Of Users and Infrastructures

CLARIN Café
25 July 2024
11:00 - 12:30 CEST

Organisers

This CLARIN Café is organised by **Maria Gavriilidou**, Chair of the User Involvement Committee

Your CLARIN hosts: **Francesca Frontini, CLARIN BoD, and Iulianna van der Lek, Training & Education Officer**

Coordination and Technical support: **Thalassia Kontino**

The event is recorded for further dissemination purposes.

Questions and comments? Put them in the chat box.

Schedule

11:00-11:10 Opening and CLARIN 101

11:10 - 11:50 Of Users and Infrastructures

11:50 - 12:30 Discussion

Of Users and Infrastructures

Kanella Pouli

Senior Research Associate, ILSP/Athena RC

CLARIN

- **What do users want?** (Main research findings: web survey & focus groups)

  "Digital Landscape" (The emerging landscape of digital work practices in the Humanities in the context of the European projects DARIAH and CLARIN, 2022-2024)

- **How to find out?** (Research methodology tips: Dos and Don'ts)

- **What do we offer?** (Do we meet the users' needs?)

- **Where do we go from here?** (Less data, more tools?)

What do users want?

"…there is a clear need for a sophisticated infrastructure model, such as a Digital Humanities (DH) center that is easily accessible, with a user-friendly interface and intelligent search engines.

This model should provide a variety of rich data and tools from a single access point, including bibliographic resources and training material to support the use of the provided resources and services. [...] it is crucial for the infrastructure to act as a hub within a network, interconnected with other infrastructures and access points.

This connectivity will enable users to find the necessary resources for their research or work and connect with other researchers. [..] this approach will save researchers time, facilitate the aggregation and reuse of data and services."

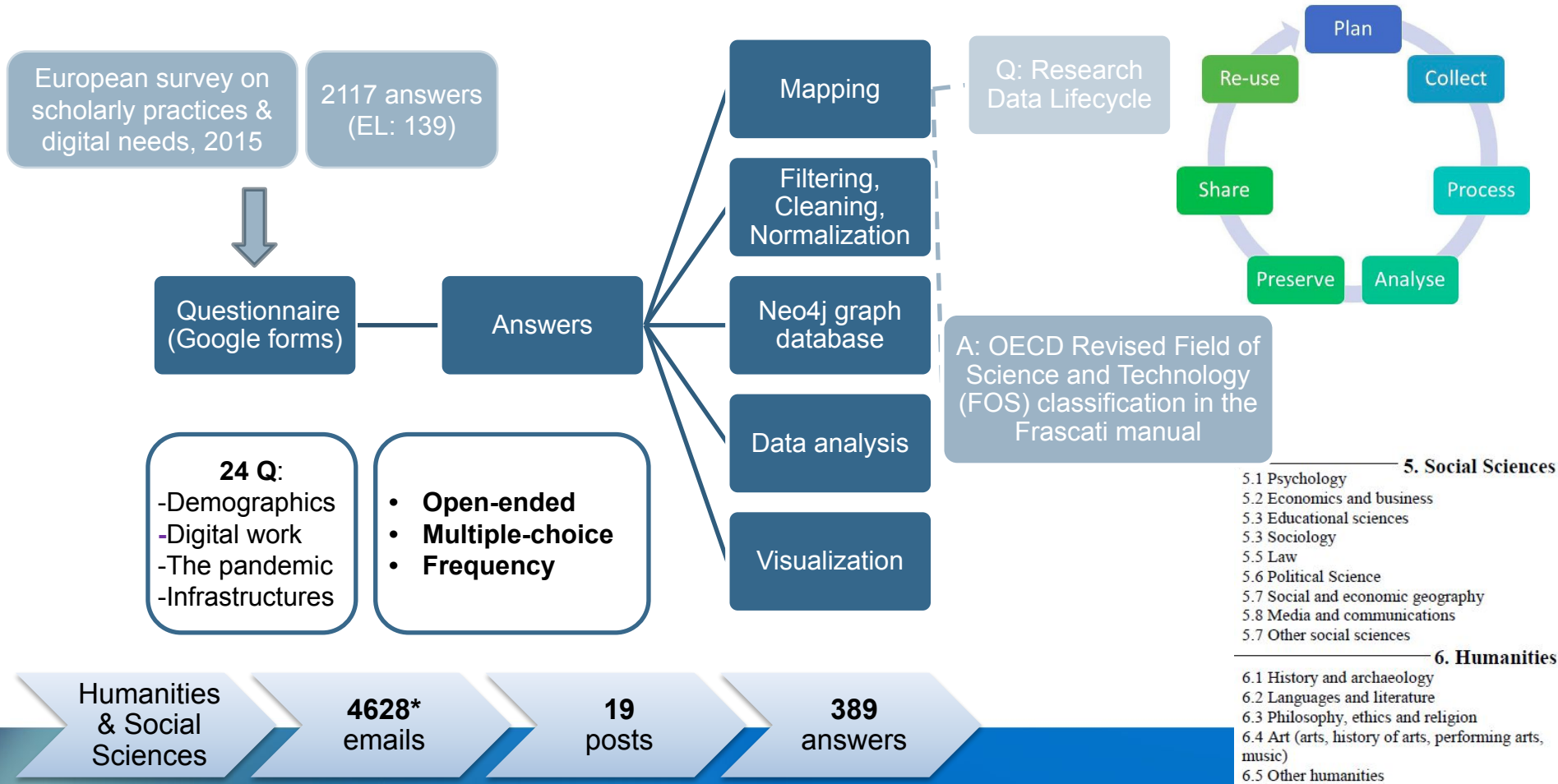*"More user-friendly search filters* would be helpful since even small mistakes can lead to *irrelevant results*."

"Visualization of results in a user-friendly manner, without unnecessary information. *Infographics* explaining the infrastructure's features and navigation."

"Currently, *digital services are spread across various websites*. It would enhance my productivity to have all these tools *centralized* in one place"

"Additional tutorials would be beneficial for those outside the field who are interested in joining, helping them integrate more smoothly. Moreover, since *many manuals are difficult to read*, incorporating video tutorials would be more effective."

European survey on scholarly practices & digital needs, 2015

2117 answers (EL: 139)

Questionnaire (Google forms)

Answers

Mapping

Filtering, Cleaning, Normalization

Neo4j graph database

Data analysis

Visualization

Q: Research Data Lifecycle

A: OECD Revised Field of Science and Technology (FOS) classification in the Frascati manual

Plan

Collect

Process

Analyse

Preserve

Share

Re-use

24 Q:
- Demographics
- Digital work
- The pandemic
- Infrastructures

- **Open-ended**
- **Multiple-choice**
- **Frequency**

5. Social Sciences
5.1 Psychology
5.2 Economics and business
5.3 Educational sciences
5.3 Sociology
5.5 Law
5.6 Political Science
5.7 Social and economic geography
5.8 Media and communications
5.7 Other social sciences

6. Humanities
6.1 History and archaeology
6.2 Languages and literature
6.3 Philosophy, ethics and religion
6.4 Art (arts, history of arts, performing arts, music)
6.5 Other humanities

Humanities & Social Sciences

4628* emails

19 posts

389 answers

Neo4j database

values (Q & A) → **Schema**: **29** classes & **31** semantic correlations → **Instances:** **1201** (classes) & **30808** (semantic correlations) → **Cypher**: **780** queries **Output**: PNG, CSV


```
topio$ MATCH (n:GeneralFrascati) RETURN n LIMIT 25
```

"n"

{"label":"Ανθρωπιστικές Επιστήμες και Τέχνες"}   Humanities and Arts

{"label":"Επιστήμες Μηχανικού και Τεχνολογία"}   Engineering & Technology

{"label":"Κοινωνικές Επιστήμες"}   Social Sciences

{"label":"Φυσικές Επιστήμες"}   Natural Sciences

```
topio$ MATCH p=()-[r:has_generalFrascati]→() RETURN p LIMIT 25
```

*(26)   Person(25)   GeneralFrascati(1)
*(25)   has_generalFrascati(25)

Q7. Indicate which of the following digital tools and services you use in your research or work

	All	Hum&A	SS	Stage
The **internet**,	96.90	97.70	98.30	acquire
Digital communication platforms (e.g. email, social networks, etc.),	89.40	89.90	93.90	acquire
Online/digital dictionaries,	70.00	74.80	65.20	acquire
Video conferencing platforms (e.g. Zoom, Skype, Webex)	86.80	88.90	92.20	acquire
Word processing software (e.g. Word),	88.37	90.27	87.83	process
Databases,	71.83	72.48	60.87	process
Image processing software,	45.22	46.31	44.35	process
Language technology tools for text,	39.79	41.61	32.17	process
Video processing software,	27.65	25.84	37.39	process
Audio processing software,	24.55	24.16	30.43	process
Geographic Database Software (GIS),	16.02	17.45	17.39	process
Design editing software,	10.59	10.74	7.83	process
Other	6.46	5.03	10.43	process
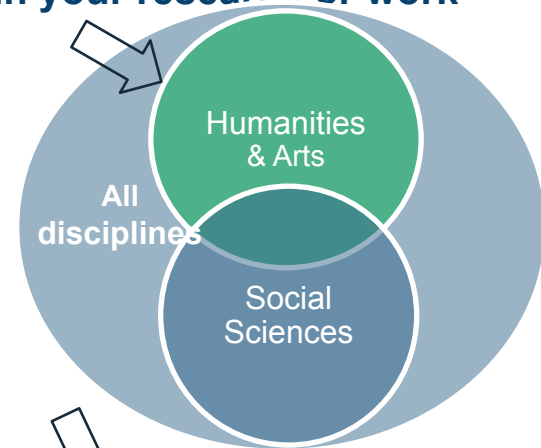
Humanities & Arts

All disciplines

Social Sciences

match (p:Person)-[r3:uses_normToolService]->(t: normalizedToolService{label:'image processing software'}) match (p:Person)-[r2:has_generalFrascati]->(f:GeneralFrascati {label:'Humanities and Arts'}) return count(distinct p)

How many respondents working in **Humanities and Arts** use **image processing software** in their work or research?

Total number of queries for Q7: 117

Queries: examples

Q7: Indicate which of the following digital tools and services you use in your research or work (multiple choice)	Cypher #Number of queries for Q7 ☐ 117	p
How many participants answered question 7?	match (p:Person)-[r3:uses_toolService]->(t:ToolService) return count(distinct p)	387
How many respondents use **image processing software** in their work or research?	match (p:Person)-[r3:uses_normToolService]->(t:normalizedToolService{label:'**image processing software**'}) return count(distinct p)	175
How many respondents working in **Humanities and Arts** use **image processing software** in their work or research?	match (p:Person)-[r3:uses_normToolService]->(t:normalizedToolService{label:'**image processing software**'}) match (p:Person)-[r2:has_generalFrascati]->(f:GeneralFrascati {label:'**Humanities and Arts**'}) return count(distinct p)	138
How many respondents working in **Social Sciences** use **image processing software** in their work or research?	match (p:Person)-[r3:uses_normToolService]->(t:normalizedToolService{label:'**image processing software**'}) match (p:Person)-[r2:has_generalFrascati]->(f:GeneralFrascati {label:'**Social Sciences**'}) return count(distinct p)	51
How many respondents working in **Humanities and Arts** use **word processing software** in their work or research?	match (p:Person)-[r3:uses_normToolService]->(t:normalizedToolService{label:'**word processing software**'}) match (p:Person)-[r2:has_generalFrascati]->(f:GeneralFrascati {label:'**Humanities and Arts**'}) return count(distinct p)	269

18.11.2022
Information
Sciences

chatGPT

5.12.2022
Literature &
Linguistics

18.01.2023
Spatial data
Sciences

30.11.2022
Folklore
Studies &
Anthropology

14.12.2022
Social &
Political
Sciences

20.01.2023
Archaeology &
History

purpose

participants

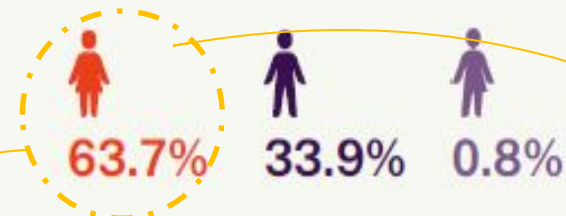moderator

discussion
guide

recoding &
analysis

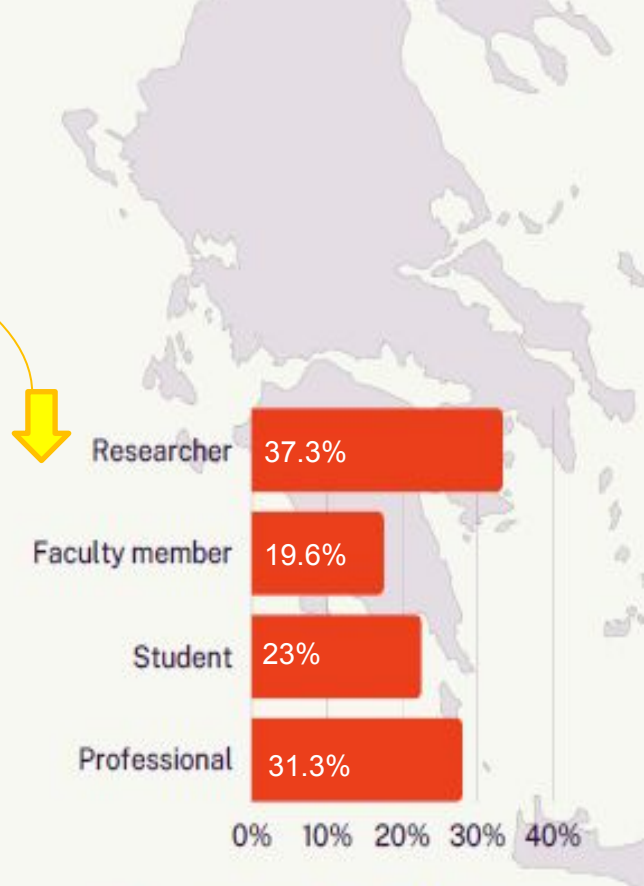Web survey: October 2022 – January2023
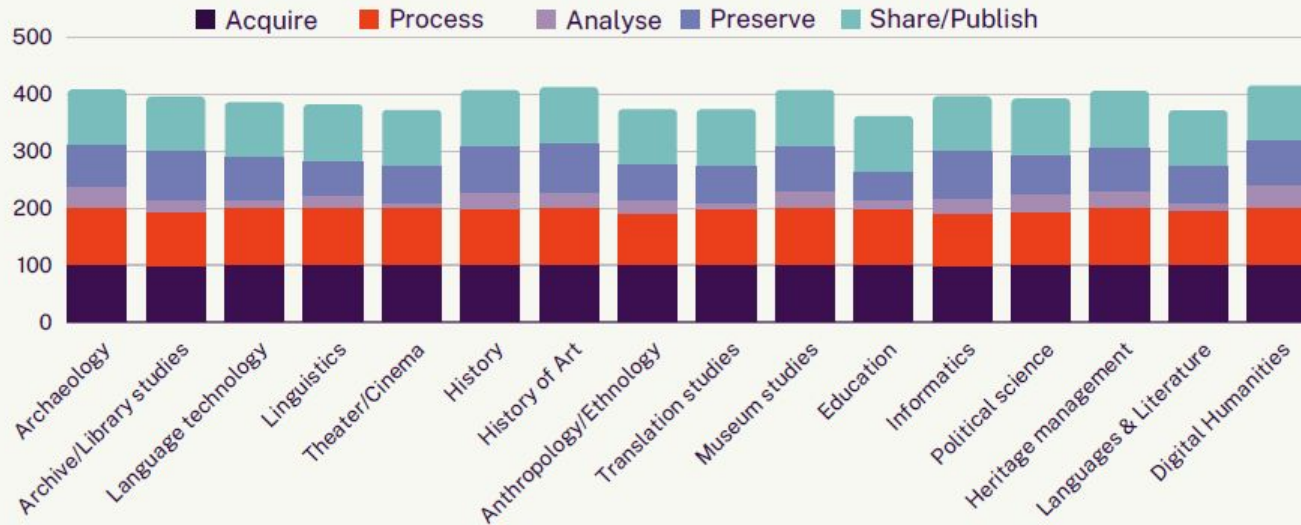
SAMPLE DEMOGRAPHICS

387
Respondents

63.7%   33.9%   0.8%

Age of respondents. N=385

>10 years
41%

<1 year
9.1%

1-3 years
20.8%

3-10 years
29.1%

Years using digital methods. N=385

Researcher — 37.3%
Faculty member — 19.6%
Student — 23%
Professional — 31.3%

Professional status. N=383

Use of digital tools/services by discipline & research data lifecycle

Disciplines and purpose are multiple-answer questions

Frequency of activities during research/work
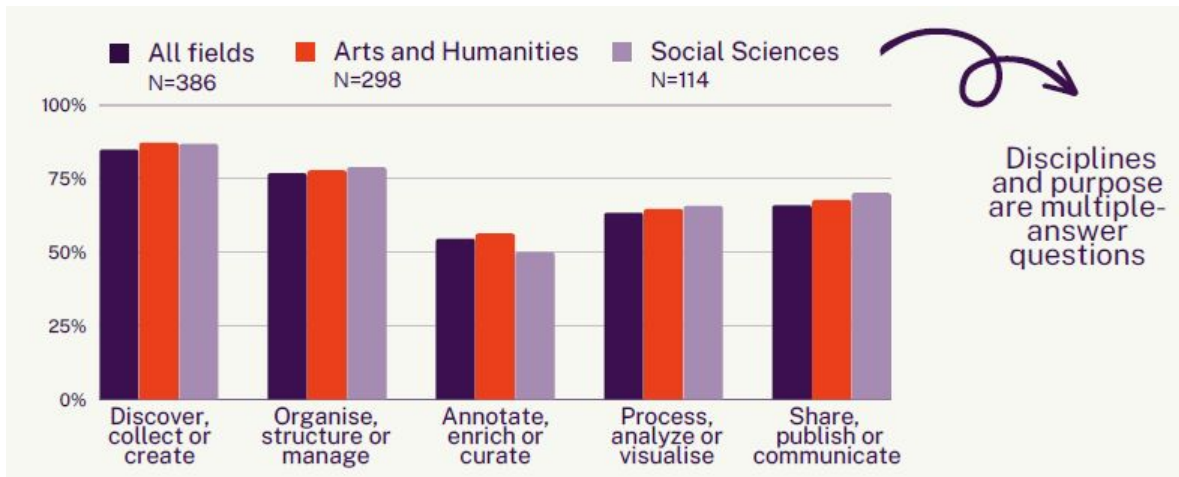
Legend: Never | Rarely | Often | Very often

Columns: All disciplines | Humanities and Arts | Social Sciences

Activities:
- Visiting (physically present) historical archives, collections or museums
- Seeking information or advice from archivists, librarians or curators
- Seeking access to digitised or digital resources
- I use a standardised catalogue or thesaurus of terms to organise my research material
- I use my own list or thesaurus to organize my research material
- I use a bibliographic management application (e.g. Zotero, etc.) to manage my references
- I use social media or discussion forums to communicate with other researchers
- I use a general-purpose infrastructure (Zenodo, github), to search for research data
- I use a special-purpose infrastructure (e.g. CLARIN:EL, DYAS, APOLONIS) to search for research data
- I use language processing tools to analyse my data
- I use quantitative tools to process my data (Excel, spreadsheet)

IMPACT OF THE PANDEMIC

21.5%

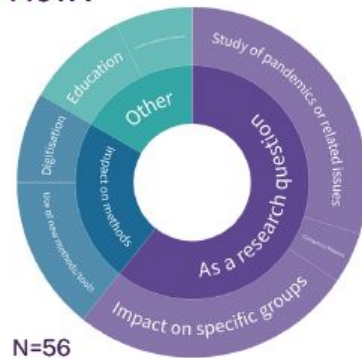Modified their research questions because of the pandemic

N=381

19.8%

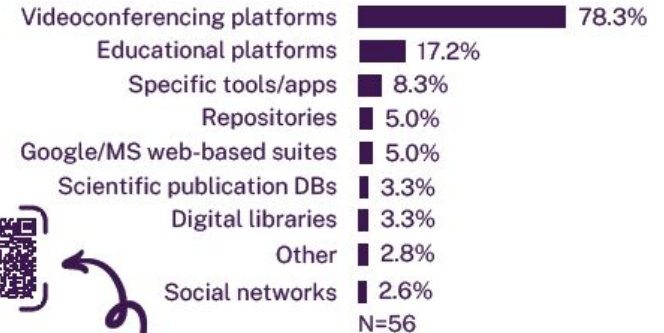Incorporated the pandemic as a parameter/part of their research

N=379

60.9%

Acquainted themselves with new digital tools/services due to the pandemic

N=381

How?

Change in research questions	57.4%
Change in data collection	32.8%
Change in data availability	11.5%
Change in the means/tools	13.1%
Change in the way of sharing	9.8%
Delay/suspend of research	11.5%

N=61

How?

Study of pandemics or related issues

Education

Other

Digitisation

Use of new methods/tools

Impact on methods

As a research question

Impact on specific groups

N=56

Which?

Videoconferencing platforms	78.3%
Educational platforms	17.2%
Specific tools/apps	8.3%
Repositories	5.0%
Google/MS web-based suites	5.0%
Scientific publication DBs	3.3%
Digital libraries	3.3%
Other	2.8%
Social networks	2.6%

N=56

Scan me

I. What services would you like a research infrastructure to offer? How would these services enhance your work conditions and outcomes?
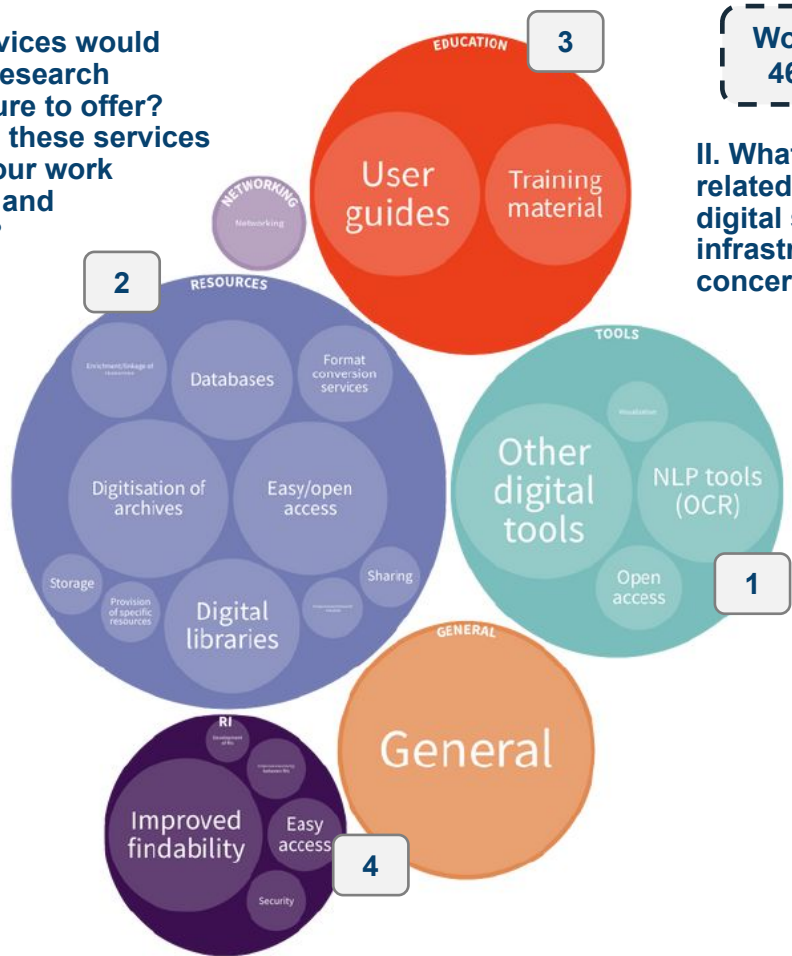
II. What other issues related to the use of digital services and infrastructures are you concerned about?

Women 46-55

What do these users want?

"I do not know" was the answer provided by 8 people

Desired digital services in research infrastructures [N=110, **29.3%**]

Concerns about digital services and infrastructures [N=87, **18.4%**]

New/ additional content

TOOLS	query	Results VLO	CLARIN:EL
Various tools (e.g. APIs)	tool*	13,379	49
NLP tools (OCR)	Nlp tool* OCR	190 734	31 0
Visualization tools	visuali?ation	55	0
Format Conversion tools	converter*	2,377	1

16,735

AVAILABILITY
- Open access
- Protection of personal data
- Copyright restrictions

☑ Ⓟ Public (328647)
☑ Ⓐ Academic (2197)
☑ Ⓡ Restricted for individual (55615)
☑ ⓘ Unspecified (306470)

INTEROPERABILITY

DATA	query	Results VLO	CLARIN:EL
Digitization of existing archives/data	archive*	119,579	44
Bibliographic sources/ Library content/ Publications	bibliiographi* librar* publication*	5,918 95,023 74,491	8 37
Databases	database*	11,329	53
Specific LRs	Social media	457	58

175,432

Training & Other issues

Yes 38.6%

[N=251]

No 61.4%

5-9 standards 6.2%

4 standards 7.2%

3 standards 15.5%

1 standard 44.3%

2 standards 26.8%

[N=97]
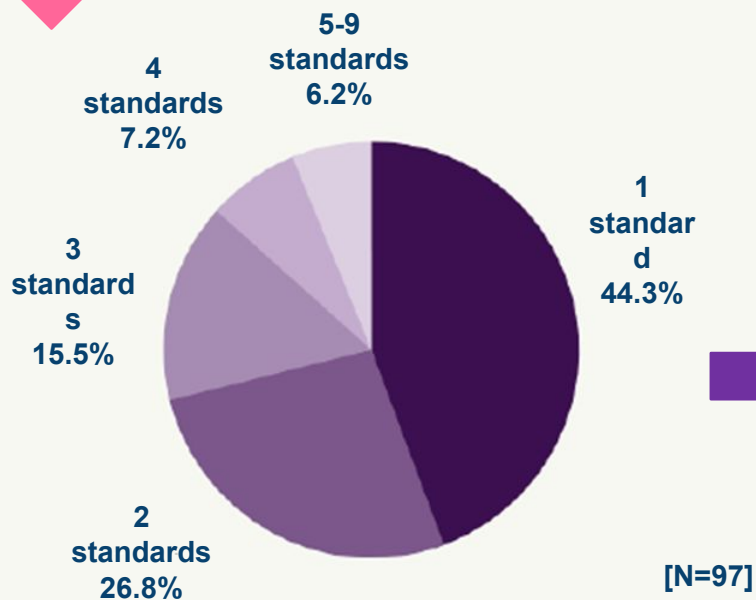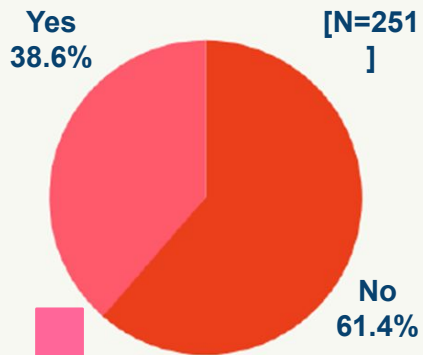
Standard		Humanities & Arts	Social Sciences
Dublin Core		21.8%	19.2%
Europeana Data Model	EDM	14%	9,6%
Text Encoding Initiative	TEI	13%	2.7%
Encoded Archival Description	EAD	13%	2.7%
CIDOC Conceptual Reference Model	CIDOC CRM	10.9%	12.3%
Machine Readable Cataloguing	MARC	6.2%	8.2%
META-SHARE metadata model		4.7%	1.4%
Data Catalog Vocabulary	DCAT	2.1%	-
Other		1.6%	2.7%
Component Metadata Initiative	CMDI	1.6%	-

"…there is a clear need for a sophisticated infrastructure model, such as a Digital Humanities (DH) center that is **1. easily accessible**, with **2.** a **user-friendly interface** and **3**. **intelligent search engines**.

This model should provide **4.** a **variety of rich data and tools** from a single access point, including **5. bibliographic resources** and **6. training material** to support the use of the provided resources and services. [...]

it is crucial for the infrastructure to act as a hub within a network, interconnected with other infrastructures and access points.

This connectivity will enable users to find the necessary resources for their research or work and connect with other researchers. [..]

this approach will **7. save researchers time**, facilitate the aggregation and reuse of data and services.

Where do we go from here?

- **2. Simplify procedures & redesign documentation using infographics instead of lengthy texts**
- **1. Ensure infrastructures are accessible to people with disabilities**
- **3. Enhance search engines with features like voice search**
- **4.1 Add new digital content, such as memes, machinima, and geospatial data**
- **5. Collaborate with other organizations (e.g., libraries),**
- **4.2. Promote & curate new content**
- **6. Develop training courses & diverse tutorials/webinars for all age groups**
- **4.3 Integrate tools for older language forms & specific media types (e.g., manuscripts)**
- **4.4 Include visualization tools**
- **7. Implement methods for automatic metadata completion**

Some observations 1/2

- **Questionnaire Engagement**: Find ways to engage more people in answering questionnaires and run smaller, targeted surveys for different age groups.
- **Question Design**: Carefully consider the placement and order of questions to avoid biases. Balance the topics and avoid leaving important questions until the end, as respondents may get tired and skip them. Keep the questionnaire concise.
- **Question Clarity**: Ensure clear boundaries between questions to prevent confusion and overlap.
- **Controlled Vocabularies**: Use controlled vocabularies or restricted answers to minimize the need for normalization later.
- **Common Language**: Establish a common language with users (digital tools vs digital platforms)

Some observations 2/2

- **Workflow Planning**: Plan the workflow in advance (the creation of the Neo4j database and familiarization with Cypher was time-consuming).
- **Focus Groups**: Conduct focus groups before creating the questionnaire to identify and address potential issues.
- **In-Person Meetings**: Whenever possible, hold all meetings in person rather than virtually.
- **Results Analysis**: Combine web survey results (quantitative analysis) with focus group discussions (qualitative analysis) and compare them to previous related work.
- **Satisfaction Levels**: Examine the level of satisfaction with digital tools or services in infrastructures.
- **Result Interpretation**: Be cautious when interpreting results - consider factors like age, digital literacy, and scientific field.

For more info:

- Ilvanidou, M., Dritsou, V., Gavriilidou, M., Pouli, K., Tzedopoulos, Y., & Souyioultzoglou, I. (2023). The "Digital Landscape in Greece" Web Survey. DARIAH Annual Event 2023: Cultural Heritage Data as Humanities Research Data?, Budapest, Hungary. Zenodo. https://doi.org/10.5281/zenodo.7973681

- Ilvanidou, M., Pouli, K., Souyioultzoglou, I., & Tzedopoulos, Y. (2023). Trends in digital methods in the Humanities and Social Sciences in Greece. Zenodo. https://doi.org/10.5281/zenodo.10797453

Go to menti.com | use code: 6534 1557

https://www.menti.com/al3n4ssvuhzp

Thank you!
~Have a great summer~

Learn more

- Newsflash (link)

- Tour de CLARIN (link)

- Impact Stories (link)

- CLARIN Learning Hub (link)

- CLARIN Value Proposition (link)

Next Café

Stay tuned: https://www.clarin.eu/content/clarin-cafe

Share your **#clarincafe** impressions with **@CLARINERIC**

Upcoming Events:

- **20 September:** Skills4EOSC Fair-by-Design Methodology in Training Development (link)

- **15 - 17 October:** CLARIN Annual Conference (hybrid) (link)