

# SSHOC: Realising the Social Sciences and Humanities for the European Open Science Cloud

---

EOSC Festival, October 2022

Tomasz Parkoła, Poznan Supercomputing and Networking Center

## Project



Horizon 2020  
European Union Funding  
for Research & Innovation

**Type of action & funding:**  
**Research and Innovation action**  
(INFRAEOSC-04-2018)

### Partners: 53

(20 beneficiaries + 27 LTPs + 6  
onboarded SSH RIs)

SSH ESFRI Landmarks and Projects  
& international SSH data infrastructures

### Project budget:

€ 14,455,594.08

### Duration: 40 months

(January 2019 – 30 April 2022)

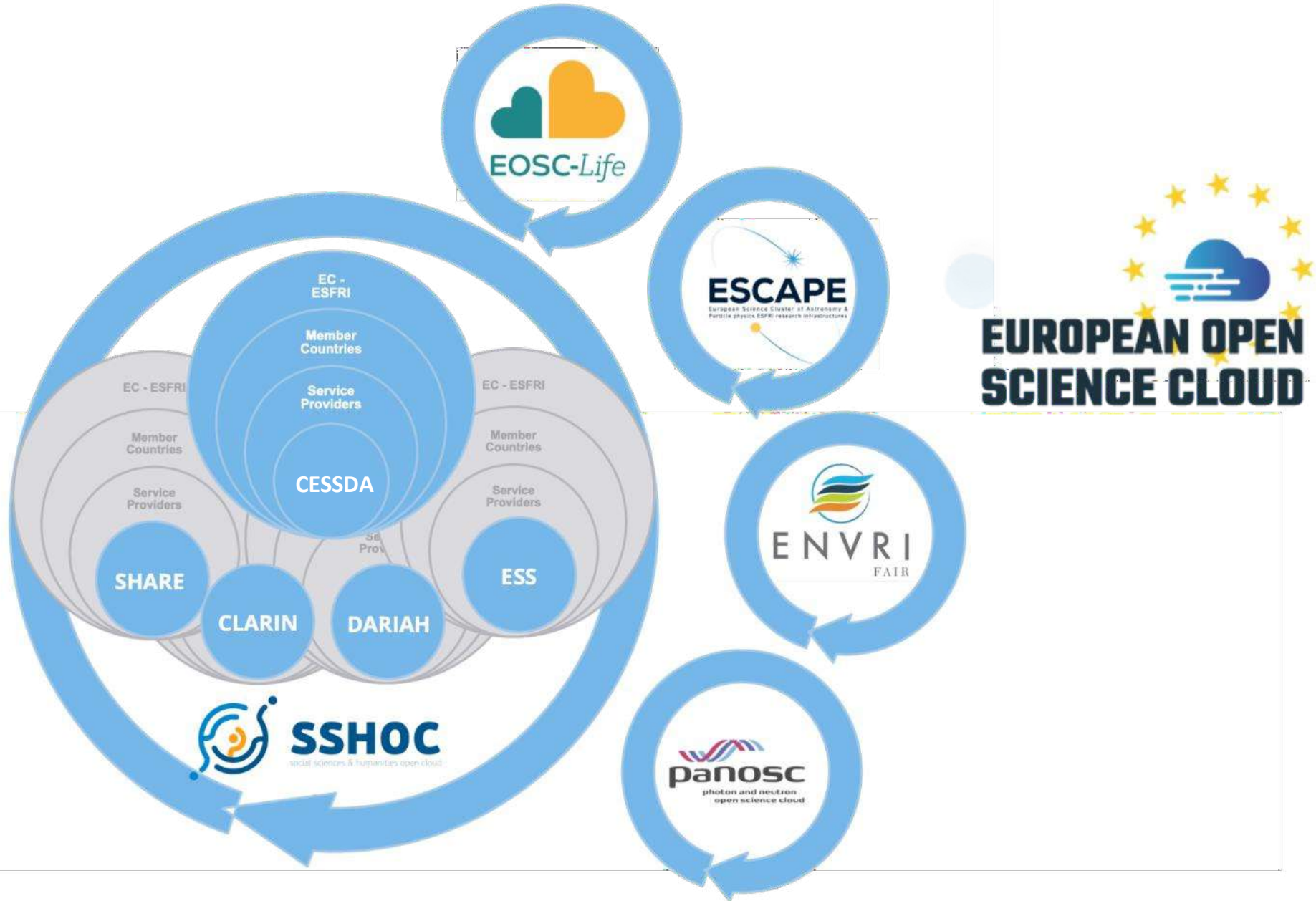
Project website:  
[www.SSHOpenCloud.eu](http://www.SSHOpenCloud.eu)



### Objectives:

- creating the social sciences and humanities (SSH) part of European Open Science Cloud (EOSC)
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

# Context



## Expected impact



The Social Sciences and Humanities are seamlessly integrated in the European Open Science Cloud



Availability of an EU-wide, easy-to-use SSH Open Marketplace, where tools and data are openly accessible



EU-wide availability of high quality "cloud ready" SSH tools and high quality SSH data



EU-wide availability of trusted and secure access mechanisms for SSH data, conforming to EU legal requirements



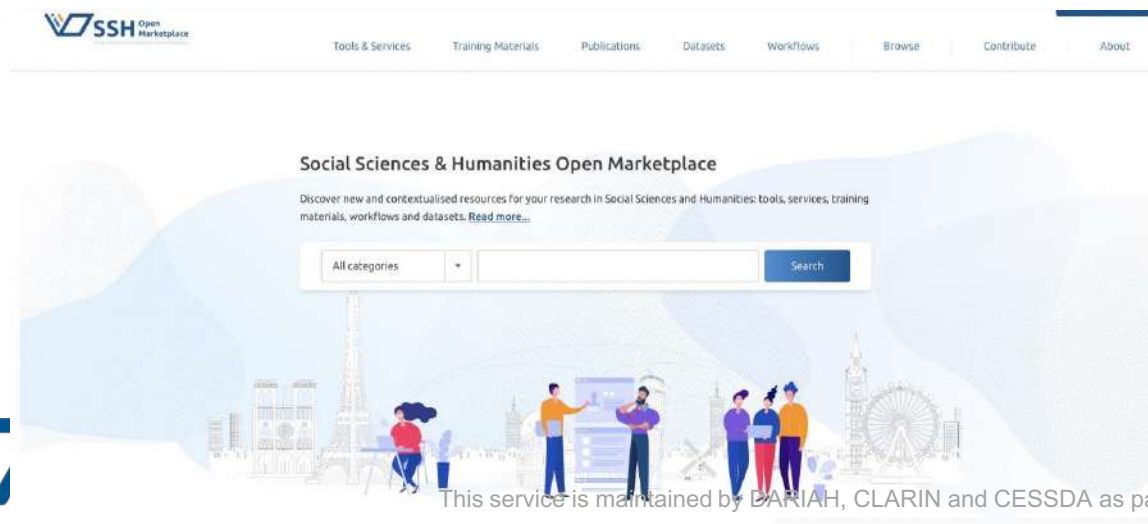
State of the art Research Infrastructure in several pilot domains advanced through dedicated SSH data pilots cluster projects



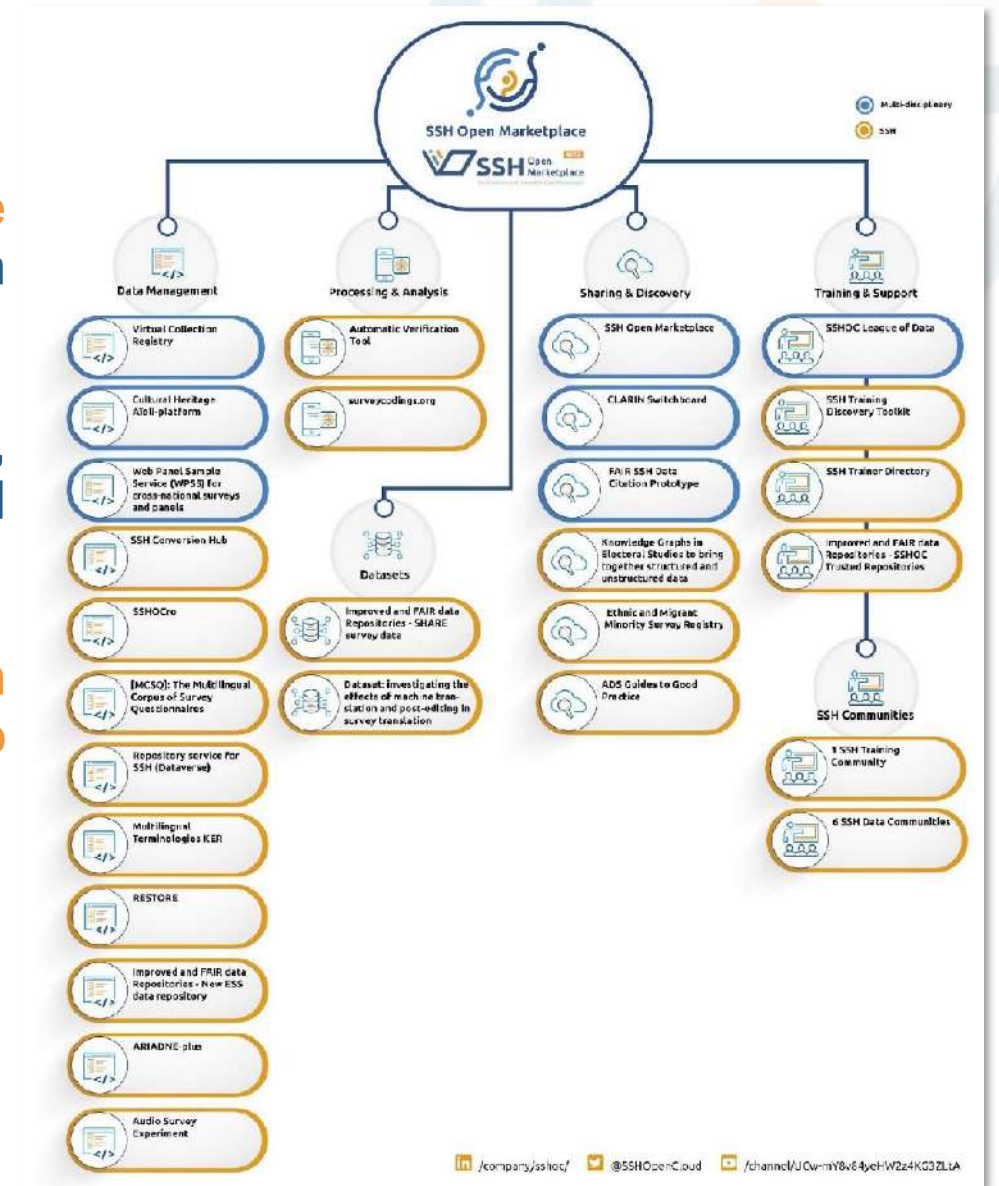
Maximising reuse through Open Science and FAIR principles (standards, common catalogue, access control, semantic techniques, training)

## Key Exploitable Results

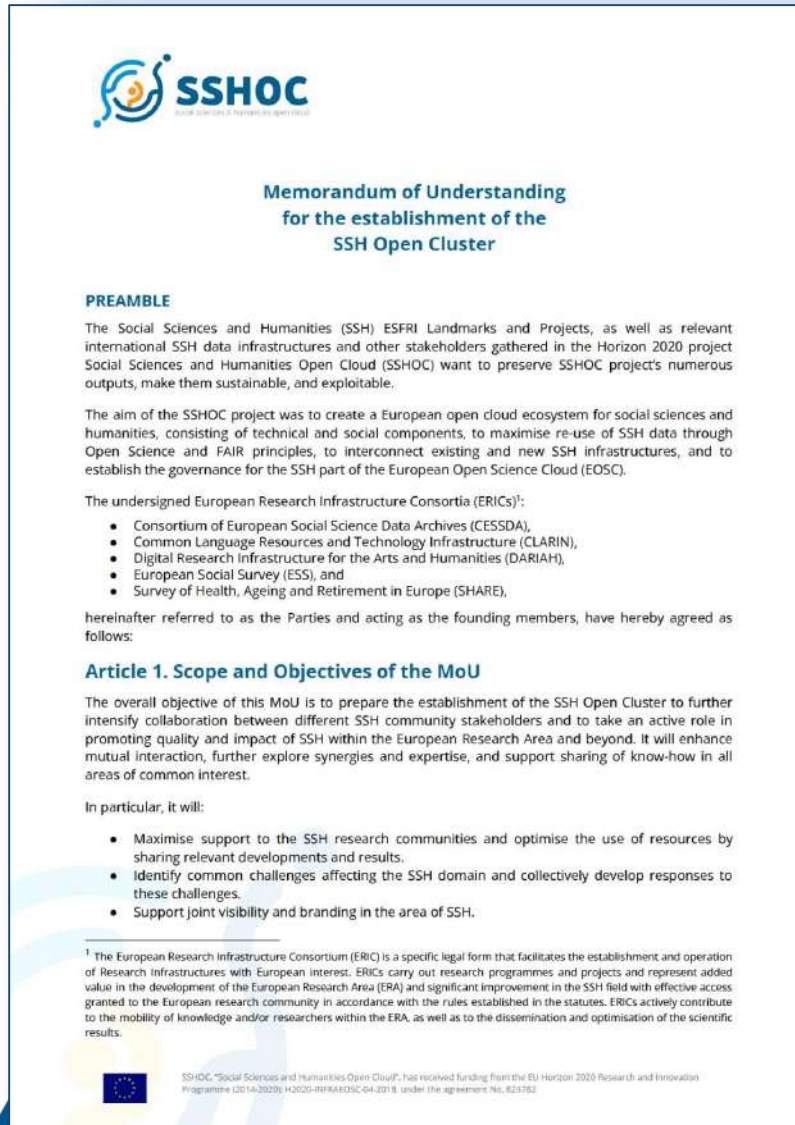
- 33 tools & services are defined as **Key Exploitable Results**, as they are available for further research activities after the end of the SSHOC project.
- Covering data management, sharing & discovery, training & support, processing & analysis and SSH communities
- All tools are being made **accessible via the SSH Open Marketplace**, and 13 of them are being **onboarded to the EOSC Portal**.



This service is maintained by DARIAH, CLARIN and CESSDA as part of the SSH Open Cluster



# SSH Open Cluster, future collaborations & MoU



- SSH ESFRI Landmarks,
- SSH ESFRI Projects,
- relevant international SSH data infrastructures,
- other stakeholders,

want to preserve Social Sciences and Humanities Open Cloud (SSHOC) project's numerous outputs - make them sustainable and exploitable.

## MoU objectives



## Social Sciences & Humanities Open Marketplace

Discover new and contextualised resources for your research in Social Sciences and Humanities: tools, services, training materials, workflows and datasets. [Read more...](#)

[marketplace.sshopencloud.eu](https://marketplace.sshopencloud.eu)

The SSH Open Marketplace is maintained and will be further developed by three European Research Infrastructures - DARIAH, CLARIN and CESSDA - and their national partners. It was developed as part of the "Social Sciences and Humanities Open Cloud" SSHOC project, European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, grant agreement #823782.





# Content types (see [about/data-population](#))



- **Tools & services** Materials or products used to perform activities: Desktop clients solutions (to be installed locally), Browser-based or command-line based resources, Mobile apps, Programming libraries or APIs, Data catalogues



- **Training materials** Tutorials, lessons or didactic resources explaining how to perform an action or highlighting the learning outcomes one would gain from engaging with the material.



- **Publications** Research results published in academic journals or repositories. The SSH Open Marketplace references only publications that can be connected to other resources and is not an exhaustive collection of papers.



- **Datasets** Single digital objects or collections of data, records, or information that is kept as a persistent unit of information in the knowledge generation process. Datasets are used as evidence for some phenomena.



- **Workflows** Sequences of operation/steps performed on research data during their lifecycle. Workflows can be achieved by using diverse tools and facilities, and useful resources are connected to each step.

  
[Search](#)[Home](#) / [Search](#)

## Search results (220)

### Refine your search

[Clear filters](#)

Sort by Last modification ▾

[◀ Previous](#) [1](#) of 11 [Next ▶](#)

#### CATEGORIES

- Tools & services 220
- Publications 1
- Datasets 1

#### ACTIVITIES

- Analyzing 592
- Visual Analysis 308
- Content Analysis 242
- Discovering 174
- ... 155

### OpenRefine [↗](#)

Activities: [Enriching](#) [Editing](#) [Data Cleansing](#)  
Keywords: [american](#) [Enrichment](#) [2010s](#) [transformer](#) [Data](#)

OpenRefine (formerly Google Refine) is a free, open-source tool for working with messy data. It enables users to clean data, transform it between a variety of formats, extend it with web services, and link it to databases. This tool is available for Windows, Mac and Linux.

[Read more](#)

### DIEF - Digital repository of the Institute of Ethnology and Folklore Research [↗](#)

Activities: [Data Visualization](#) [Disseminating](#) [Enriching](#) [Migration](#) [Searching](#)  
Keywords: [Feminist studies](#) [Memory Studies](#) [music](#) [Theatre](#) [data archive](#)

# Search results (36)

## Refine your search

Clear filters

Sort by Relevance

◀ Previous 1 of 2 Next ▶

### CATEGORIES

Datasets

36

### KEYWORDS

cultural-analytics

36

historical

26

literature

24

linguistics

21

historical-data

18

book-history

17

metadata

16

novels

15

english

14

fiction

14

More...

### SOURCES

Humanities Data

36



### Self-Repetition and East Asian Literary Modernity, 1900-1930



Keywords: fiction literature cultural-analytics japan china

Corpus metadata, extracted feature sets, stoplists, and Python and R code associated with this study.

[Read more](#)



### Stable Random Projection



Keywords: cultural-analytics machine-learning

Replication material and code libraries for the paper "Stable Random Projection." Frozen version of files available online at <https://github.com/bmschmidt/SRP-replication>. Binary files hosted at Northeastern University

[Read more](#)



### Dataset: Faces extracted from Time Magazine 1923-2014



Keywords: images cultural-analytics machine-learning magazines gender

The data presented here consists of three parts: Dataset 1: In this set, we extract 327,322 faces from our entire collection of 3389 issues, and automatically classified each face as male or female. We present this data as a single table with columns identifying the date, issue, page number, the coordinates identifying the position of the face on the page, and classification (male or female). The coordinates identify

[Read more](#)

## Steps (9)

1 Define the characteristics of the outcome

2 Define the characteristics of the image

3 Survey existing experiences

4 Choose engine based on the type of the content

5 Training the model

6 Test on a subset and assess quality

7 Correct output

8 Re-train the model with corrected output

9 Produce OCR output in standardized format

Expand

4 Choose engine based on the type of the content

Collapse

If you choose a classic commercial OCR engine like ABBYY FineReader, the customization would be limited. Others tools allow for more options (training a model for instance). It may be interesting also to use separate tools to perform different steps of the recognition process (Layout and character recognition)...

### Related (5)



Report on the comparison of Tesseract and ABBYY FineReader OCR engines

The aim of this report is to compare OCR accuracy of two well known OCR engines: Tesseract 3.0.1 and FineReader10...

[Read more](#)



Tools for text digitisation

More than 250 state-of-the-art tools for text digitisation. 61 results in group text recognition

[Read more](#)



Tesseract OCR

This package contains an OCR engine - libtesseract and a command line program - tesseract. Tesseract 4 adds a new\...

[Read more](#)



Transkribus

Transcribe. Collaborate. Share... ..and benefit from cutting edge research in Handwritten Text Recognition!

[Read more](#)



Språkbanken

A free cloud service for OCR The project 'En fri molntjänst för OCR' - A free cloud service for OCR, funded by the...

[Read more](#)

# SSH Open Marketplace Governance “scheme”

## 3 ERICs service agreement

DARIAH, CLARIN & CESSDA signing a collaborative and binding agreement and giving resources to maintain and further develop the SSH Open Marketplace after the end of the SSH OC project.

Service provision ensured by ACDH, CH and PSNC



## Research communities / User contributions

The SSH Open Marketplace is designed as to support community curation. Contributors can add or enrich existing content via a user-friendly interface.

Hands-on sessions are the main means envisioned to support contributions.

## Editorial Board & the day-to-day operations

It is made of members and administrators who take care of the maintenance, data quality and duration of the SSH Open Marketplace. They are also in charge of the community engagement.

Clara Petitfils, Suzanne Dumouchel, Nicolas Larrousse, Edward J. Gray, Laure Barbot, Arnaud Roi, Matej Ďurčo, Klaus Illmayer, Stefan Buddenbohm, & Tomasz Parkola. (2021). D7.5 Marketplace - Governance. <https://doi.org/10.5281/zenodo.5608487>

## PSNC contribution to the SSH Open Cluster and SSH part of EOSC

### Collaboration

- Represented in the Joint Research Committee of the DARIAH-EU
- Actively involved in DARIAH-PL e-Infrastructure developments (<https://lab.dariah.pl/>)
- Founding member of the Time Machine initiative (<https://www.timemachine.eu/>)
- More than two decades of actions in the cultural heritage and digital humanities (Europeana, IMPACT Centre of Competence, Open Preservation Foundation, DataCite)



This service is maintained by DARIAH, CLARIN and CESSDA as part of the SSH Open Cluster

# Thank you!

<https://marketplace.sshopencloud.eu/>

[sshopenmarketplace@sshopencloud.eu](mailto:sshopenmarketplace@sshopencloud.eu)

[tparkola@man.poznan.pl](mailto:tparkola@man.poznan.pl)

