appear to have been the right ones. Sometimes too much faith was put in the promises of other projects to deliver components the CLARIN could adopt.

All the software is developed as Open Source.

DARIAH

Existing expertise within DARIAH played a key role in determining the choice of implementation languages, with reliability and previous knowledge being the guiding principles when choosing third party tools/frameworks.

It is not possible to say in retrospect if these choices we the right ones, because of the diverse nature of the infrastructure.

All the software is developed as Open Source.

ESS

The desire to standardise was a major factor in choosing the implementation languages. Third party tools/frameworks choices are based on the general technology strategy.

Java is the main implementation language. Solr and XML are also used. Content is managed with Open CMS. Nesstar is used with MS SQL and MySQL.

Open source not relevant for this project.

SHARE

Blaise is used as the implementation language for the questionnaires. This is because Blaise is a well-known accepted package for scientific questionnaires. Java and PHP are used for the panel management and translation packages, because these are free to use and have a large user community.

SPSS and STATA are used because they have a large user group in the target group of SHARE.

The researchers are happy with the SHARE data, which is a major factor in vindicating the technology choices that were made.

There are no plans to make the software Open Source, as it is very specific to the SHARE project.

Stakeholders

Questions asked:

- which segments of the stakeholder population were consulted as part of the requirements gathering phase? (e.g. data producers, depositors, users)
- which segments of the stakeholder population are addressed by plans/activities for each of training, help/support and outreach?

CESSDA

Data users (researchers, students, teachers), data producers/depositors (within the research community and some NSI's) were consulted during requirements gathering.

CLARIN

All stakeholders (data producers, depositors, users) were consulted during requirements gathering.

All segments of the stakeholder population are addressed by plans/activities for each of training, help/support and outreach, provided their needs are relevant and time is available.

DARIAH

The preparatory phase included a very detailed user requirements gathering phase. Stakeholders from all three segments were consulted, whereby most attention was paid to research users.

All three stakeholder segments (data producers, depositors, users) are targeted by training, help/support and outreach activities.

ESS

Use cases and stories are provided for different stakeholder types. Queries are recorded by type and issues arising are fed into the development process.

Training courses are not organised but there are a number of web-based support materials including videos - this could be expanded in future. Links are included to appropriate, endorsed materials elsewhere.

User consultations go to some 500,000 registered users, a recent survey elicited information for impending design changes.

Google analytics are used to inform decisions about enhancements.

SHARE

Fieldwork agencies, researchers, data users, data cleaners, sample designers, managers have a role in the requirements gathering phase of a new wave.

Everybody involved in the project is involved through the training. SHARE developed EASYSHARE for users/researchers (even for starting students) to make the data available in an easy understandable way.

Summary

Questions asked:

- if you were specifying, designing and implementing you system again from scratch, what would you do differently and what would you do the same?
- what are the plans for sharing code/methodology/components?

CESSDA

Any new system is likely to be built differently but would not be expected to risk losing key functionality and novel developments.

CLARIN

In future, perhaps be less optimistic about some external developments (such as eduGAIN).

Everything CLARIN develops is open and available for others. We plan especially to share with projects such as DASISH.

DARIAH
It is too early to say if the right technical choices were made.

DARIAH possesses an integrated software development infrastructure, open to everyone associated with DARIAH to use.

ESS
The new system is being built differently to the existing version.

SHARE
Except for minor issues, SHARE would proceed in a similar fashion.

There is an interface planned for scientific users to share their statistical programs. No other plans have been made at the moment.

5. Conclusions

The first phase of analysis of the published documentation revealed that 'How' and 'What' were addressed frequently by three ESFRIs each, and 'Who' and 'Why' by one ESFRI each. That is to say, there was little commonality across the board of the topic areas covered. Despite that, it did help identify the documents and sections in which to find relevant technical information.

The second phase showed that (subjectively at least) only the CESSDA documentation set provided detailed information about some of the topic areas, with CLARIN, DARIAH and SHARE providing some deep, medium breadth coverage. Again this suggested that there was little detailed information available that could be used for a firm comparison of approaches taken by two or more ESFRIs.

The interviews with experts from each of the five ESFRIs confirmed some of the initial findings and helped to fill in a few of the gaps. The findings for five key areas are detailed below.

The lack of communality means that suggestions of potential architectures that could be utilized in the Reference Architecture are not forthcoming, though one or two candidate component technologies are mentioned.

5.1 Architectural Commonalities

Discovery and Access
Present evidence suggests that there is currently relatively little provision for the discovery of digital tools available to researchers in the Social Sciences and Humanities. This is supported by the phase two analysis, which stresses a general bias within the infrastructures towards narrowness; although as noted above there is not sufficient consistency to draw any firm conclusions. By focusing on documenting smaller areas of activity in great detail, present infrastructures appear better placed to serve those already engaged with software tools than in exposing existing tools to wider user groups. A forthcoming report for the UK Arts and Humanities Research Council's

Connected Communities research strand (S. Dunn and M. Hedges, *Engaging the Crowd with Humanities Research*) stresses the importance of social media in engaging such users.

Tools, workbenches, data transformation

There is little evidence at present to suggest that there are many workbenches and tools repositories serving the social science and humanities research communities. The *www.arts-humanities.net* portal and Bamboo Dirt (*http://dirt.projectbamboo.org*) are two examples, but both (by necessity) privilege the gathering and presentation of data and metadata over the enabling of virtual collaboration. However, as is clear from the expert interviews, all the European infrastructures have in place advocacy, user consultation and collaboration layers to one degree or another, and this would seem to be a key area for coordination in the future.

Data acquisition

Unfortunately the ESFRIs have not much information available about their data acquisition tools and technologies. An explanation could be that this task requires much hands-on work with many partners involved that is not easily documented as part of the technical architecture.

Only SHARE has some documentation available in the form of PowerPoint slides (*Architectural overview of the software*). This material is used to train partners that will be responsible for the data acquisition. In SHARE the central organization supplies local fieldwork agencies with tools to do the complete data acquisition task in a homogenous way. Although the creation of these software packages requires an investment, SHARE benefits from this investment because data processing can be done more easily.

Growing/Sharing Expertise

Growing communities of use, broadening the stakeholder base and sharing expertise could be considered core to the development and sustainability of a research infrastructure, therefore we would expect this to be an important aspect to all 5 ESFRIs, though the approaches are likely to be diverse.

DARIAH has put this topic at the core of its infrastructure as it has been built in to the Virtual Competency Centres (VCC) structure. VCC2 is on Research & Education Liaison, which has tasks centred around:

- Understanding Research Practices. The aim is to conduct research on current and emergent scholarly information practices and needs. This forms the content of an up-to-date knowledge base and use registry.
- Training and Education Programme. An international summer school programme, development of online training and documentation materials, shared syllabi, and a registry of university courses amongst other things.
- Community Engagement. A series of expert seminars, workshops and publication and working paper are proposed for this task.
- Virtual Research Environment (VRE). The proposal here is to encourage modular, lightweight VRE systems, where the appropriate outside tools can be plugged in to the platform. DARIAH expects to produce a blueprint that can be used as a guide for the development of domain or project specific VREs.

As part of the VCC on Advocacy, Outreach and Impact, DARIAH proposes to undertake a broad-based analysis of skills and gaps, looking at methodological, technical and tools based knowledge in identified stakeholder groups.

The concept of a Knowledge Sharing Infrastructure (KSI) is at the centre of CLARIN's plans for developing and sharing expertise. Part of the KSI focuses on disseminating information about activities in CLARIN through newsletters, video tutorials, encouraging scholarly publications, exploitation of social media tools for communications and advertising campaigns (in other infrastructures this is handled separately, however integrating public relations strategies in to a knowledge network may give additional benefits in sharing expertise widely). There are plans for Help Desk facilities that are both automatic and human mediated to offer support to users who would like to engage with CLARIN technology and infrastructure. The Human Intermediated Question and Answer Service aims to allow users to directly interact with CLARIN experts. The Community Creation and Consolidation Service (KSI-CCCS) has the objective of publicising potential ideas for research projects and to find collaborating partners. The last component of the KSI is Continuous Education and Training Facilities. This aims to establish an educational alliance with universities for the distribution of e-learning materials through the alliance. Overall the KSI could be considered to be CLARIN-centric rather than a series of initiatives to support a network of expertise and knowledge.

The objectives of task 6 of the CESSDA-PPP are: to support capacity-building through developing the skills, knowledge and abilities of less-developed and less-resourced CESSDA organisations, by means of staff training and exchange programmes;
to foster and develop emerging CESSDA organisations through the provision of a complete 'tool kit' of standards, operational tools and expertise, allowing effective knowledge transfer. Although it is not certain that these activities will be conducted under the auspices of the CESSDA ERIC, it is worth noting that as part of Data without Boundaries[1] (DwB) the concept of a European Service Centre for official statistics (ESC-OS) has been proposed. This institution will be established on the basis of the existing CESSDA-ERIC network. "The ESC-OS online platform will attempt to build a community of contributors and serve as an important forum for advice, discussion and instruction on all matters pertaining to European official statistics (OS) data. A primary objective is to establish a mutual support network among users via a range of online social tools. Furthermore the ESC-OS proposes to implement a schedule of training courses, which are tailored to the needs of the research community, where there is an identified strong demand.

SHARE has two main strategies: 'train the trainer sessions' and local trainings in which participating agencies get trained in how to undertake a SHARE survey, with an interviewer manual in which everything is documented; easySHARE, which is a *"simplified data set for researchers who are less experienced in the quantitative analyses of complex panel data and for student training and teaching"*[2].

[1] http://dwbproject.org Combination of CP & CSA project funded by the European Community Under the programme 'FP7 - SP4 Capacities' Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics
[2] http://www.share-project.org/data-access-documentation/easyshare.html

ESS EduNet is a training resource mainly developed for use in higher education. The ambition is to create a social science laboratory where theoretical questions can be explored using high quality empirical data based on the European Social Survey. ESS also provides 2-day training courses 'ESS Train' that focus on key aspects of the survey lifecycle, with the specific aim is to equip researchers with the skills and knowledge they need to improve the rigour and equivalence of cross-national survey research in the European context. *"The courses also provide researchers new to comparative research with a unique opportunity to meet others also entering the field and to interact with acknowledged experts in cross-national survey design and implementation."*

Training researchers in e-Science techniques and the analysis of data is a common theme across all five ESFRIs, delivered online and/or face to face, however there are numerous approaches to developing and extending the user communities. It remains to be seen which of the approaches prove to be most effective as some are hypothetical at this point in time.

Innovation

Almost all ESFRIs talk about innovation in their project, however not always in the same way. CESSDA reported that innovation equates to creating a one stop shop in which data from different archives was combined. Highlights include the facility for researchers to query this data portal by using their own native language. Also the provision of a virtual catalogue of the archives' holding (i.e. the actual data remained distributed) was innovative. These innovations can be very useful for similar projects where data archiving is involved. However the individual innovations (like the native language query tool) could be used in many other situations.

CLARIN innovates in a similar way. They plan to query their data with complex metadata searches. Using semantic maps and relation registries they enable researcher to query data over different resources in a targeted way. Since the archive system knows much more about the data, the search results of a researcher can be much better. This technique could be used in complex data sources that are available in all of the ESFRIs.

ESS is innovative in a different way, they report on using advanced techniques in their system. For example HTML 5 is used in their websites which makes them suitable for modern browsers. An advantage of this choice is that it would be easier to make the data available for alternative devices besides PC's. Smartphones and tablets can render HTML5 pages in a more suitable way for the specific device. This opens new ways of publishing the data to the users. This is again an innovation that can be useful for all ESFRIs.

SHARE reported another way of being innovative - via their questionnaires. Not only is typical face-to-face interviewing used in SHARE to collect data, but also biomarkers like dried blood spots are collected and used for analysis. Innovative features are included within the interviews too. The third wave included a Life History Calendar interview, which made a reconstruction of the respondent's life in a structural way. In the fourth wave a social network module was tried in which information about the respondent's social connections was collected. These innovations are more specific to the type of research conducted in SHARE and are less likely to be applicable other ESFRIs (unless they do similar research, as ESS does).

DARIAH did not mention any specific innovations in their project. This however does not mean that no innovation takes place.

5.2 Potential Architectures for the Reference Architecture

In essence, no uniform view emerged across the ESFRIs as the technical approaches taken are too varied. One or two emerging/best practice technologies were referred to (such as HTML5 and CSS3 to support adaptive and/or responsive design, Solr for Enterprise-strength search) which are already well known outside of the ESFRIs. However the documentation review provides a firm platform for the next task (Task 2.2 – Reference Architecture) to build on.

6. Appendix A – ESFRI Overviews

European Strategy Forum on Research Infrastructures (ESFRI), the Social Sciences and Humanities ones are: CESSDA, CLARIN, DARIAH, ESS and SHARE.

"…mission of ESFRI is to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe, and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and international level."

Source:
http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfr

1.1 CESSDA

Council of European Social Science Data Archives

"An umbrella organisation for social science data archives across Europe. Since the 1970s the members have worked together to improve access to data for researchers and students."

Source: http://www.cessda.org/

2.1 CLARIN

Common Language Resources and Technology Infrastructure

"… committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities."

Source:
http://www.clarin.eu/external/index.php?page=about-clarin&sub=0

3.1 DARIAH

Digital Research Infrastructure for the Arts and Humanities

"… infrastructure will be a connected network of people, information, tools, and methodologies for investigating, exploring and supporting work across the broad spectrum of the digital humanities."

Source:
http://www.dariah.eu/index.php?option=com_content&view=article&id=3&Itemid=114

4.1 ESS

European Social Survey

“an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations.”

Source:
http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=23&Itemid=318

5.1 SHARE

Survey of Health, Ageing and Retirement in Europe

“a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social and family networks of more than 55,000 individuals from 20 European countries aged 50 or over.”

Source: http://www.share-project.org/

7. Appendix B - ESFRI Documentation

The publically-available documents that were examined and formed the input to Appendix C - Spreadsheet Analysis are listed below:

6.1 CESSDA

Reference	Title
D4.4	Recommendations and Guidelines Report
D5.5	Logging of portal data use, for statistical reporting and evaluation purposes
D5.6	Single Sign On Discussion document
D5.7	A CESSDA Enhanced Data Portal
D8.2a	Funding models for the future development of metadata standards and software tools
D9.2a	Functional and Technical Specifications of 3CDB
D9.3	Technical Specifications for a European Question Data Bank
D12.1	Version Control Final Prototype
D12.3	Authentication Prototype

7.1 CLARIN

Reference	Title
D1C-1	Implementation plan for nationally funded projects
D2R-1a	Centres Network Formation
D2R-1b	Centres Network Formation - Centre types
D2R-2a	Federation Foundation - LRT
D2R-3b	Federation Building - v2
D2R-4	Registry Requirements
D2R-5b	Registry Infrastructure - v2
D2R-6b	Web Services and Workflow Requirements - v2
D2R-7b	Web Services and Workflow Creation - v2
D2R-8	Service and Application Building
D2R-9a	Cost Estimates - v1
D2R-9b	Cost Estimates - v2
D3C-1.1	Working Group Formation Report and Activity Plan
D3C-3.2	Humanities Organizations, Initiatives and Projects Report
D3C-6.1	Integrated strategic plan for supporting SSH research
D5C-1	Report about outcome of liaison with other European projects and initiatives
D5C-2	Language Resources and Tools Survey and taxonomy and criteria for the Quality assessment
D5C-3	Interoperability and Standards
D5C-4	Description of the BLARK, the situation of individual languages
D5R-2	Usage Scenarios and Interoperability case studies
D5R-3a	Integration of LR - v1: Linguistic processing chains as Web Services: Initial linguistic considerations
D5R-3b	Integration of LR - v2
D5R-4	Validation of technical standards and infrastructure prototype
D6C-2.1	Revised Web-Site
D6C-4.1	Recommendations for future help-desk and advice infrastructure
D7S-2.1	A report including Model Licensing Templates and Authorization and Authentication Scheme
D7S-3.1	Collaboration Plan between CLARIN and external services

Reference	Title
D7S-4.1	Set of Federation Agreements for CLARIN centres
D8S-1.1	Requirements and best practice overview for governance
D8S-1.2	Analysis and proposal(s) for governance
D8S-1.2a	The shape of CLARIN (annex to D8S-1.2)
D8S-2.2	Financial plan for construction and exploitation phase
D8S-3.1	Requirements and best practice for transnational coordination and collaboration with third parties
D8S-3.2	Analysis and proposal(s) for coordination

8.1 DARIAH

Reference	Title
AP1	Technical Architecture (no date or version)
N/A	DARIAH Authorization and Authentication Infrastructure 2011-12-15
N/A	DARIAH Bit Preservation API 2012-12-22
N/A	DARIAH Business Plan D5.2 v1.0 February 2011
N/A	DARIAH Implementation Roadmap D8.2 November 2010
N/A	DARIAH Schema Registry (M1.2.1) 03/16/2012
N/A	DARIAH Storage API 2012-02-09
N/A	DARIAH Technical Report - overview summary (Nov 2010)

9.1 ESS

Reference	Title
N/A	Report to DASISH WP2, Task 2.1 'SOA Report'

10.1 SHARE

Reference	Title
N/A	Architectural overview of the software (PowerPoint)
N/A	CentERdata website (http://centerdata.nl/en/TopMenu/Wat_doen_we/ICT-toepassingen/questasy.html)
N/A	DDI Alliance website (http://www.ddialliance.org/sites/default/files/QuestasyDocumentingAndDisseminatingLongitudinalDataUsingDDI3.pdf)
N/A	Project website (http://www.share-project.org/data-access-documentation.html)
N/A	Technical and Scientific Description
N/A	Technical Description

8. Appendix C - Spreadsheet Analysis

The cell values in Figure 1 indicate number of occurrences of information of the specified type found in the available documentation for a given ESFRI.

Cell values indicate number of occurences of information of specified type found in available documentation for given ESFRI

CESSDA	Interrogative			
Abstraction	How	What	Where	Who
Diagrams		4	2	
Management		1		1
Policy/Vision	2	8	1	1
Grand Total	**2**	**13**	**3**	**2**

ESS	Interrogative	
Abstraction	How	What
Enterprise	1	
Management	3	
Software	1	1
Grand Total	**5**	**1**

CLARIN	Interrogative					
Abstraction	How	What	When	Where	Who	Why
Code	6	5	1	1	1	1
Diagrams	6	3		1	2	1
Enterprise	1	1	1	1	1	1
Management	7	5	1	1	1	1
Policy/Vision	8	8		2	4	1
Software		1	1	1	1	1
Grand Total	**28**	**23**	**4**	**7**	**10**	**6**

SHARE	Interrogative					
Abstraction	How	What	When	Where	Who	Why
Code	10	2				
Management	1	7		1	1	
Policy/Vision		6	1	1	1	2
Diagrams	2					
Software	3					
Grand Total	**16**	**15**	**1**	**2**	**2**	**2**

DARIAH	Interrogative			
Abstraction	How	What	When	Why
Code	5			
Diagrams	13			
Enterprise	1			1
Management	8	1	2	1
Policy/Vision	2	3	1	10
Software	1			
Grand Total	**30**	**4**	**3**	**12**

	How	What	When	Where	Who	Why
CESSDA	2	13		3	2	
CLARIN	28	23	4	7	10	6
DARIAH	30	4	3			12
ESS	5	1				
SHARE	16	15	1	2	2	2

Figure 1: Phase 1 analysis summary

Figure 2 identifies areas that had narrow and deep coverage, by making a subjective assessment of the breadth of coverage (Narrow, Medium, Broad) and the depth of coverage (Shallow, Medium, Deep).

CESSDA	Breadth
Depth	Narrow
Deep	2
Medium	14
Shallow	2

ESS	Breadth		
Depth	Broad	Medium	Narrow
Shallow	2	2	2

CLARIN	Breadth		
Depth	Broad	Medium	Narrow
Deep		7	5
Medium	4	6	16
Shallow	2	4	27

SHARE	Breadth		
Depth	Broad	Medium	Narrow
Deep		1	1
Medium	2	8	7
Shallow	11	7	

DARIAH	Breadth		
Depth	Broad	Medium	Narrow
Deep			3
Medium	1	3	19
Shallow	6	6	11

Figure 2: Phase 2 analysis summary

9. Appendix D – Interview template

The interview template used to elicit additional information about each of the ESFRIs was as follows:

DASISH WP2 - State of the Architectures Task

Follow up interviews
Aim of this activity is to flesh out the initial findings (see DasishT21_Phase2_0-6.xlsx) of the analysis of public technical documentation for the five ESFRIs.

The following sections and supporting text are intended to act as a guide for people conducting interviews with ESFRI technical experts.

Introduction

1. Overview: cover the objectives of DASISH WP2 task 1 (and the work package)
- produce report on common infrastructure requirements that feeds in to:
- (Reference Architecture report)
- (Tools and Services Knowledge Registry)

2. Methodology
- We are taking a grid-based view of the technical materials, namely
Policy/Vision; Management; Diagrams; Code; Software; Enterprise vs. What, How, Where, Who, When, Why
    - We are looking for areas that have narrow and deep coverage, as we believe that is likely to indicate practical experience, rather than a theoretical approach
    - As a technical expert, we would be grateful if you could help us fill in some of the gaps for your ESFRI

Think about IP angles for deliverable e.g.
- Discovery and access,
- Tools, workbenches, data transformation
- Data acquisition
- Growing/sharing expertise (in what, data analysis, tooling ….)
- Innovation – can it be done with tried and tested approaches/tools/methods?

Interview
3. Systems Architecture
Assuming there is a formal architecture:
- name (and briefly describe) the key components
- is there a split between tried and tested and novel approaches. If so please elaborate
- who decided on the architecture, and what was the process?

If no architecture, why was it thought unnecessary?

4. Policy/Vision (aka Scope)
- Can the policy/vision be summed up in a single paragraph? (if so, what is it)

- How does the systems architecture relate to/support the policy/vision (is there any formal traceability between them)
- what plans exist to expand infrastructure?

5. Management (aka Concepts and Relations)
- how does management know that the built system implements the planned systems architecture?
- is a formal review of the systems architecture planned (or has one taken place) to evaluate its effectiveness/suitability?

6. Diagrams (aka Design)
- have the system designs been updated between planning and implementation of the architecture to reflect reality (i.e. what are the documents showing us)?

If not, are updates planned?

7. Code (aka Specification)
- was a formal design methodology adopted (why/why not)?
- if yes, what was it, and was it fit for purpose (why/why not)?
- any published APIs so others can access the services/data?

8. Software
- What influenced the choice of implementation language(s)
- What influenced the choice of third party tools/frameworks (commercial and/or OS)?
- in retrospect, were the right choices made (and why do you say that)?
- any plans to make components available as Open source?

9. Stakeholders
- Which segments of the stakeholder population were consulted as part of the requirements gathering phase? (e.g. data producers, depositors, users)
- Which segments of the stakeholder population are addressed by plans/activities for each of training, help/support and outreach?

10. Conclusion
- If you were specifying, designing and implementing you system again from scratch, what would you do differently and what would you do the same?
- What are the plans for sharing code/methodology/components?