DARIAH requirements and roadmap in EGI

Davor Davidović*, Eva Cetinić, Karolj Skala

Ruđer Bošković Institute

EGI Community Forum 2015

Bari, Italy

EGI-Engage is co-funded by the Horizon 2020 Framework Programme of the European Union under grant number 654142

Contents

What is DARIAH?

DARIAH, the Digital Research Infrastructure for the Arts and Humanities…

…aims to enhance and support digitally-enabled research and teaching across the humanities and arts.

It is a connected network of tools, information, people and methodologies for investigating, exploring and supporting research across the digital arts and humanities for researchers and humanists.

About DARIAH

- Established as European Research Infrastructure Consortium (ERIC)
- 18 member states:

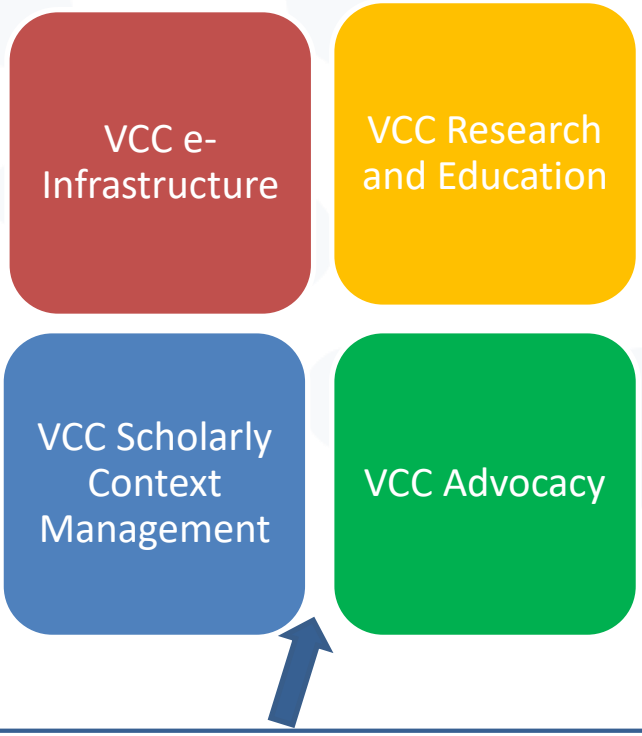
  Austria, Belgium, Croatia, Cyprus, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, Netherlands, Poland, Portugal, Serbia, Slovenia, Switzerland

- National DARIAH organization: DARIAH-DE, DARIAH-IT,…
- In-kind contribution + associated projects

ARIADNE CENDARI COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE CHARISMA DASISH

DARIAH Organization

Virtual Competency Centres



Cover strategic areas and topics, provide sustainability and incorporate the outcomes of working groups

15 Working Groups:

Text and Data Analytics
Natural Language Processing
Training and Education
Digital Annotation
Visual Media
Guidelines and Standards
…

Dynamic and flexible units with specific goals and outcomes, related to one or more VCCs

What are the DARIAH requirements?

- No unique answer on that questions
- Very heterogeneous community with numerous research disciplines, applications and tools utilized, types of media objects, etc…
- DARIAH has not conducted any comprehensive survey on the member's technical requirements
- Therefore…

…***hard to define DARIAH needs and provide the right solutions!***

The goal of the survey

To collect information on e-Infrastructure **requirements**, **experience** and **needs** of the A&H research community

- We want answers to the following questions:
  - Who are the targeted research group
  - What is their technical background
  - Which application/services/tools they use and how
  - What are their current and future needs for e-Infrastructure
  - How the digital objects are stored, managed and shared

Collection process

- Online survey rounded with the interviews (link)
- Tools: LimeSurvey and skype
- Collection period:

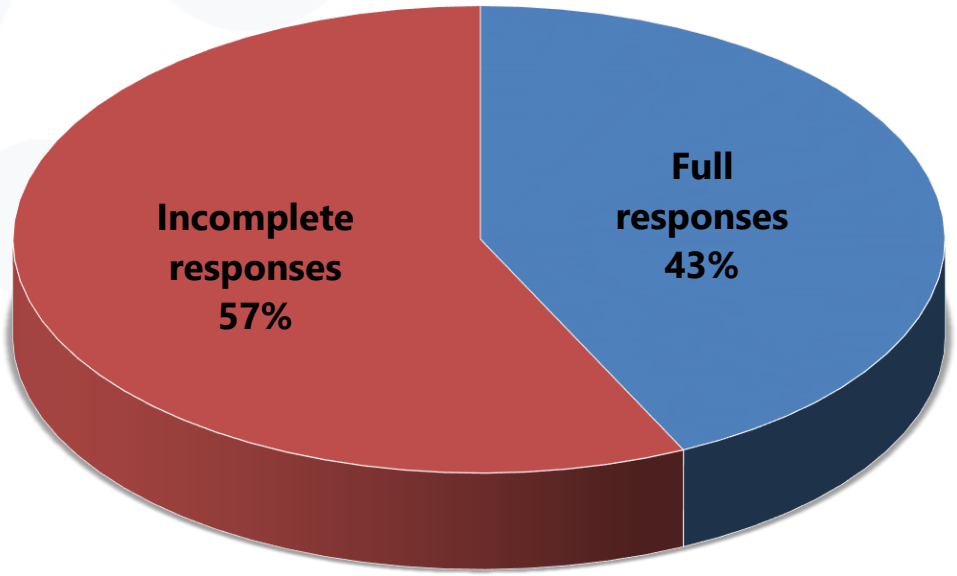
  15th August – 30th September 2015
- Total number of questions: **65**
- Questions divided into groups:
  - About participants (3)
  - Experience with e-Infrastructure (3)
  - Authentication and authorization (6)
  - Digital Arts and Humanities assets (4)
  - Data management – sharing and accessing data (17)
  - Services and applications for data analysis (9 x 3)
  - Future planning (3)
  - Contacts (optional)

Collecting process – target population

- 3 different user roles:
  - ***Application/service developers***, i.e. computer scientists who design, develop and/or implement applications and services used by other DARIAH members
  - ***Application/service providers***, i.e. computer specialists who are responsible for providing applications/services to other DARIAH members
  - ***Researchers*** in digital arts and humanities, i.e. consumers of the applications and tools

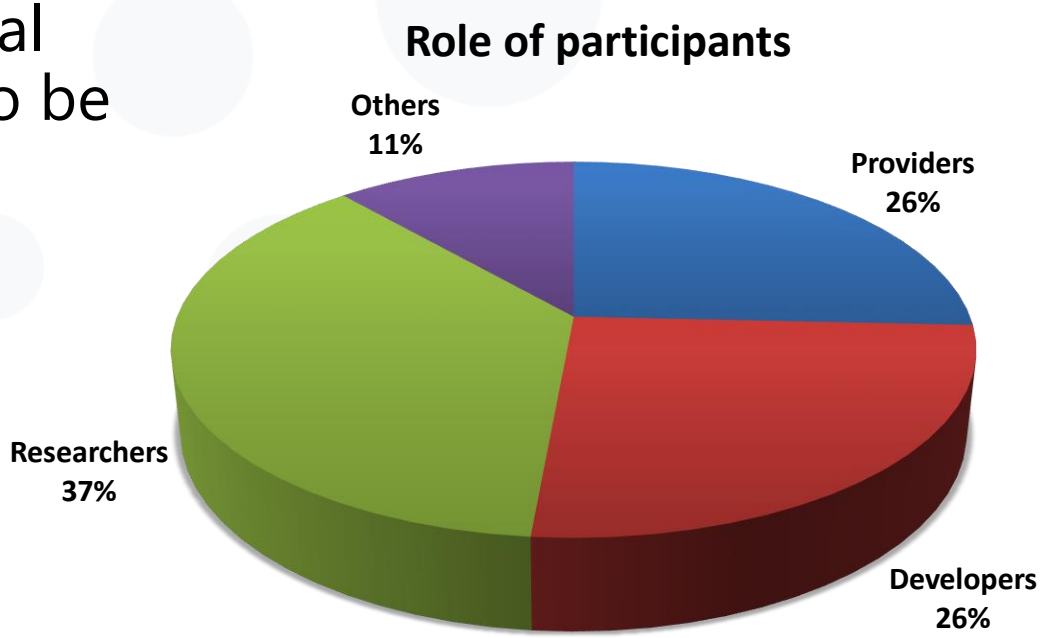
Survey statistics

- Full responses: 15
- Incomplete responses: 20
- Total responses: 35
- 2 interviews

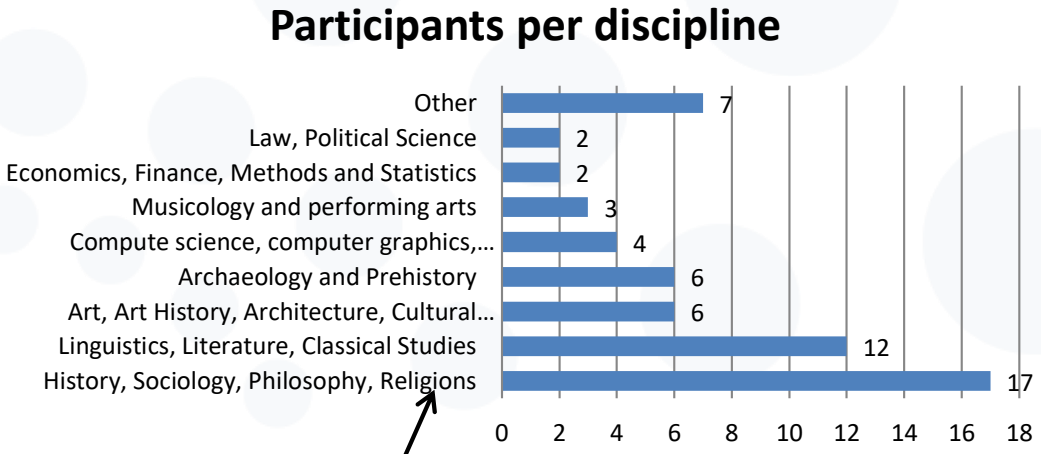


Who are participants?

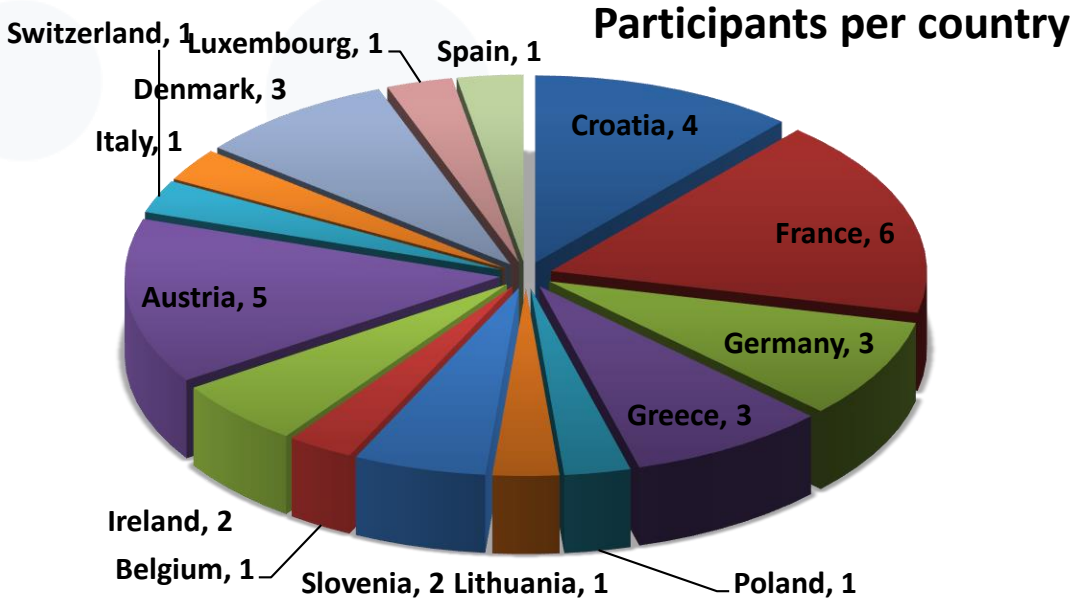
- A significant number of research scientists (52%)
- ‘Others’ -> two roles
- Better technical background to be expected



Who are participants?



'History'-> emphasis is on storage not computation



Experience with e-Infrastructure

- Only 5 positive reponses on “Do you know what e-Infrastructure is?”
  - Requires more effort to be put in dissemination
- The infrastructure services used:

Digital Repositories	Computational resources (e.g. computer cluster, grid, or cloud))	Authorization / Authentication services	Web-oriented services
6	4	4	4
33%	22%	22%	22%

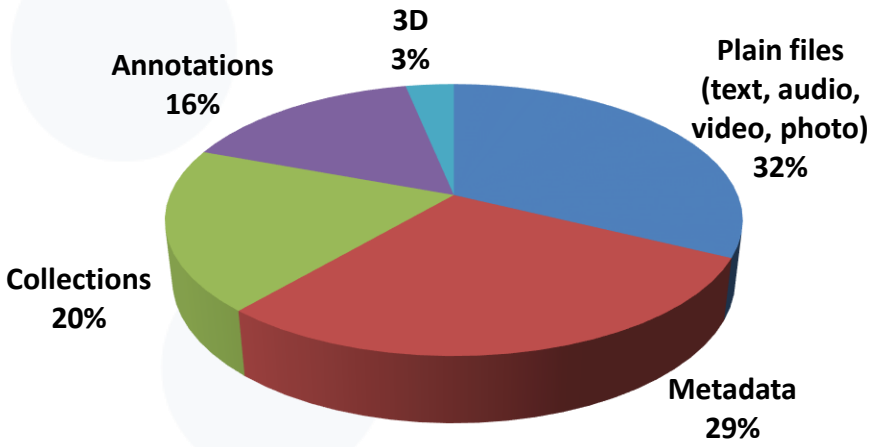
- Resource they have (or wish) access to
  - Mostly repositories and digital archives (55%)

Authentificatin and Authorization

- 12 (34.29%) of participants is aware of Identity Federations
- 7 (20.00%) of participants institutions are part of national Identity Federations
- Main barriers/challenges to join an identity federation
  - Lack of trust (8.57%)
  - Lack of Manpower (11.43%)
- DARIAH level
  - DARIAH IdP based on SAML, member of Edugain

Digital A&H assets and data management

Data types used



Data generated per year (in GB)	Number of responses
1	1
360	1
500	1
1000	4

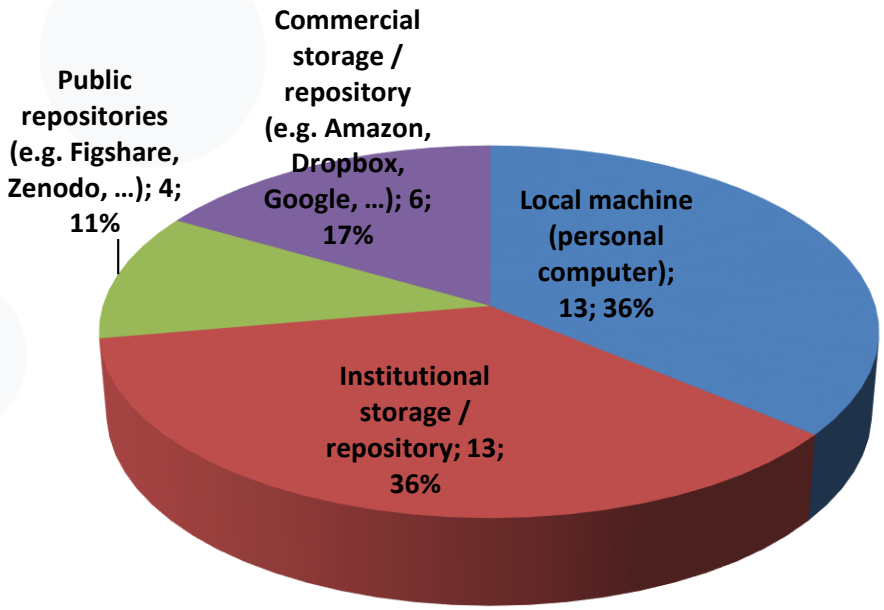
Digital A&H assets and data management

Storing and sharing

- **37%** use local machine and **37%** institutional storage/repositories
- Only **11%** share their research data

Data generated per year (in GB)	Number of responses
1	1
360	1
500	1
1000	4

Where data are stored



Services, applications and tools

Services

Gallica
Virtual library of the libraries France

Fulir
Institutional repository of scientific production

Google, Google drive, Dropbox

histoGraph
graph based visualisation to explore the collectios

Frameworks

Koha
Library management system

Python

HAL-SHS
archiving and dissemination of scientific literature

Tools

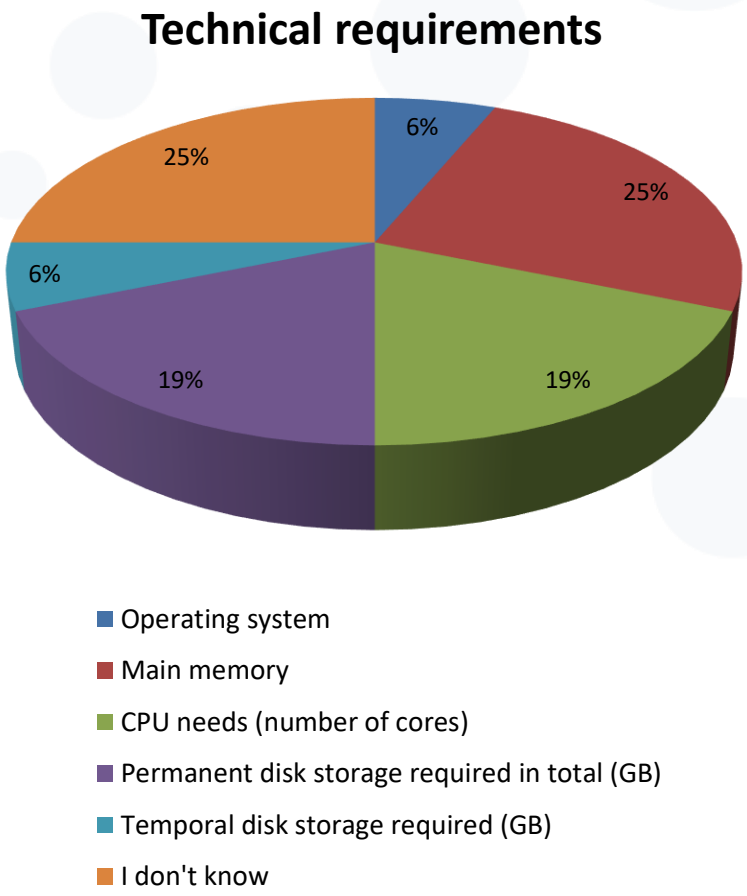
ArcGis
GIS dana analysis and visualisatin

MeshLab
Processing 3d sampled data

Topic Modeling
Generating word embedding vectors

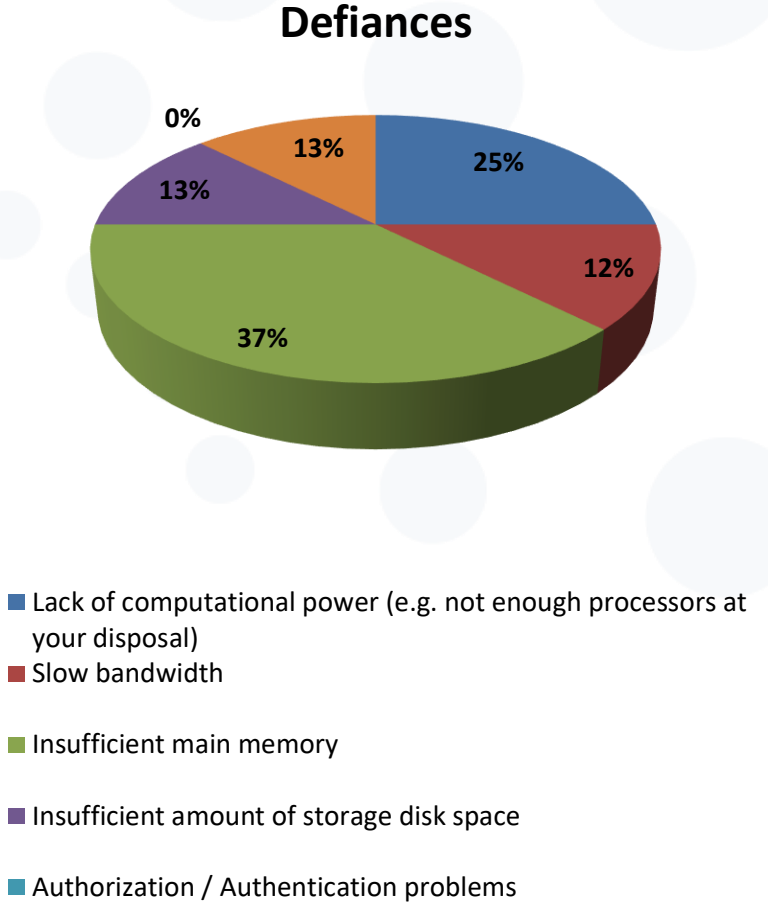
Photoscan
scene 3D reconstruction software

Infrastructure requirements



- Data intensive
- Two mayor issues
  - Computational power
  - Main memory
- Other needs:
  - GPU (TopicModelling) -> GPGPU cloud access

Infrastructure requirements



- Data intensive
- Two mayor issues
  - Computational power
  - Main memory
- Other needs:
  - GPU (TopicModelling) -> GPGPU cloud access

Examples

histoGraph

- Multiple independent jobs
- Requirements:
  - Main memory/storage
  - CPU power
- Defiance:
  - Main memory

http://histograph.eu

Topic Modeling

- MPI-based application
- Requirements:
  - CPU power
  - Main memory/storage
- Defiance:
  - Computational power
  - Main memory
- Requirements: GPGPU

Survey conclusion - requirements

Resources	• Processing power • Storage/main memory • Other computational resources – clusters/GPU
Data sharing and accessing	• Local or institutional repositories • Repositories and digital archives -> portals/gateways • Low storage capacities
AAI	• DARIAH IdP • Local access (username/password) • Problem with accessing EGI resources
Support and training	• Benefits of Cloud services

DARIAH roadmap in EGI

Per use-case approach is required!
Sustainable long-tail user support and training is obligatory

Resources	• Provide EGI Grid and Cloud resources (virtualization) • Object storage for archiving • EGI GPU cloud access (?)
Data sharing and accessing	• EGI long-tail of science platform • Science portals/gateways for accessing EGI services • gLibrary, CDSTAR, WS-PGRADE, other
AAI	• DARIAH IdP – interoperability with EGI VOMS • Virtual organization -> vo.dariah.eu (robot certificates • Collaboration with EGI AAI
Support and training	• Promote the usage of EGI training framework • Provide user support (on EGI level) • Demonstrate successful stories on DARIAH related events

Thank you for your attention.

Questions?

www.egi.eu