Long-term data for Europe

EURhisFIRM

M9.2: Recommendations for technical requirements
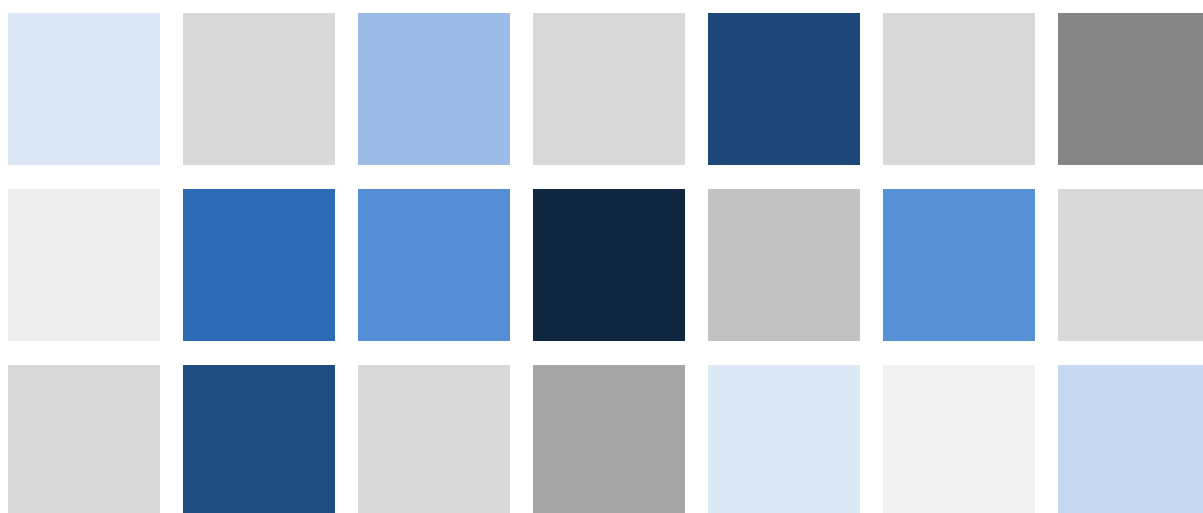
Deliverable	M9.2: Recommendations for technical requirements
Due Date of Deliverable	Month 36
Work Package	WP9: Infrastructure policy and architecture
Tasks	T9.1, T9.2, T9.3, T9.4
Type	Milestone

REVISION:

1.0, 14 May 2021

AUTHORS:

Coen Fierst van Wijnandsbergen, Onno Lissenberg, Stefan van Kampen, Abe de Jong (ROTTERDAM SCHOOL OF MANAGEMENT, ERASMUS UNIVERSITY)

APPROVED IN 2021 BY:

Jan ANNAERT (UNIVERSITEIT ANTWERPEN)
Wolfgang KÖNIG (JOHANN WOLFGANG GOETHE-UNIVERSITÄT FRANKFURT AM MAIN)
Angelo RIVA (ÉCOLE D'ÉCONOMIE DE PARIS)

Table of Contents

1. Introduction

This report is the final document of Work Package 9 (WP9): Infrastructure policy and architecture. After the completion of previous reports, this timing will enable us to reflect on project developments of the past three years and the architectural proposals made in other work packages.

Following the EURHISFIRM proposal of 2018, WP9 has the intention to describe a concise technical infrastructure and its policies for implementation of required functionalities. This report would be used as input by an implementation partner consecutively to estimate and build a working software solution while the consortium members take care of the organisation and processes. During the project timeline, emphasis was put on enhancing the existing systems to approach the EURHISFIRM proposal's ambition. At the same time, the promise of SSHOC as a functional contributing host environment became most relevant.

At this point, the EURHISFIRM project still has these choices to make:

- What level of centralism or federation will the final infrastructure have for this European RI?
- How shall the future EURHISFIRM infrastructure be operationally organised (governing bodies; operational, maintenance, and research personnel)?
- What functionality will SSHOC bring, and what will have to be developed within EURHISFIRM?

In the next chapter, we sketch the options and the consequences that each of these choices brings. First, we want to interpret the term "federation" in our context:

A federation consists of **autonomous entities** "*United by compact under a **central organisation**, as governments or commercial organisations.*"[1]

A federated architecture *"(…) allows interoperability and information sharing between **semi-autonomous de-centrally organised** lines of business (LOBs), information technology systems and applications. (…) The pattern emphasises a controlled sharing and exchange of information among autonomous components by communication via **messages**. Highest possible autonomy shall be given to the different cooperating components. In return they are expected to **adhere to common models by using defined interfaces**."*[2]

The concept of federation is supplemented by the concept of syndication: "*Syndication is a kind of **central authority** being able to interpret the federated model and compile meaningful information out of it. (…) Common to all such systems and organisations is a common semantic model and protocol, to which each participant agreed to adhere and behave like to a law.*".[3]

---

[1] https://www.wordnik.com/words/federated In the D5.1 report, the term **decentralisation** is used, which is interpreted to be synonymous with federated.

[2] https://en.wikipedia.org/wiki/Federated_architecture

[3] https://en.wikipedia.org/wiki/Federated_architecture

Key element in these definitions is the autonomy of systems, while ensuring interoperability via adhered interfaces and common models. This allows to decide that either a syndicate authority takes care of overarching functionality (like access by data consumers) or have the autonomous systems include this at their level.

2. Recommendations for technical requirements

2.1   Options for federation

For EURHISFIRM, the options below apply. Please note that the terms used are taken from D9.1, section 3.1.[4]

Option	Description
1	Local competence centres and locally maintained software. Nodes need to develop into Network Integration Centres (NICs) in future to support EU goals.
2	Starting from option 1, have one appointed competence centre develop into role of supporting central functionality instead of implementing support within all nodes (implement the syndication aspect in a single place).
3	Implement common EURHISFIRM functionality with a dedicated service provider like SSHOC. This option relies on the availability of SSHOC WP3-WP9. DARIAH might be used for archiving and publication of source documents.

2.1.1   Option 1 consequences

This option allows asynchronous development of local centres without the need to design the final RI first. The growth happens organically and the costs of development and maintenance are absorbs within hosting institutions. In practice, groups of researchers developing local databases feel more confident and controlled compared with data producers. Easily support stage 0, 1 and 2 (nationally consolidated).

This option is less efficient when arriving at stage 3 (CDM compliant) and 4 (EU consolidated). To de-duplicate data with other centres, all members in the federation need to write interfacing software and check duplicates with all other centres. With each new centre in the landscape, the complexity will increase, due to the multitude of interactions.

Data consumers either need to travel to gain access to data, unless all centres need to develop a web frontend and downloading functionality or replicate data from other centres to give local access. If mapping to authentic sources (like EUROFIDAI) is required, all centres need to set up subscriptions and maintain interfaces.

---

[4] D9.1 Research infrastructure policy 1.1, 3 July 2020.

The option to organically grow functionality locally will comply the least with the EURHISFIRM infrastructure policy document D9.1 and CESSDA software maturity level guidelines.

2.1.2   Option 2 consequences

The second option is to stimulate one of the contributing parties to develop parts of the centrally useful functionality, which will relieve local centres to have to create this. It is most attractive to have NIC functionality offered by this party. The development of this option may be taken step by step, and in an order that has been prioritised by the EURHISFIRM community (which is the syndicate). It needs to be decided whether one NIC will be sufficient.

The de-duplication at the EU level and the mapping to authentic sources could be taken care of by the NIC, as a service to local centres. This NIC will attract data consumers the most, as it will have in its database material that is the end goal of EURHISFIRM. These consumers may want to do further research on the data (like transformations) and return this to the RI. This option may be more in line with the business development ambitions expressed in WP10 when compared with option 1.

The NIC requires a granular implementation of access rights, agreed upon by the contributing local centres. The appearance of a central NIC in the landscape also requires the implementation of stages with the data. Some local centres may want to de-duplicate beyond their own database content from the start (such as the Data extraction system of WP7), but this would not be required for all. Implementation of stages within the metadata helps to identify the de-duplication work already done. Local centres may regain their rich data models and remain autonomous with respect to selecting datasets shared with the NIC.

Taken further, this option could develop functionality for aspiring institutes without any local system as well wanting to produce data at stages 1, 2, and 3. In that sense, the NIC could behave like a SAAS provider for Data Submission Units (DSU's): only a browser needed.

A concern with this option is that the NIC providing institute will have to devote relatively more budget and resources than the other centres unless a compensation scheme is put in place.

2.1.3   Option 3 consequences

In this option, comparable functionality of SSHOC replaces the functionality of federated centers, or of the syndicate NIC described in option 2. This option implies a "buy and interface" instead of "make" approach. The arguments to use SSHOC functionality instead of new development are (1) the scalability, (2) potential for economies of scale, and (3) the EURHISFIRM domain's exposure to other social sciences participants. This make or buy decision is relevant when the EURHISFIRM community requires a specific functionality to move ahead, and SSHOC has comparable functionality on offer.

SSHOC may require participants to adhere to standards before allowing usage of SSHOC SSH marketplace elements. SSHOCs guidelines and standards may supersede these of CESSDA and bring in other functional advantages free of charge.

The main disadvantage of this option may be the reduced autonomy in functional development.

Also participating teams that produce data may have concerns that they cannot control access by third parties. However, we expect the opposite in case SSHOC supplies an access rights model to all local centres, because data access and auditing will be better controlled than when developed locally.

Another issue may be usage costs of SSHOC components, which will be the biggest obstacle in writing a business case document for needed changes.

2.2   List of policy and usability aspects within the options

We've made a high-level assessment of WP9 infrastructure policy elements and the ability to implement these within the three options. We have ranked these from * (least compliant) to *** (most compliant). We see this as a starting point for discussion within the EURHISFIRM community and with the SSHOC SSH marketplace stakeholders.

Section from D9.1	Option 1	2	3
ISO25010 (2.1) And CESSDA software maturity levels (2.2)	* Most incompatible IT landscape. Each local centre must address software legacy to keep up and running.	** When NIC implemented at one centre (make).	*** When NIC implemented at SSHOC (buy).
FAIR (2.8)			
Completeness of stages and support for reaching stages (3.1)	* Stages implicit and known within centre.	*** Stages made explicit at NIC.	***
Support user types and granular access rights (3.2)	* Not offered. Access requires physical presence unless a web frontend is developed.	** Using standard components (make).	*** Access control with EU wide or possibly worldwide standard identification of users (buy).
Confidentiality Integrity Availability (3.3)	** Locally managed, legacy	** Locally managed. the CESSDA software maturity level guidelines indicate what agreements need to be made to ensure this towards the local centres.	*** Expect structural resources.
Security measures (3.4)	**	**	***

Section from D9.1	Option 1	2	3
	when implemented (make).	when implemented (make).	Expect structural resources.
GDPR support (3.5)	*** Requires only local support.	*** Requires coordination in syndicate.	*** Expect privacy organisation.
Usage auditing (3.6)	* Implicit in IT components, requires development of reports.	* Implicit in IT components, requires development of reports.	*** Design as part of SLA.
Time behaviour (3.8)	* Not explicitly defined.	** scalability at NIC level possible when using service provider.	** Design as part of SLA.
Controlled by hosting institute and staff.	*** Via IT department.	** Via IT departments, the CESSDA software maturity level guidelines indicate what agreements need to be made to ensure this. Communication platform for EURHISFIRM community controlled autonomously [5]	* Design as part of SLA, Communication platform embedded in SSHOC.

[5] As proposed in D9.2 Users project development unit design.

2.3   Recommendation

The 2021 situation for EURHISFIRM focuses on data production at the national level endorsed by local institutes. Identifiers are shared in bulk for de-duplication purposes. In this situation, option 1 works well. Adding new local centres with copies of software and databases exhibits this option's disadvantage. Where software gets forked, at least some of the disadvantages apply. [6]

We recommend setting a common understanding of needed functionality for each community member and plotting this into the future (in other words, a backlog). We also suggest creating a prioritised list of desired technical requirements (e.g. show a roadmap), supported by the community, and (collaboratively) implement these in an iterative and transparent process.

Further, we recommend having the most promising institute to host NIC functionality to benefit all members. We see these elements as necessary:

- The institute has access to funds or IT service providers to host common functionality. Note this may also mean this institute is able and willing to serve as a stakeholder within an existing RI to implement a functionality (like DARIAH for hosting and publishing scanned sources).
- The institute has access to IT development- and support staff to help write business cases to implement needed functionality. IT experts may suggest opportunities to collaborate/share infrastructure resources (knowledge, personnel, and possibly the actual infrastructure).

In short, we propose to adopt **option 2**, fully controlled by the prioritised requirements of the participating community members. We suggest turning a copy of the institutes' current database into the Common Data Model format and detail designing APIs for interfacing. [7]

The NIC could initially help develop the interfacing for the local centre systems as well. Start with one API to enable connections (e.g. linking algorithms) between the federated databases and demonstrate how 2 (national) competence centres could de-duplicate. Adding common functionality makes EURHISFIRM a more attractive platform to join for aspiring members.

---

[6] https://en.wikipedia.org/wiki/Fork_(software_development)
[7] See D9.4 Document on technical requirements, page 26 for an overview of needed interfaces.

https://eurhisfirm.eu

The SSHOC marketplace will develop solutions to some of the requirements prioritised within the community, allowing for a make-or-buy decision. As a SSHOC stakeholder, The EURHISFIRM community could put an effort in using our backlog to help shape SSHOC priorities.

This approach allows asynchronous developments locally as well as within the NIC. A disadvantage of this may be that some developed functionality may have a shorter lifespan because a published SSHOC product impacts the business case for make and maintain. A second disadvantage may be that this approach helps existing EURHISFIRM members more than aspiring new members without any local centre software, looking for just a browser and account to get going. If the backlog and business case mechanism works well, the NIC may develop just such functionality for a new member, based on funds introduced by them.