

# DELIVERABLE 4.1

## SKILLS GAP ANALYSIS

Authors: Lisanne M. van Rossum, Artjoms Šeļa  
Date: February 2022



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

Project Acronym: CLS INFRA

Project Full Title: Computational Literary Studies Infrastructure

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Grant Agreement No.: 101004984

### Deliverable/Document Information

Document Title: Deliverable 4.1 Skills gap analysis

Authors: Lisanne van Rossum, Artjoms Šeļa

Dissemination Level: Private

### Document History

Version/Date	Changes/Approval	Author/Approved by
V.0.1	Ciara L. Murphy	Artjoms, Lisanne
V.0.2	Vera Maria Charvat	Artjoms, Lisanne

## INDEX

<b>i. Objective and scope</b>	<b>4</b>
<b>ii. Executive summary</b>	<b>5</b>
<b>PART I. SKILLS MATRIX</b>	<b>8</b>
<b>1.1. Setting up a skill matrix</b>	<b>8</b>
1.1.1. Cycles or self-contained areas?	9
1.1.2. Understanding of cycles	9
<b>1.2. Cycle / skill taxonomy</b>	<b>10</b>
1.2.1. Theory and research set-up	12
1.2.2. Collection	12
1.2.3. Analysis	13
1.2.4. Delivery	13
<b>PART II. MAPPING SUPPLY LANDSCAPE</b>	<b>15</b>
<b>2.1. Annotating supply in schools and courses</b>	<b>15</b>
<b>2.2. Annotation procedure</b>	<b>16</b>
2.3. Supply overview	18
<b>PART III. MAPPING DEMAND LANDSCAPE</b>	<b>21</b>
<b>3.1. Surveying demand in research community</b>	<b>21</b>
3.1.1. Target audience	21
3.1.2. Objectives	21
3.1.3. Dissemination	22
<b>3.2. Survey design</b>	<b>25</b>
3.2.1. Introduction	25
3.2.2. Opening questions	26
3.2.3. Scale statements	27
3.2.4. Closing questions	30
<b>PART IV. SURVEY RESULTS</b>	<b>33</b>
<b>4.1 Participants</b>	<b>33</b>
4.1.1. Gender	33
4.1.2. Career stage	34
4.1.3. Academic background	35
4.1.4. Geographical span	37
4.1.5. Involvement in CLS and computational experience	38
<b>4.2. Skills</b>	<b>40</b>
4.2.1. Scores	40
	2

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

4.2.2. Score variability	42
4.2.3. What do literary scholars want?	44
<b>4.3. Open questions</b>	<b>45</b>
4.3.1. Opportunities and challenges for CLS training	45
4.3.2. Response distribution per career stage	49
4.3.3. Missed areas in survey	52
<b>PART V. GAP ANALYSIS</b>	<b>54</b>
<b>5.1. Methodological steps</b>	<b>54</b>
<b>5.2. Final heatmap</b>	<b>56</b>
<b>5.3. Conclusion: where approaches intersect</b>	<b>58</b>
<b>PART VI. SUPPLEMENT</b>	<b>60</b>
<b>6.0. Materials</b>	<b>60</b>
<b>6.1. References</b>	<b>60</b>
<b>6.2. Annotation sources</b>	<b>62</b>
<b>6.3. Survey response data normalization and encoding</b>	<b>63</b>
6.3.1. Career stage	63
6.3.2. Academic background	63
6.3.3. Geographical span	64
6.3.4. Training challenges / opportunities	64
6.3.5. Missed areas in survey	66
<b>6.4. DARIAH survey consent form</b>	<b>68</b>
<b>6.5. Entropy and demographics</b>	<b>71</b>
<b>6.6. Distribution of scores per skill</b>	<b>72</b>
<b>6.7. Abbreviation list</b>	<b>73</b>

## i. Objective and scope

This report was produced for Task 4.1 of Work Package 4 of the Computational Literary Studies Infrastructure research project (“CLS INFRA”). [CLS INFRA](#) is a European Commission-funded project which aims to create unified and easy access to a wide range of European and national infrastructures for the Computational Literary Studies (“CLS”) community. This includes engaging literary researchers with the development of tools and services for computational literary studies through training activities, an ambition which is formalized in Work Package 4.

Accordingly, Work Package 4 will organize a series of international Training Schools (Task 4.3) and produce corresponding training materials that will be hosted in close collaboration with the [DARIAH campus](#) platform (Task 4.2). Within this infrastructural context, it is the objective of Task 4.1 to produce a gap analysis to strategically inform Work Package 4’s supplementation initiatives of the current training offer in computational methods for literary research.

The gap analysis presented in this 4.1 report is based on the creation of a skills matrix that can be used as a framework to map existing resources in CLS training in the form of materials, events, and opportunities. The resulting index of supply is then juxtaposed with a survey of the training demands in skills for computational research by scholars with various disciplinary backgrounds, nationalities, and experience levels.

This research identifies areas of over- or undersaturation in skills for computational literary scholarship through a dual approach: it combines a quantitative heatmap that closely maps to the skill matrix produced with qualitative accounts to amplify the voices of members of the global CLS research community.

## ii. Executive summary

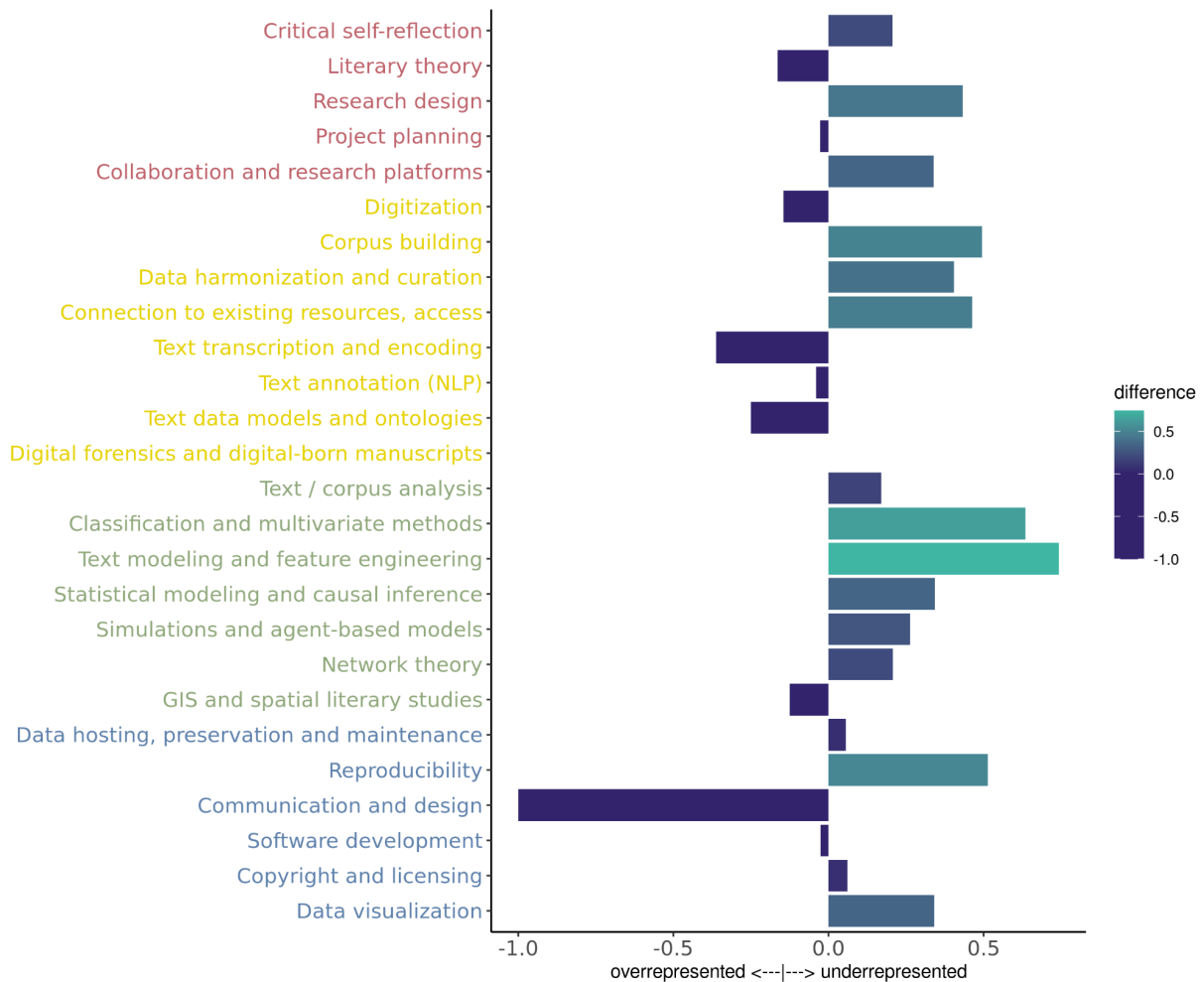
The main task of this deliverable was to explore current gaps in teaching of research skills for computational literary studies. This knowledge would then inform CLS INFRA's own approach to training schools, but also would chart the territory for broader communities of practitioners and scholars. While the landscape of humanities computing skills became a well-traversed and studied area in the recent decades, we offer a contemporary look - through a lens of literary scholarship and well-defined object of study - at the paths both taken and underexplored.

We approached the task primarily in a quantitative manner to be able to uniformly bring major tendencies in teaching and skill demand to a common denominator. This required an explicit mapping of 1) existing teaching practices (“**supply**”) and 2) opinions of the practitioner community (“**demand**”) to a single grid of skills, where it would be possible to compare both parts and identify the gaps directly.

Skills in a grid were derived from four broadly defined stages in a research cycle: 1) **Theory and research setup**, 2) **Collection**, 3) **Analysis**, 4) **Delivery**. In defining skills we aimed at the middle level of abstraction: skills do not relate to particular implementations, platforms or software, but embody general practices and activities, while remaining useful and distinct (e.g. corpus building, classification, statistical modeling). To understand supply we have manually annotated current offers in a sample of European university courses in Digital Humanities and summer school workshops, tying them to the grid. At the same time we set up an online survey to ask the community to evaluate each skill from the grid based on its perceived future prospects in the field and teaching (1-5 scale response, 118 participants).

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Figure 1. General differences between demand (survey results) and supply (schools and courses annotations).



After value normalization from both sides of demand and supply, it was possible to identify general areas of teaching saturation and undersaturation from the perspective of computational literary research. The coarse averaged picture of demand/supply discrepancy (Figure 1) shows that some areas of **Analysis** and **Collection** can be the focus of training in future, especially advanced text modeling, classification, practical corpus building and access frameworks to existing collections. Across **Research setup** the focus is on research design principles, while **Delivery** shows underrepresentation of knowledge on reproducibility. Further detailed analysis revealed some important tendencies that are hidden by the average trend: for example, summer schools on Digital Humanities do not tend to have statistics workshops, while the community demand for **statistical modeling and causal inference** is among 5 highest scoring. Overall, we see that collective opinion favors the practical research-oriented backbone of the grid: starting at

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

research design principles, through data preparation and enrichment, to text modeling and analysis to statistical inference, ending in a reproducible package.

The survey also offered a chance to observe the demographic structure of the CLS community. Most of the responses came from early career scholars (PhD candidates, assistant professors and postdocs), indicating a new generational wave within computational literary studies. While the gender of participants appears to be fairly balanced, there are more late career (professors) men than women and men are more represented in disciplines with technical backgrounds (computational linguistics, computer science). Self-reported involvement in CLS and experience in computational skills was likewise on average lower for women participants. There was an overall positive relationship between involvement in the field and proficiency in computational skills, which suggests that involvement in a computational field and corresponding methodological knowledge often are paired in current practice.

Open questions built into the survey showed further nuance in community lines of thinking about opportunities and pitfalls for training and schooling in CLS. Among the topics most frequently mentioned in the responses were a quantitative lack of (centralized) training, concern about the discipline's positioning in relation to its peers in arts and science, and the lagging behind in institutionalization of computational research skills in university curricula.

To a lesser degree, participants addressed the lack of resources to learn, such as time, money, or access. Several participants reported missing the opportunities to pursue a career in CLS. Others felt overwhelmed by the unorganized offer in CLS training and material, or described a qualitative lack in schooling beyond introductory modules.

Data set subdivision by career level suggested that participants' needs shift over the course of their career. Institutionalized training and learning skills are among the most encountered issues by (under)graduates and PhD students. A main concern of early career researchers seems to be (professional) opportunities in the field, while professors show increased interest in a solidified mission and research practice for CLS.

When asked about missed areas in the survey, participants suggested more focus on the heterogeneity of the current textual landscape in CLS (e.g. minority languages, translations, historic corpora) and the discipline's connections within and beyond academia, such as the job market. The reported core issues in CLS training overlap thematically with the suggestions made for improvement of this survey. It underlines how this research is not intended to close the gap between demand and supply in CLS training, but instead to take a step towards identifying it.



# PART I. SKILLS MATRIX

## SUMMARY

- PART I documents the establishment of a foundational skills matrix to take inventory of supply in CLS training.
- The skills matrix is loosely derived from the life cycle stages of a research project; 1) Theory and research; 2) Collection; 3) Analysis and 4) Delivery.
- These four units were subdivided into 26 smaller units termed “skills,” outlined and exemplified in this chapter.
- To provide future opportunities for a more complex and network-driven approach to our grid, each skill was mapped to one or more concepts in the [Taxonomy of Digital Research Activities in the Humanities](#) (TaDiRAH, Borek et al. 2016).

## 1.1. Setting up a skill matrix

Originally we started to derive the skill matrix from "scholarly primitives" (Blanke & Hedges 2013; Unsworth 2000) that were formulated for building a digital infrastructure for humanities that would be as discipline-independent as possible. Future work in creating activity taxonomy in digital research also followed the broad discipline-agnostic perspective (Borek et al. 2016). Similarly, surveys and reviews of the teaching practice, needs of the field and skill bases were done from the perspective of large infrastructural projects like CLARIN and DARIAH, focusing on the infrastructural role in mediation of user requirements and accessibility challenges (see Drude et al. 2016; Tasovac et al. 2018).

In the case of CLS INFRA this was not the approach we wanted to take: our infrastructure is focused on the clear object of inquiry and can be directly related to several disciplines (literary theory, computational linguistics, NLP, narratology, etc.). This is both a privilege, to be able to do a tunnel-vision analysis, and a disadvantage, since our results are not directly relatable to a broader Digital Humanities movement.

That is why we shifted our understanding of the initial “primitives” into a broadly defined cycle-based structure, while dramatically repurposing and transforming the initial units. Boundaries between steps in many realistic cases of mapping the supply/demand might not be clear-cut, but knowledge production is complex and looped on itself, so it was a necessary simplification.

The four steps are: **1) Theory and research setup, 2) Collection, 3) Analysis, 4) Delivery.** This report uses a consistent color-coding scheme for each of the research cycle phases: “red”, “yellow”, “green”, and “blue.”

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Each of the areas or steps was then split to smaller units. We call these units “skills” throughout this study. Skills stand for areas of research activities that are more general than particular methods, operations and their implementations (e.g. “stylometry with R”, “word association measures”, “lemmatization”), but at the same time are well-defined to be pointing at more or less distinct groups of tasks and subfields of research (e.g. “corpus enrichment”, “statistical modeling”). By managing the level of abstraction of the research sub-units we make sure they remain useful in mapping the skills gap both from the perspective of a practitioner answering the survey and an annotator making connections. The underlying structure of the grid and our thinking about it was informed mainly by computational research on (literary) text corpora, literary history and scholarly editing. It does not account for a broad range of activities connected to non-textual media (e.g. 3D modeling and VR, video and sound analysis, multimodality in general).

### 1.1.1. Cycles or self-contained areas?

We propose that the main four research phases can be conceptualized both as 1) linked steps in the research cycle and 2) independent areas of research (which would obviously have their own cycles). So our skill matrix can potentially account for both linear and nested structure, but should not be read as intrinsically one or another.

### 1.1.2. Understanding of cycles

First, we draft our understanding of the four areas in general and then list derived “skills” across those categories. We also provide a brief description of each of the skills and their corresponding TaDiRAH concepts / activities.

**Theory and research setup:** The first step in our simplified research cycle, it encloses the discovery process that happens before the active research phase (but it could happen anytime): engagement with theory, exploration of the knowledge gaps, critical understanding of a discipline, research design practices and preparations. Platforms and frameworks for making collaboration and crowdsourcing work are included here, too.

**Collection:** Often the most laborious, step two groups skills that are related to access to, reuse,- or creation of datasets, queries, data models. Digitization, text encoding, corpus enrichment and natural language processing are also included here.

**Analysis:** While there is no watershed between data collection and analysis, step three focuses on executing research and solving questions. It includes skills related to operationalizing questions and complex concepts, modeling relationships (chronology, author, genre), classification and clustering, statistics.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

**Delivery:** Delivery includes presentation of research outcomes in their many forms, ensuring access and reproducibility, delivering papers, publishing digital editions with HTML & CSS. Includes data visualization and software development practices.

## 1.2. Cycle / skill taxonomy

Table 1 summarizes the final skill taxonomy: it follows a simple two-level - one parent and many children - structure. We recognize the artificiality of boundaries between skills and the potential relationship that can be formed across the grid, but we explicitly aimed for the linear, unnested representation to streamline annotation and response collection. To provide future opportunities for a more complex and network-driven approach to our grid, we map each skill to one or more concepts in the Taxonomy of Digital Research Activities in the Humanities (“TaDiRAH,” Borek et al. 2016). Table 1 reflects this mapping, while explicit links to [vocabulary entities](#) are available in our [Gitlab repository](#). Note, that TaDiRAH connections were made across its original hierarchy, so that one skill can be simultaneously related to a parent and a child concepts in taxonomy.

Below we provide our expanded understanding of skills (mainly example-based) that also formed the basis of the survey descriptions.

*Table 1. Final taxonomy of skills grid with mapping to TaDiRAH concepts.*

Step	Skill	Code	TaDiRAH-related concepts
Theory and research setup	Critical self-reflection	T1	Contextualizing; Theorizing
	Literary theory	T2	Theorizing
	Research design	T3	Conceptualizing
	Project planning	T4	Organizing
	Collaboration and research platforms	T5	Collaborating
Collection	Digitization	C1	Data Recognition
	Corpus building	C2	Collecting
	Data harmonization and curation	C3	Data Cleansing
	Connection to existing resources	C4	Gathering

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

	Text transcription & encoding	C5	Transcribing; Encoding; Annotating
	Text annotation	C6	Natural Language Processing; Enriching
	Data models and ontologies	C7	Modeling; Querying
	Digital forensics and digital-born manuscripts	C8	Bit Stream Preservation; Discovering
Analysis	Text / corpus analysis	A1	Structural Analysis
	Classification and multivariate methods	A2	Relational Analysis
	Text modeling and feature engineering	A3	Machine Learning; Extracting
	Statistical modeling and causal inference	A4	Modeling; Reasoning
	Simulations, agent-based models	A5	Modeling
	Network analysis	A6	Network Analysis
	GIS and spatial literary studies	A7	Spatial Analysis
Delivery	Data hosting, preservation and maintenance	D1	Preserving; Sharing
	Reproducibility	D2	Replication
	Communication and design	D3	Communicating; Web Development; Writing
	Development of applications and tools	D4	Programming; Designing
	Copyright and licensing	D5	Sharing
	Data visualization	D6	Data Visualization

### 1.2.1. Theory and research set-up

1. **Critical self-reflection.** Nature and design of tools, methodological considerations, reflection of digital impact on literary studies, biases in visualizations, algorithms and data.
2. **Literary theory.** Awareness of current debates in the discipline, key concepts, connections between computational research and established theories (history, narrative, genre), exploration of related theories outside of literary studies (e.g. social sciences, linguistics).
3. **Research design.** E.g. being able to come up with a relevant CLS research question, choosing an appropriate scope, understanding general research directions (inductive or deductive, exploratory or hypothesis-testing)..
4. **Project planning:** E.g. being able to propose a research plan for a CLS project, acquire funding, manage a sustainable research life cycle.
5. **Collaboration and research platforms:** E.g. organize collaboration (Git, Open Science Framework), establish a crowdsourcing framework (tagging, transcription).

### 1.2.2. Collection

1. **Digitization.** Getting from images to machine-readable texts, finding solutions for recognition of printed and handwritten texts, and optimizing post-recognition error correction.
2. **Corpus building:** Building and organizing a collection of texts, using a crawler to get texts from the web, knowing how to store and document texts (referential tables, databases).
3. **Data harmonization and curation.** Cleaning or standardizing data with regular expressions, adding and curating metadata, familiarity with common metadata standards.
4. **Connection to existing resources.** E.g. drawing information from existing collections , connecting to libraries and other open collections using an API (“application programming interface”).
5. **Text transcription and encoding.** Encoding structural and linguistic elements of a text in machine-readable way, using mark-up conventions (TEI) to represent texts for scholarly editing or analysis.
6. **Text annotation.** (with special regards to the use of Natural Language Processing (“NLP”). Enriching text with semantic, morphological, or syntactical information, using NLP models to extract locations, individuals, and institutions, building domain-specific NLP models (e.g. for historical texts).

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

- 7. Text / data models and ontologies.** Creating and using data model standards and ontologies (formulating an underlying scheme for your data), querying of structured information and transformation between different data representations (Resource Description Framework (“RDF”)), establishing connections to the Linked Open Data cloud.
- 8. Digital forensics of digital-born manuscripts.** Preserving digital-born data (forensic copies), working with digital manuscript bitstreams, discovering remnants of the writing process in the bitstream (revision archaeology).

### 1.2.3. Analysis

- 1. Text / corpus analysis.** Comparing word or other linguistic unit usage between collections, making a concordance out of a collection, examining which words co-occur more frequently than others, using established metrics for text analysis, like readability scores or lexical richness.
- 2. Classification and multivariate methods.** Measuring similarity between a group of texts and clustering results, using supervised approaches, learning properties of textual classes and then predicting unseen data, evaluating unsupervised clustering or supervised predictions. Many would call it “machine learning”.
- 3. Text modeling and feature engineering.** Finding solutions to extract and analyze concepts that are not immediately recovered from unstructured data (like characters, plot progression curves, narrative events, etc.).
- 4. Statistical modeling and causal inference.** Building a regression to understand relationships between variables, making causal claims and testing them, treating possible confounds.
- 5. Simulations and agent-based models.** Building a historical model where simple agents are interacting over time, exploring the use of simulations for literary history.
- 6. Network analysis.** Visualizing networks and analyzing network structures; temporal networks that change over time; using network representation in literary studies.
- 7. GIS and spatial literary studies.** Using geo-referenced data to create and analyze maps, any research that has geographical dimension to it.

### 1.2.4. Delivery

- 1. Data hosting, curation, preservation, and maintenance.** E.g. where to host a dataset or corpus, obtaining a Digital Object Identifier (“DOI”) for further ease of access, ensuring preservation and long-term archiving.
- 2. Reproducibility.** Knowing the good practices of documentation of data and software, so others can reuse and reproduce work(flows).

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

3. **Communication and design.** Using writing tools or frameworks (including LaTeX and Pandoc) and reference managers, publishing static pages on Web, communicating research and results (blogs, social media).
4. **Software development.** While software development is itself a complex life cycle, here we understand it as primarily a delivery practice -- that makes resources out of your software (e.g. building a simple command-line tool, a Python script or library, an interactive Web application). Good-enough practices for development for researchers are covered here too.
5. **Copyright and licensing.** Awareness of benefits of licensing in publishing data or results, possible copyright pitfalls, knowing which Creative Commons license to use for your own data, and your responsibilities when reusing licensed data or software.
6. **Data visualization.** Producing publication-level images and graphs, making interactive data visualizations and deploying them on the web..

## PART II. MAPPING SUPPLY LANDSCAPE

### SUMMARY

- PART II presents how the skills grid established in PART I informed the approach to cataloging supply in CLS training.
- The mapping strategy relied on annotation of summer schools (n = 8) and on university courses sample (n = 20), between 2010 and 2021.
- This data set was largely derived from the [DARIAH registry of Digital Humanities courses](#) and supplemented by manual research.
- The two annotator's personal strategies were cross-checked and modified according to a mutually accorded "focus over nuance" principle.

### 2.1. Annotating supply in schools and courses

It is not easy to approach the question of supply for skills in computational literary studies. First of all, CLS as a field is poorly institutionalized, itself reflecting the fuzzy, decentralized existence of Digital Humanities. In the situation of low integration of computational skills into humanities and literary studies curricula, training schools and short-term workshops still remain the workhorses of skill transfer and teaching, sometimes serving as a first exposure to computational and digital skills for researchers. On the other hand, in the last decade education institutions increased the presence of organized Digital Humanities programs and elective courses, started dedicated centers and laboratories, with more institutionalized possibilities of learning for students on-site.

In indexing the current offer in CLS-related skills we wanted to bring the two perspectives - university courses and school workshops together. We relied on the existing [DARIAH registry of Digital Humanities courses](#) (that also include schools) to build up our sources on current teaching practices (Wissik 2020). We have pulled a list of 11 summer schools that we aimed to cover as full as possible, while following certain limits for a year of the school to not introduce an additional chronological confound to our results (year 2010 was set as a soft threshold). To this



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

list we also added [Digital Humanities Summer Institute](#) (“DHSI”), as one of the most established recurring learning opportunities in DH.

It was not possible to cover all courses available in the registry in the similar manner and instead we decided to take a random sample of 20 courses (out of 90 available). The initial sample was updated after few courses were discarded upon observation, because of inaccessibility or irrelevance for CLS: full reproducible sampling procedure is available in the repository.

Initially we also planned to annotate a random samples of DH conference papers (Weingart 2020) and published papers in computational literary studies (corpus provided by CLS INFRA WP3, Task 1), adding a dimension of “practice” to our grid, but it quickly became apparent that this path is not feasible. Papers are already products of research cycles and they are rarely concerned with skill transfer, so their overall connection to the skills grid is weak. Instead of producing annotations that would be, at best, poorly justified, we decided to focus on courses and workshops.

We started at Digital Humanities, since computational literary studies might be seen as a subset of the activity in a larger tent. DH is historically heavy text-oriented, getting its first ground in many literary departments, so we build on the previous surveying work of a community to use a wider discipline as an entry point to charting activities that are CLS-specific. In short, our small sample is both heavily dependent on DH and heavily leaned towards the European landscape.

## 2.2. Annotation procedure

The two authors of this manuscript, Artjoms Šeļa (AŠ) and Lisanne M. van Rossum (LS), were also the supply annotators: it was a deliberate choice, since we designed the skills grid together and had a shared vision of it. LR was responsible for course coding, while AŠ was responsible for schools. LR and AŠ did first align their coding and understanding of the grid, making a joint trial annotation of workshops of European Summer University in Digital Humanities 2020 (“ESU (DH)”, held in 2021).

Annotation itself followed few explicit principles:

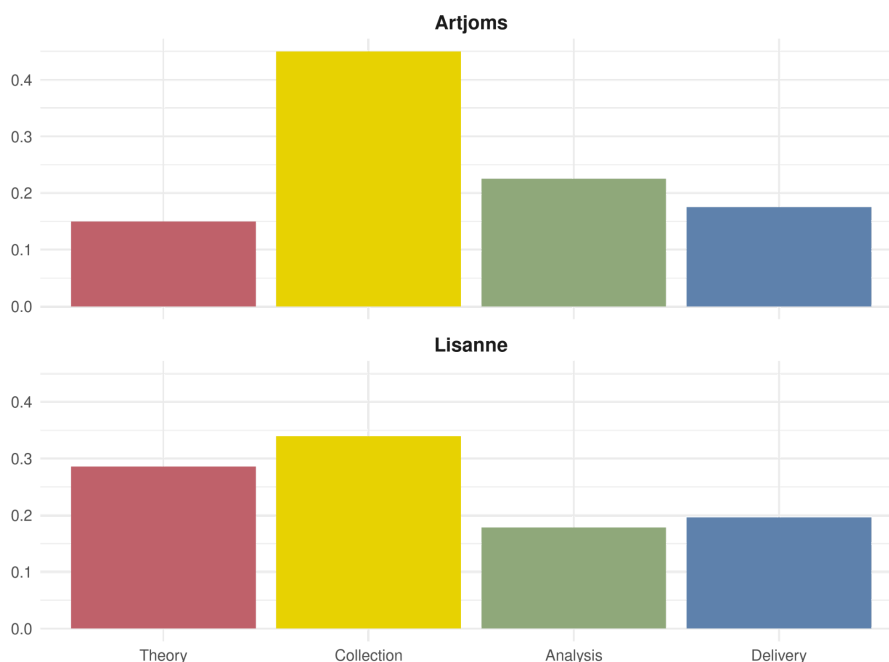
- **One entity (course or workshop) could be registered to a few skills on a grid** to allow flexibility in describing teaching. Skills often come in bundles (e.g. text-markup workshops often deal with data models and web publishing), and university courses usually offer constellations of DH-related knowledge;

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

- We **preferred a summary** or an abstract of a workshop as a basis for annotation. Title-based annotations were done only in clear-cut cases and the choices are reflected in the annotation materials;
- To be annotated, a workshop or course should have had an **explicit connection to CLS**, which usually means - an explicit connection to text as an object of inquiry / analysis. Exceptions were possible for generalizable areas (critical DH, statistics, project management).
- Only CLS and grid-related workshops/courses were annotated. This is especially important for schools, since our annotation data has gaps and **cannot be used to study the structure of summer schools**, or make any generalizations about it, outside of CLS context.

Since we had no ground-truth annotation and 1:1 entity-skill relationship, any assessment of annotation precision and inter-annotator agreement was problematic. Trial annotation of ESU 2020 workshops revealed overall similar profiles at the level of research steps, with noticeable discrepancy in **Theory** (Figure 2). This also accounted for the increase in annotation volume: LR had 56 “hits” across the skill grid, while AŠ, who underestimated theory-related workshop activities, had 40. The correlation between proportions of matching individual skills in two annotation sets was 0.5, which might be considered suboptimal. The overall approach was renegotiated between LR and AŠ afterwards, with both annotators agreeing on “focus over nuance” principle in their view of workshop descriptions (each workshop should be described by as minimum positions on a grid as possible, so that we annotate its main focus, not its academic flair).

Figure 2. AŠ and LR annotations of ESU 2020 profiles bundled across research steps.



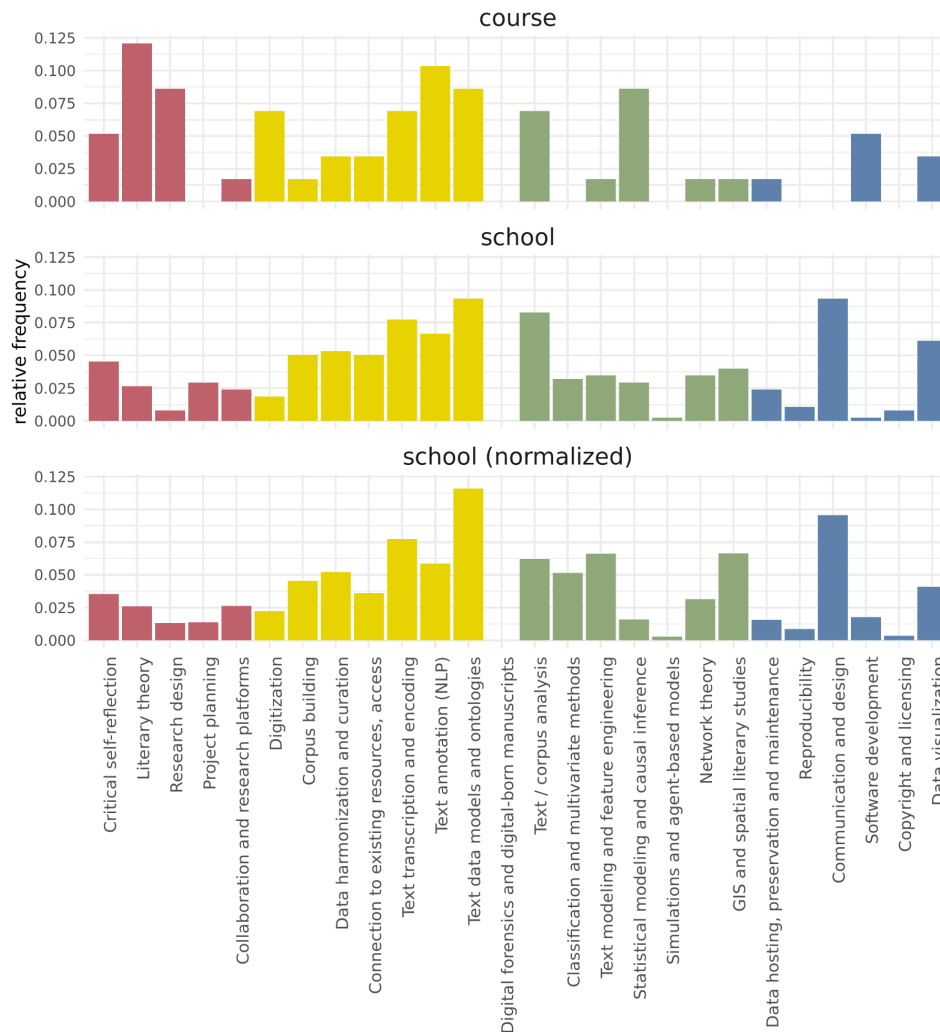
## 2.3. Supply overview

Annotation results are presented on Figure 3:

- **course** panel shows relative numbers of all course “hits” across the grid;
- similarly, **school** panel shows relative numbers of all course annotations taken together;
- **school (normalized)** reflects an attempt to normalize annotations per each school. Absolute number of annotations differ per school - large ones that also have data for many years, like ESU, would skew the picture. Normalization included three steps: 1) deriving relative number of annotations per skill for each school independently; 2) stacking these numbers together (sum); 3) deriving the proportions again. As a result, a small school that had 20% of hits in Network theory would equally contribute to the global average.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Figure 3. Proportions of annotations per each skill in the grid.



We see that the largest mass of all annotations is distributed in the **Collection** part both in schools and courses. This is expected, as data collection and wrangling are both crucial parts of research, frequently independent from the research object (text mining can serve multitude of questions and frameworks); also its transferability, suitability for workshops, is high. We see that high **Theory** values in university courses and low in short-term schooling might also depend on transferability discrepancy: workshops, by design, focus on practical skills, having less options for systematic teaching than semester-long courses.

**Analysis** and **Delivery** skills look more fragmented (especially in courses), but normalization allows to see that analysis is evened out: with one exception of **statistics**, that schools tend to avoid. Dedicated statistics workshops (including the single one in simulations) also mostly come from a relatively recently established school (*Digital Methods in Humanities and Social*

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

*Sciences*, University of Tartu, run since 2018) with its origins in the linguistics department. Similarly, normalization shows that some of the **collection** skills are present unevenly and are dominated by data models and text transcription.

There wasn't a single workshop or course in our sample dedicated to **digital forensics** and **digital-born manuscripts**. University courses also leave a lot of the grid unmapped: more data would be needed to make better generalizations on this basis; what we have now is only a small part of the picture that is also subjected to annotation bias.

## PART III. MAPPING DEMAND LANDSCAPE

### SUMMARY

- To index the demand side of the skills gap, researchers were surveyed for their needs and interests for training in computational skills for literary analysis.
- PART III reports on the production, promotion, and set-up of this survey, titled “Mapping skills landscape for Computational Literary Studies (CLS): a survey.”
- The survey was launched in January 2022 and had a run time of 4,5 weeks. It was widely disseminated using local networks, the [CLS INFRA project website](#) and [Twitter page](#), and several international digital newsletters and mailing lists.
- It was designed to correspond closely to the skills matrix but also asked a number of thematic open questions to complement the quantitative gap analysis.

### 3.1. Surveying demand in research community

#### 3.1.1. Target audience

To index demand for education in skills for computational literary analysis, we used a survey for widespread dissemination in the CLS research community. Our target audience were members of the CLS community, preferably with earlier involvement in the field, but we also wanted to target newcomers to gain insight into their experiences. Therefore we wanted to limit the time needed to fill out the survey, ensure it had a simple structure, and to not presume previous knowledge about computational methods for textual analysis by limiting jargon and providing explanations when broaching technical subject matter.<sup>1</sup> After some deliberation about flexibility, accessibility, and cost, we designed the survey in Google Forms.

#### 3.1.2. Objectives

The resulting survey was titled “Mapping skills landscape for Computational Literary Studies (CLS): a survey,” and intended to closely map to our skills grid to work towards a detailed comparison in the resulting gap analysis. As such, each of the skills corresponded to one entry in the survey, 26 “skill statements” in total. These core statements were preceded by two segments: an introductory segment and a first set of opening questions about the participant’s

---

<sup>1</sup> Sincerest thanks to the colleagues who provided us with valuable feedback on these fronts by doing a series of (timed) trials and/or by reviewing draft versions of the survey: Vera Maria Charvat, Ciara Lynn Murphy, Justin Tonra, Silvie Cinkova, Julie Birkholz, Serge Heiden, Julia Dudar, Ingo Boerner, Christof Schöch, Salvador Ros, Joanna Byszuk, Botond Szemes, Antonina Martynenko, Laura Hernández Lorenzo, and Maciej Eder.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

experience, career and gender information, which would allow us to control the results along these axes. They were followed by a series of open questions about the participant's experiences with training and schooling in CLS subjects, and to allow the participant to reflect on the comprehensiveness of the survey.

Summed up, the aims of the survey were to:

- Gain insight into our intended broad audience across variables such as gender, career stage, experience, and nationality, using a series of demographic questions;
- Map demand for CLS training and schooling in a manner closely related to the matrix used to index supply, using a series of scale statements;
- Take into account gaps beyond the skills matrix produced for this research, using a series of open questions.

### 3.1.3. Dissemination

With the intention of improving the accessibility of the survey for a wider audience, we recorded a short video as an introduction in which we explained the goal of the survey in broad strokes and including subtitles, while avoiding jargon and abbreviations as much as possible (text box 1). This video clip was uploaded on [YouTube](#) for embedding in the survey and was also featured in the promotion of the survey to provide an attractive audiovisual element for dissemination.

*Text box 1. Script of introductory video.*

Hello and thank you for your interest in our survey!

We are working on setting up an infrastructure for computational literary studies in Europe, and we want to hear from researchers, teachers, students, and practitioners; from you!

This survey asks questions about the current situation and prospective directions for teaching computational and research skills.

By completing it, you'll have the power to steer the field closer to the things -you think- are currently underrepresented.

This survey is mostly aimed at those who have some experience with computational literary studies. However, we would be happy to hear from anyone who is interested in the topic. Your participation will help us shape the future of CLS!

Thanks again for taking the time to complete the survey and don't forget to leave your email at the end if you would be interested in reading the final report.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

The survey was launched on the project Twitter, [www.twitter.com/CLSINFRA](https://www.twitter.com/CLSINFRA) (Figure 4), and the project website, [CLS INFRA Project Outputs](#) (Figure 5), on 20 January 2022 around noon CET.

Figure 4. Twitter launch of the survey.

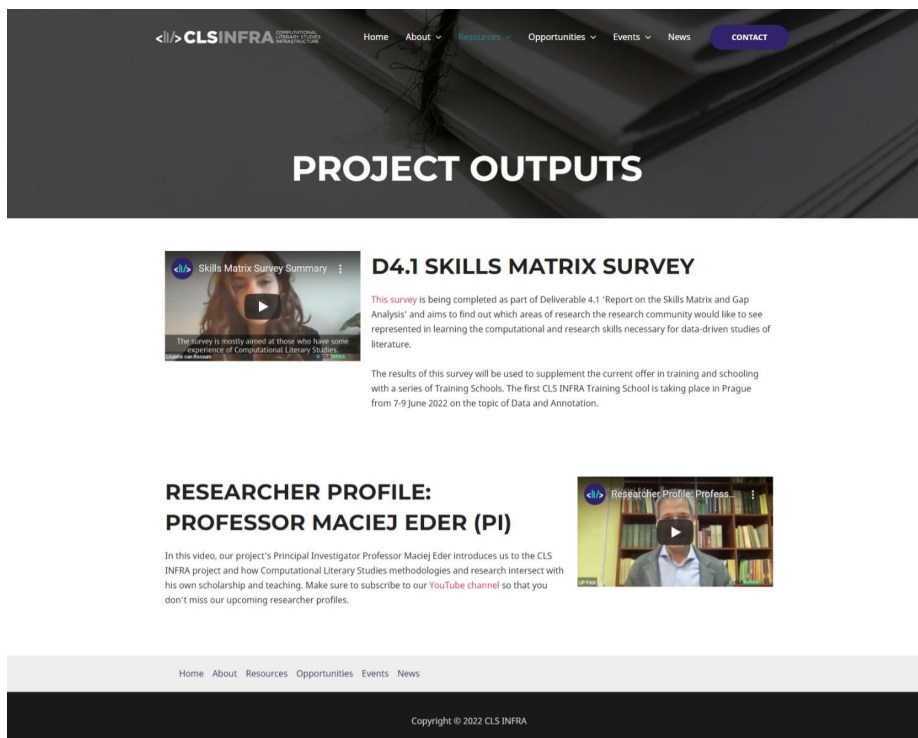
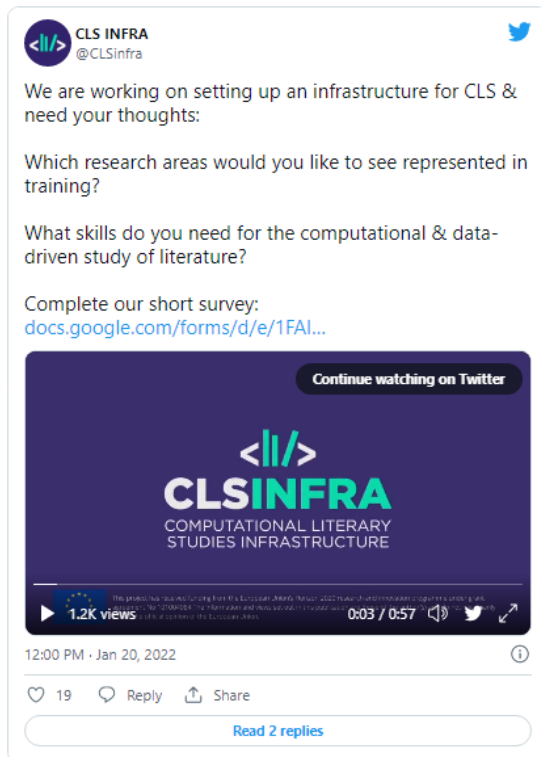


Figure 5. CLS INFRA project website launch of the survey.



Next to web and social media promotion strategies, the survey was also disseminated in the personal networks of CLS INFRA project members and Digital Humanities newsletters<sup>2</sup> using an e-mail message template (text box 2). A full overview of the promotion networks used can be found in the [Work Package 4 communications plan](#).

*Text box 2. Email template used to disseminate the survey.*

*[Subject]: Help CLS INFRA improve training in computational and research skills*

Dear colleagues,

CLS INFRA (<https://clsinfra.io/>) is a European Commission-funded project which aims to create unified and easy access to a wide range of European and national infrastructures for the Computational Literary Studies community. This includes: connecting the building of resources and infrastructure, providing training environments and networks, and promoting the theoretical considerations of these resources and infrastructures.

With this survey, CLS INFRA is looking for your help to find out which areas of research you would like to see represented in learning the computational and research skills necessary for data-driven studies of literature. Everyone with interest in the subject is invited to partake in the survey, regardless of experience level. The survey takes about 10-15 minutes to complete.

#### **What's in it for me?**

You will help us shape the future of computational literary studies!

The results of this survey will be used to supplement the current offer in training and schooling. CLS INFRA is organizing a series of Training Schools in the directions outlined by the research community. The first school is taking place in Prague from 7-9 June 2022 on the topic of Data and Annotation.

[Take me to the survey.](#)

Thanks in advance for your time – and for forwarding the survey to those who might be interested!

---

<sup>2</sup> Acknowledgements to Eliza Papaki for her help with disseminating the survey.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Best wishes on behalf of the CLS INFRA project team,

Artjoms Šeja and Lisanne van Rossum

Twitter: @CLSinfra / <https://twitter.com/clsinfra>

E-mail: [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl), [artjoms.sela@ijp.pan.pl](mailto:artjoms.sela@ijp.pan.pl)

After 33 days, or 4,5 weeks of run time, the survey was closed on 22 February 2022 around noon CET with a total of 118 responses.

## 3.2. Survey design

### 3.2.1. Introduction

The written statement introducing the survey and accompanying the promotion video featured a thank-you to the participant, together with an estimate of the time needed to complete the survey (text box 3). This time estimate was based on timed trials kindly provided by our Work Package members, who reported completing the survey between 10 and 15 minutes. In this segment, the participant first was guided through an introductory statement about the research objectives and theoretical background of the survey, together with a lay-out of the survey design as loosely following the stages of a research lifecycle.

Moreover, the segment included contact information for the CLS INFRA project and the authors, such as project website and Twitter handle, and a hyperlink to the [Task 4.1 task repository on Gitlab](#), which allowed participants to read additional information about the work informing the survey.

The introductory text was followed by a personal data protection statement and direction to a consent form about the processing of personal data in compliance with the European Union General Data Protection Regulation ('GDPR'). The privacy statement was drafted using the [DARIAH consent form wizard](#) and can be referred to in the supplement of this report. Only upon the participant's clicking of a [next] button, were they led to the first question portion of the survey.

*Text box 3. Introductory statement.*

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

***Dear participant, thank you for taking our CLS INFRA survey to index demand for training and schooling in computational literary studies. Expect this survey to take 10-15 minutes to complete.***

*The resulting report will be produced to better understand the landscape of demand and supply for computational and research skills necessary for doing data-driven studies of literature. This information will also be used to steer CLS INFRA Training schools in the directions outlined by the research community. The first school is taking place in Prague from 7-9 June 2022 on the topic of Data and Annotation.*

*Specific skills-related questions are organized in four large blocks that loosely follow a research life cycle: 1) Theory and research setup; 2) Collection; 3) Analysis and, finally, 4) Delivery. Knowledge production is complex and non-linear, so it is not necessary to assume those blocks and enclosed skills are linearly connected, but it might be useful to keep the overarching cycle in mind. You will find a general description of each block at the section's beginning. In case you need more information about the skills grid and our working process, you can visit our Gitlab repository listed below. Any and all of your experiences are greatly appreciated to help further our community.*

*Best wishes,*

*Lisanne van Rossum and Artjoms Šeļa,  
on behalf of the CLS INFRA project team*

*Web: <https://clsinfra.io/>*

*Twitter: @CLSinfra / <https://twitter.com/clsinfra>*

*Task repository: <https://gitlab.clsinfra.io/cls-infra/wp4/t4-1-skills-grid>*

*CLS INFRA values your privacy and processes your personal data in compliance with the EU General Data Protection Regulation (GDPR). Read more here: <https://file.io/C3kMclv7XwLR>*

### 3.2.2. Opening questions

The first section, the “Opening Questions,” asked a series of optional personal questions. The first was “In what career stage are you currently?”, providing a range of multiple choice options from undergraduate, to (post)graduate, PhD candidate, Postdoc, Lecturer/assistant professor, reader/associate professor, to professor, and also included the options “Retired” and “Other”, the latter of which allowed participants to self-identify in a short open response. We based these generalized career stages on a modified combination of Commonwealth/British and American academic career ladder conventions.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

The next multiple choice question, "What is your gender?", featured the options male, female, and non-binary/gender non-conforming. We realized that these three options were reductive, and to combat this we also added the options "Prefer not to say" to allow participants to explicitly opt out of the gender question next to the possibility of skipping it entirely. We also allowed participants to self-identify with an open question in case that was their preference.

The following question, "In which country are you located?" was a short-answer open question as to not limit the participant's freedom in identifying from any country or region, and to avoid a long multiple choice list of possible options.

The two final questions in this segment were a set of Likert scales from 1 to 5 that addressed the participant's own reflection on their involvement in Computational Literary Studies (1 "just curious" to 5 "it's my job"), as well as their perceived experience with computational methods (1 "beginner" to 5 "expert"). We decided on an uneven range from 1 to 5 to allow for a neutral middle (3) that the participant could use to relate their response to. The two scale statements were grouped together to be able to make observations about the expertise level of professionals versus casual "passers-by."

### 3.2.3. Scale statements

The core of the survey, the statements to measure the specific demand for the skills outlined in our research cycle grid, were set up along similar 1 to 5 Likert (1 "agree" to 5 "disagree") scales and subdivided by the same four categories Theory, Collection, Analysis and Delivery (Figure 6). Progress was indicated to the participant at the start of each section. Each section included a respective working description of the category, along with repeated suggestions how to approach the statements from the perspective of personal demand, but the participants were also asked to include their wider perspectives on the field when possible. This is illustrated by the N.B. in Figure 6.

*Figure 6. Example of category header.*

**Theory and research setup (1/4)**

The first step in our simplified research cycle, it encloses the discovery process that happens before the active research phase (but it could happen anytime): engagement with theory, exploration of the knowledge gaps, critical understanding of a discipline, research design practices and preparations. Platforms and frameworks for making collaboration and crowdsourcing work are included here, too.

**NB:** Please feel free to answer from the perspective of your individual interests, but we also invite you to answer from your perspective on the current state of CLS training. Note that we do not ask to judge the relative or intrinsic importance of a certain skill: all knowledge is important!

Figure 7. Example of a skill scale statement.

I want to see more opportunities to learn about **data harmonization and curation**  
 E.g. cleaning or standardizing data with regular expressions, adding and curating metadata, familiarity with common metadata standards.

1            2            3            4            5

Disagree                        Agree

As exemplified in Figure 7, each skill in the skills grid was translated into a statement starting with "I want to see more opportunities to learn about...". We used this particular phrasing to avoid inquiry after personal proficiency or experience, which was already addressed more broadly in the opening section of the survey. We also refrained from asking for relative importance of or hypothetical interest in a certain skill that would be implied with a "I (would) want to learn..." statement – it would undoubtedly warrant fascinating responses, but those fall outside of the scope of this research. We underlined our focus on pragmatic general or personal demand in the N.B. suggestions under each category. The responses, however, did display a comparable bias in our responses, which will be explored at further length in the analysis chapter.

We focused on each skill with an individual scaled statement rather than embedding them in a multiple choice grid per category, even though this would significantly reduce the length of the survey and lead to a faster workflow for the participant. Our main motivation for this is that we

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

felt that our approach to "skills" was not an universal given but rather an interpretative construct imposed on the respondent. Therefore we preferred a more elaborate format in which we could supply additional information in the form of example-based skill descriptions (chapter 1.2.). In drafting these examples we received helpful feedback from colleagues inside and outside the CLS INFRA project (table 2). Google Forms did not allow for an optional drop-down menu in case a participant needed clarification, so these examples had to be attached for every skill surveyed. To preserve visual emphasis on the statements themselves we placed them in and used manually generated bolding in unicode to highlight each skill because Google Forms is restrictive in its formatting options for headers.

However carefully worded, it still does appear from our responses that the illustrations chosen did not in every case map correctly to our participant’s understanding of what particular tool or method belongs under which skill label. We elaborate further on this issue in Chapter 4.3. And Supplement 6.3.5.

*Table 2. Version record of accompanying explanations for skill "software development" (Delivery category).*

"I want to see more opportunities to learn about <b>software development</b> "	
Initial draft version	Delivering usable, packaged tools, applications and interfaces.
First version with phrasing based on research questions	How to make resources out of your software (e.g. command-line tool, library for Python, or interactive apps on Web, e.g. with Shiny).
Second version	E.g. making resources out of software (command-line tool, Python library, interactive Web apps with Shiny).

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

<p>Third version with [commentary] from reviewer A</p>	<p>E.g. making resources out of software (command-line tool, Python library, interactive Web apps with Shiny).</p> <p>[In order to develop software you already have to know the basics (use of the command line, programming languages, etc.)...]</p> <p>Discard the question or rephrase it e.g., "I want to learn more about the basic computational background skills required for Digital Humanities, e.g., programming languages such as R and Python, using a versioning tool such as Git, ..."]</p>
<p>Fourth version with [suggestion] from reviewer B</p>	<p>While software development is itself a complex life cycle, here we understand it as primarily a delivery practice -- that makes resources out of your software (e.g. building a simple command-line tool, a Python [script or] library, an interactive Web application). Good-enough practices for development for researchers are covered here too.</p> <p>[remark: "scripting" (or soft software development) could be an interesting catch word]</p>
<p>Final version used in survey</p>	<p>While software development is itself a complex life cycle, here we understand it as primarily a delivery practice -- that makes resources out of your software (e.g. building a simple command-line tool, a Python script or library, an interactive Web application). Good-enough practices for development for researchers are covered here too.</p>

### 3.2.4. Closing questions

The final section of the survey, termed "Closing questions", featured a set of 5 open questions. The first open question was required and asked participants about their academic discipline(s) to obtain a clearer picture of which source communities were represented in this picture of demand for CLS training. We deliberately left room for more disciplines than one to acknowledge that this is not a straightforward issue – demonstrated by the ample use that our participants made of this option. Exactly half of our participants (59) listed more than one discipline and as many as 7 disciplines were reported by the same respondent.

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

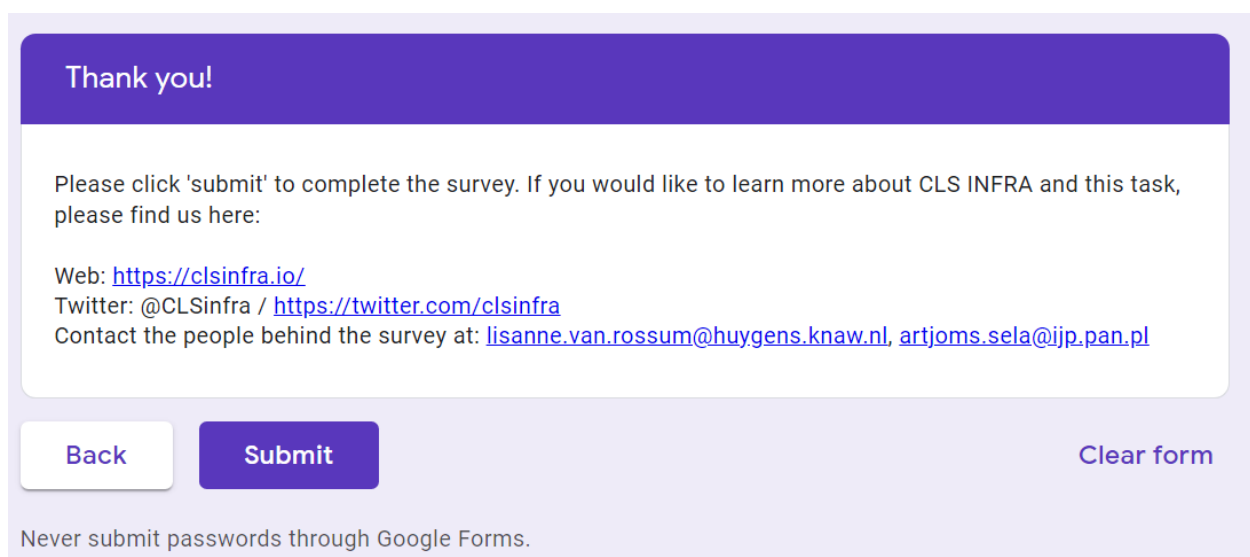
The following set of questions was optional and inquired after the participant's broader considerations for training and schooling in the field of CLS beyond the perspective of this survey. The first question, "What opportunities/challenges do you see for training and schooling in Computational Literary Studies (if any)?", gave the respondent the opportunity to identify areas of concern. In an earlier iteration of the survey, this subject was broached in two questions, one about opportunities and one about challenges. We ended up merging them because we preferred not to overstructure such a general open question and realized that the distinction between opportunity and pitfall is not always clear-cut.

We added the final question, "Are there any topics or areas that we missed in this survey?/Anything else to add?" because we realized that this survey with such an elaborate formulation process could in no way be complete, and to index which areas of research skills people would flag as important by their own initiative.

The final optional question asked the respondent for their email address in case they would like to receive a single update on the final report of the survey. The e-mail addresses collected will be used to send a newsletter to the respondents in March 2022.

Upon completion of the survey the participant was directed to the final window depicted in Figure 8, where we expressed our thanks and allowed them to review or submit their answers. To afford the participant another chance to engage with CLS INFRA after partaking in the survey, we repeated the project website, project Twitter, and work package team members' emails.

*Figure 8. Final submission window.*



Thank you!

Please click 'submit' to complete the survey. If you would like to learn more about CLS INFRA and this task, please find us here:

Web: <https://clsinfra.io/>  
Twitter: @CLSinfra / <https://twitter.com/clsinfra>  
Contact the people behind the survey at: [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl), [artjoms.sela@ijp.pan.pl](mailto:artjoms.sela@ijp.pan.pl)

Back Submit Clear form

Never submit passwords through Google Forms.



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

## PART IV. SURVEY RESULTS

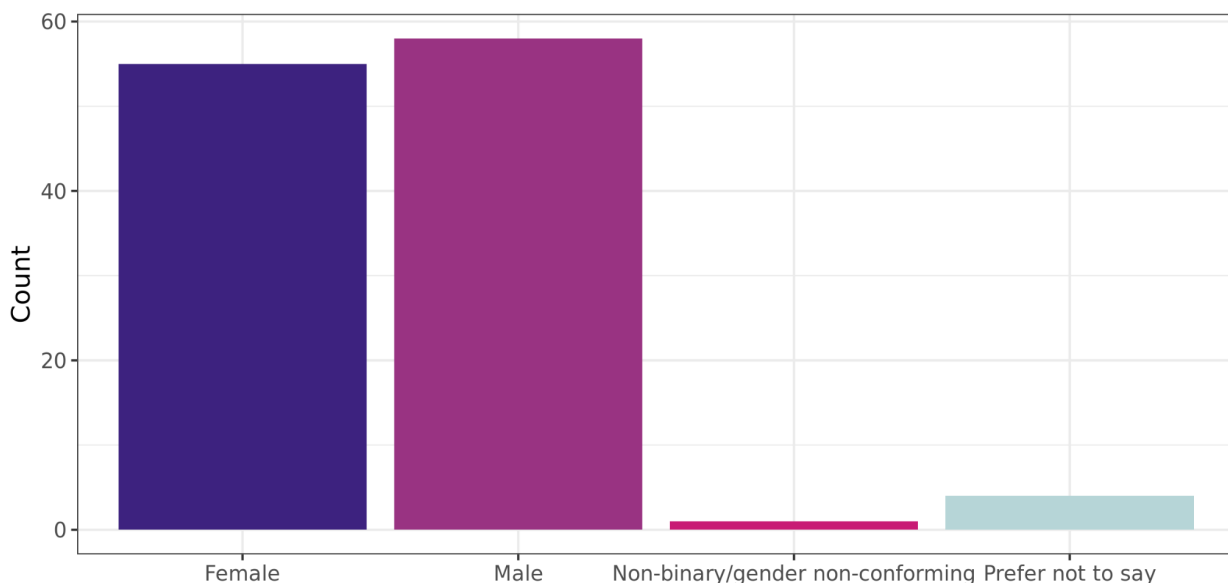
### SUMMARY

- PART IV reports on the results of the survey to index demand for training and schooling in CLS.
- The survey gathered a total of 118 participants from 26 different countries. The respondent pool seems to be superficially gender balanced, but the results do betray some indirect gender biases in the field.
- Respondents who report professional involvement with CLS also score themselves higher in computational skills. While unequally distributed, all academic career stages are represented in the survey. Qualitative analysis of open responses points to several thematic concerns for CLS (schooling) such as equal access, job opportunities, shared vision, and lack of (specialized/institutionalized) training. These concerns differ per career level.
- Analysis-based skills are among the highest scoring in demand while Delivery-based skills are among the lowest scoring in demand, although ratings were universally high.
- The lowest-scoring skills show traces of *disagreement* and opinion clusters. We invite a future exploration of the potential controversies.
- Respondents mainly have an academic background in 1) literary studies, 2) linguistics and 3) digital humanities. Relatively, literary scholars show decreased interest in analytical skills that are not immediately apparent for the analysis of literature, such as statistics.

## 4.1 Participants

### 4.1.1. Gender

Figure 9. Gender distribution of survey respondents.

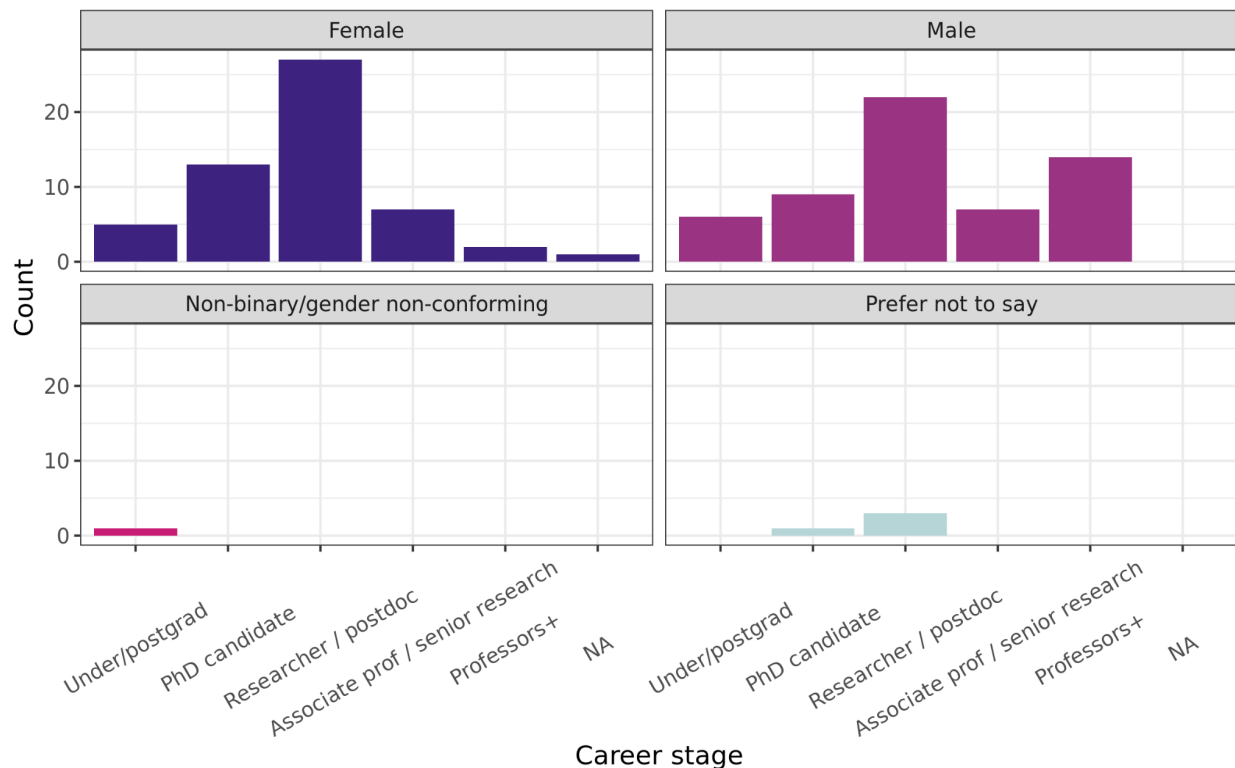


DELIVERABLE 4.1 SKILLS GAP ANALYSIS

All 118 participants chose to answer this question. 55 respondents self-identified as female as opposed to 58 male respondents, 1 non-binary or gender-nonconforming respondent, and 4 who preferred not to say (Figure 9). With a 46.6% / 49.2% female / male ratio, it appears that at first glance that our survey respondents have a near-equal gender balance, but we will also take gender into account when looking at distribution along other variables such as career stage. Note that our approach to gender here is largely pragmatic and focused on the male/female divide because it allows us to make the most quantitatively informed observations. One non-binary or gender-nonconforming respondent may for instance indicate that this is a relative minority in the field or a failure on our part to reach a non-binary or gender-nonconforming audience with the survey, but the available data is simply too sparse.

4.1.2. Career stage

Figure 10. Career stage distribution of survey respondents.



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

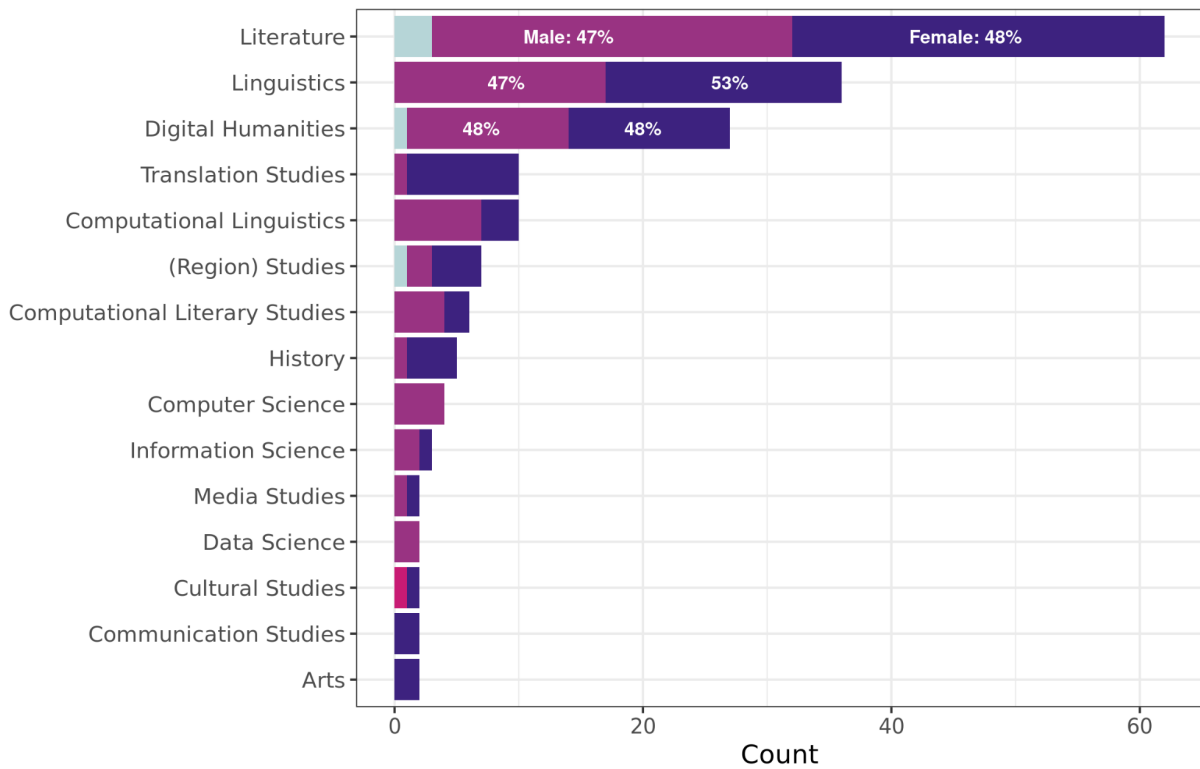
The question on the topic of career stage gathered 117 responses in all. Data was sorted by implicit career "seniority" (see Supplement 6.3.1). Career stage distribution by gender (Figure 10) looks similar for men and women, but a bias is present. Female respondents are in earlier career stages than their male counterparts, most notably PhD candidates and Researchers / Postdocs. Moreover, there is an overrepresentation of male professors.

The main pool of respondents to this survey is early career or pursuing a PhD, which suggests something about generation dynamics within the CLS community. CLS may clearly be a young field, but we are missing information about the wider career stage distribution in academia to be able to make relative statements. It is important to note that response expresses *demand in training*, and is not a direct inventory of CLS *membership*. As such, those who may have little interest in training due to advanced proficiency may have not participated in the survey.

### 4.1.3. Academic background

The open nature of the discipline question in the survey resulted in a lot of post-processing of responses and finding common labels (see Supplement 6.3.2). For additional clarity in summary presentation we have further collapsed all mentions of (national or regional) literature under one label, and did the same for linguistics and region studies. We do this to roughly de-sparse data to get a general insight into CLS INFRA main target audiences (literary scholars, linguists, digital humanities community). Figure 11 shows the absolute frequency of discipline mentions that had appeared at least twice (87% of all 209 mentions), with the gender distribution across the categories. Note that the overall count does not correspond to the number of participants here, but instead reflects the distribution of mentioned disciplines.

Figure 11. Discipline mentions across participants split by the underlying gender distribution.



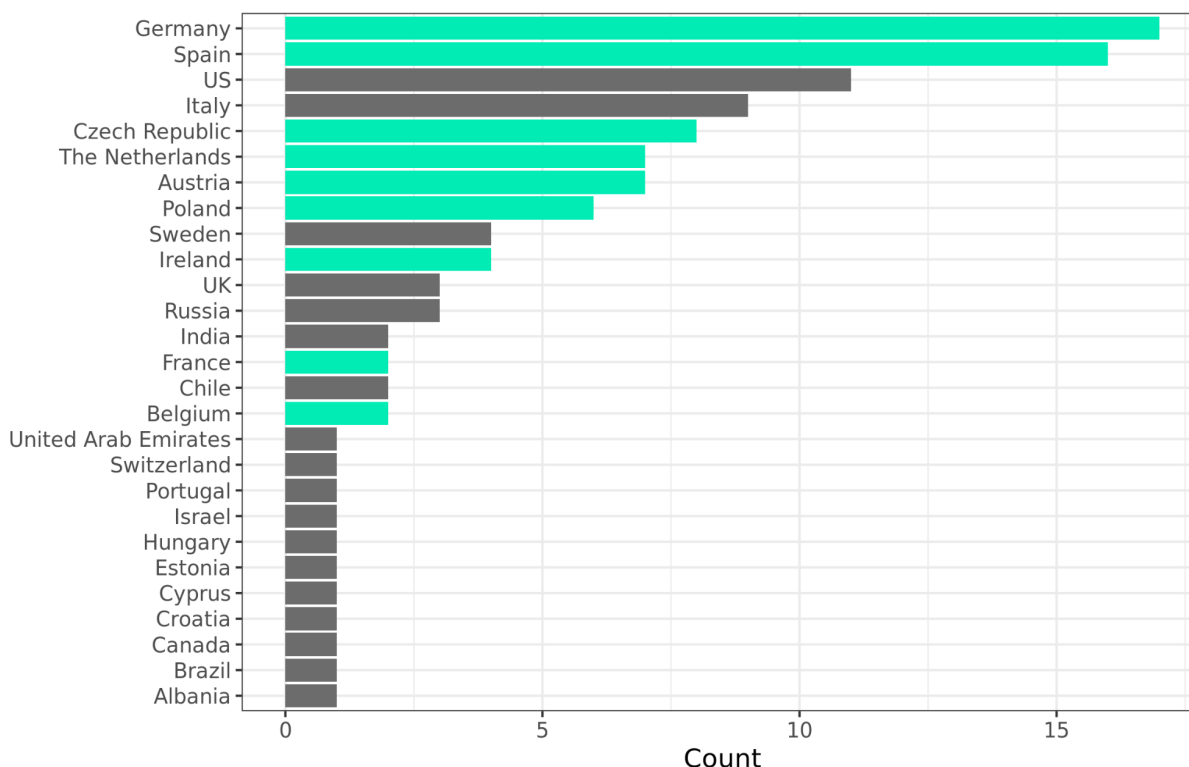
Two most frequently used disciplines for self-identification are unsurprising, considering the target audience of the survey: literary studies and linguistics. More interesting is the third place: Digital Humanities is commonly used as an academic background identification. Even moreso: 8 respondents listed DH as their sole discipline. Comparatively, computational literary studies is listed in the same fashion (sole identity) only once, and even then the word “computational” appeared in brackets.

Disciplines that were mentioned once include: 1) **subfields** like narratology, stylometry, discourse analysis, Shakespearean studies or genre studies 2) **digital prefix** and new media studies (Digital Literacy, Digital Storytelling, Digital Pedagogy), 3) **adjacent fields** (sociology, psychology, cultural evolution, cultural analytics).

Gender distribution in three main disciplines reflects the general tendency across the survey for surface-level balanced representation. However, the mentions of technical backgrounds (computational literary studies, computational linguistics, computer science, data science) are male dominated, which possibly points to biases that continue to affect education and academia around STEM / humanities divide.

#### 4.1.4. Geographical span

Figure 12. Geographical span distribution of survey respondents. Colored bars mark CLS INFRA partner countries, 4 responses with absent location were excluded from the plot.



The 114 respondents answered the open question about geographical location, reporting 26 different countries (after manual normalization, see Supplement 6.3.3.). Figure 12 displays the survey's geographical span, which has reached non-EU respondents from the United Kingdom, United States of America, India, Chile, the United Arab Emirates, Canada, and Brazil. Among our respondents, Germany, Spain, and the United States of America are among the most represented, while Eastern European countries are among the least represented, next to France, Belgium, Switzerland, and Portugal. Apart from Sweden, Northern European countries are absent. Evidenced by the open responses to this survey, we do not interpret underrepresentation in this metric at face value as a lack of demand for training in these countries, but rather as asymmetry in access or connection.

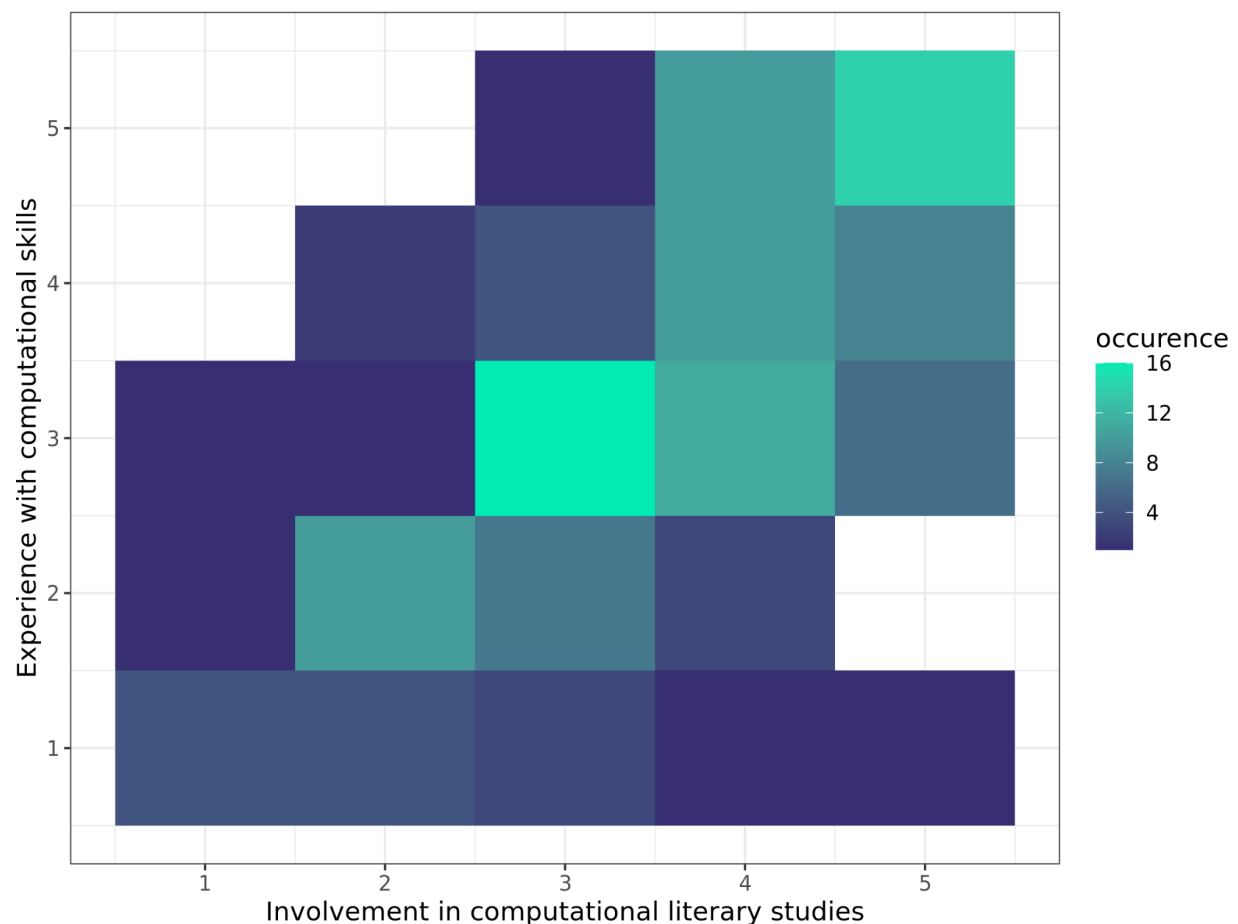
Speaking of connections, other European member countries of CLS INFRA besides Germany and Spain score well, such as the Czech Republic, The Netherlands, Austria, and Ireland. Because the survey was disseminated through the project member's local research networks this may have skewed the data in CLS INFRA member countries' favor.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

### 4.1.5. Involvement in CLS and computational experience

How important is the knowledge of computational skills for the involvement in the CLS and computation-adjacent humanities? While this complex and much-debated question, closely related to code literacy (see Bleeker et al. 2021; Bowers et al. 2021; Klein 2022) lies outside of the scope of this paper, we can at least observe patterns in the current practice among survey participants. We investigated the relationship between participant’s self-reporting of experience in computational skills and perceived involvement in CLS, both responses on a Likert scale from 1-5 (1 least, 5 most).

Figure 13. Heat matrix of per-participant combinations of responses on involvement and experience scales.



As seen in Figure 13, there is an overall positive relationship between the two self-reported values (Kendall’s tau: 0.56). Respondents rarely work in CLS without computational skills and vice versa: people who have computational skills are involved in CLS, which is likely the result of survey selection bias.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

The average participant involvement in CLS was 3.55 and average reported experience in computational skills was 3.21. When divided by gender, we can see a clear difference in the distribution of both involvement and experience by male and female respondents (table 3).

*Table 3. Average self-reporting CLS involvement and computational experience per respondent category on Likert scale from 1 (least) to 5 (most).*

	Average involvement	Average skill experience
total	3.55	3.21
<b>female</b>	<b>3.23</b>	<b>2.75</b>
<b>male</b>	<b>3.98</b>	<b>3.77</b>

One of the possible explanations for this difference is gendered self-reporting bias, which has been previously explored, e.g. in the context of math performance. Bench et al. found that men tend to overestimate their own abilities when self-reporting performance (2015). The same effect may be at play here, further substantiated by the finding that 12 out of 14 total participants who rated themselves with 5 in both CLS involvement and computational skills ("5x5") were men (table 4).

*Table 4. Participants who awarded themselves the highest score of 5 (out of 5) on CLS involvement and on skill experience ("5x5").*

	5x5 occurrences
total	14
<b>female</b>	<b>2</b>
<b>male</b>	<b>12</b>

These numbers can be confounded however by the predominance of males in late(r) career stages. We suspect that the career and self-reporting biases interact, for example, among the 5x5 scorers were 3 male professors (out of 14) while the 2 female professors among our respondents rated themselves 2x2 and 3x3, respectively.



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

In other words, we could be encountering a gendered self-reporting bias and/or an effect of gender asymmetry in career level, but it is ultimately outside the scope of this report to tease out any causal reason for the gender discrepancy in our data.

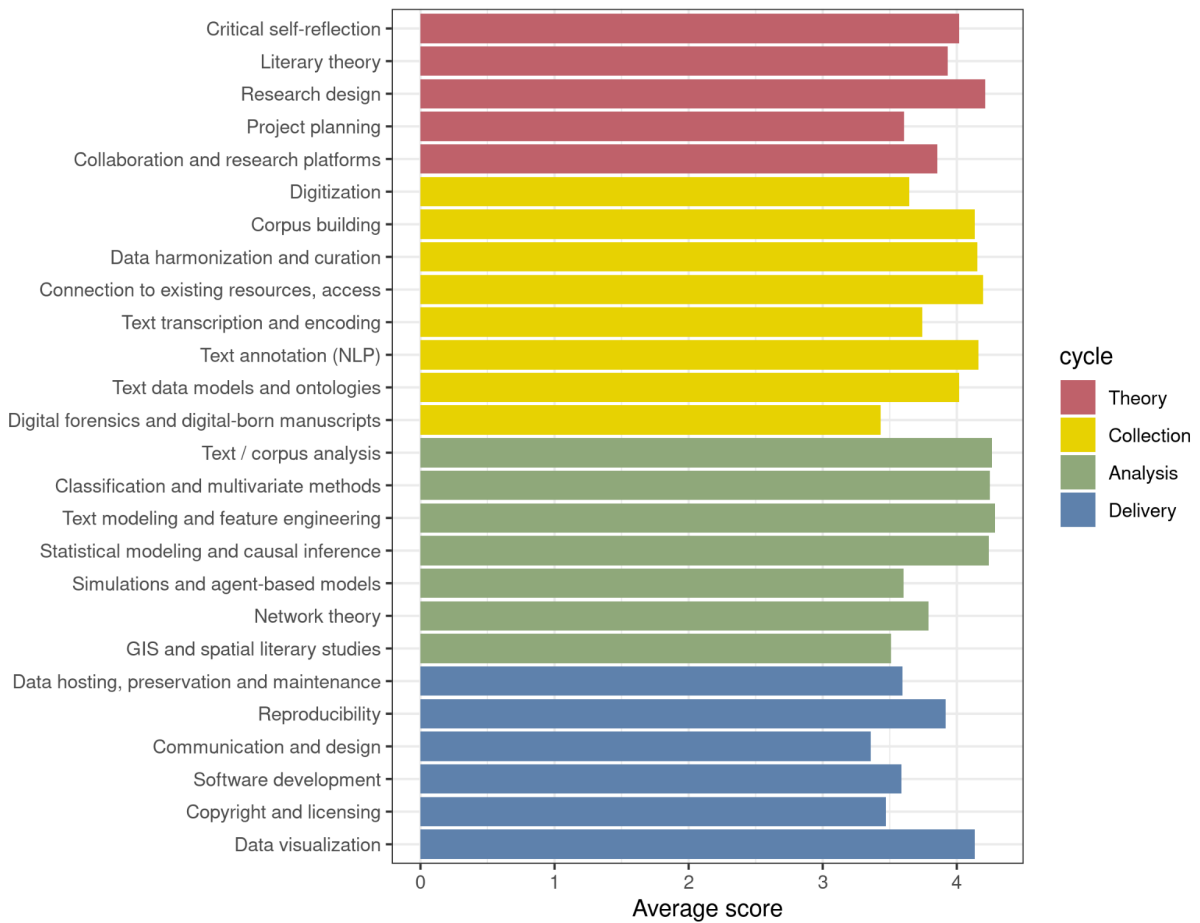
## 4.2. Skills

### 4.2.1. Scores

Overall, each skill in the surveyed grid scored high. The average score across the grid was **3.9**. The largest scoring skill was **Text modeling and feature engineering** (4.28), the lowest scoring - **Communication and design** (3.36), with less than 1 point in average difference. The distribution of survey answers was heavily skewed towards 5 and variability in some groups of participants was low (see next section). Figure 14 shows distribution of average scores around the grid and a really small span of differences between skills. In this context, it is not very insightful to speak about individual best / worst performing skills, but it invites us to look at groups of skills that are performing similarly.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Figure 14. Survey results: Average score of skills.



A look at the 5 highest and lowest scoring skills reveal the general centers of gravity of practitioners' opinions (Table 5). We see that the highest scores are driven by the direct text analysis procedures, statistics and research design questions, while the low scoring skills could be seen as supplementary to the research cycle: few of them originate in Delivery, relate to research-adjacent problems (legal questions, software development, forensics and preservation). If we take into account that four next highest scoring skills all come from Collection (access, NLP, data harmonization and corpus building), what collective opinion seems to be shaping out of our grid is a **backbone research kit**: focus on research design & procedures, data preparation, text analysis and statistical inference.

Table 5. Highest and lowest scoring skills in skills matrix, color coded by category.

Highest scoring	Lowest scoring
Text modeling and feature engineering	Communication and design
Text / corpus analysis	Digital forensics and digital-born manuscripts
Classification and multivariate methods	Copyright and licensing
Statistical modeling and causal inference	GIS and spatial literary studies
Research design	Software development

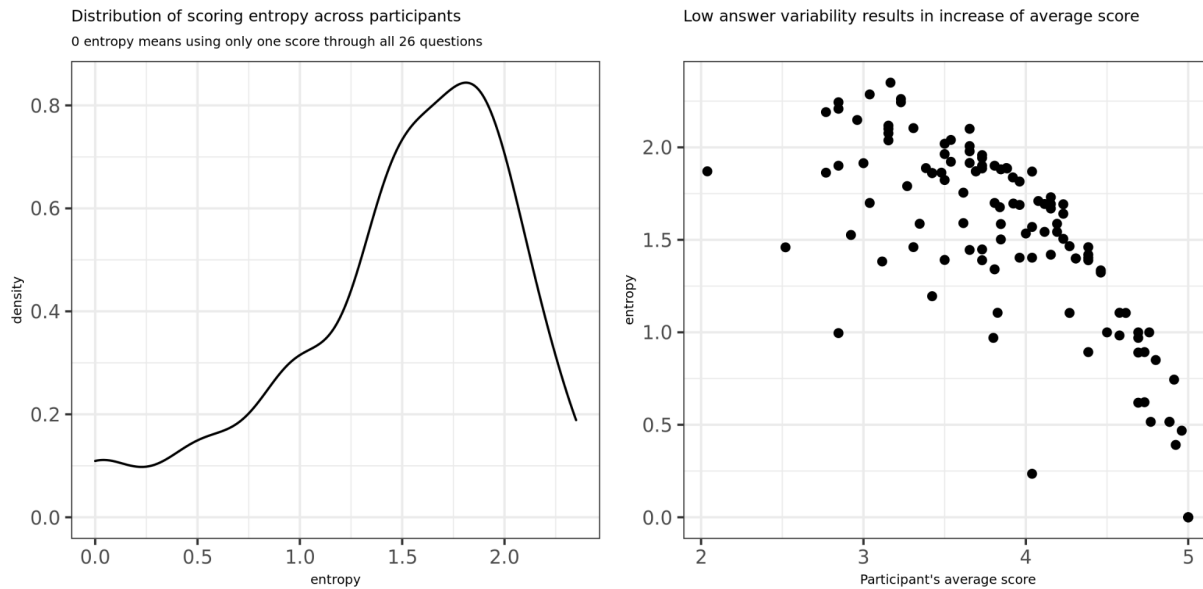
### 4.2.2. Score variability

Data above suggest that we are dealing with high uniformity in answers (many people score high for many skills). The effect of individual inner scale might be in play (people can use 3 points as the lower end of their evaluation). There could be an effect of novelty/saturation: for newcomers all skills might seem important, but practitioners would have developed a relational system of views in which some skills might be seen as already well-represented. Finally, the survey design itself was restrictive: we asked a prospective question about the future situation with certain knowledge domains and offered one uniform scale for every answer. Low scores would meet a natural resistance here.

To have a quick estimate of how participants tend to score the grid, we calculate Shannon's entropy for the probability distribution of each participant's score. If a person only votes "5", their entropy would be 0, if a person uses all scores equally, their scoring entropy would be 2.3 - exactly an entropy of a fair five-sided die.

The average entropy across participants was 1.49, the distribution of entropy scores also show a lean towards lower entropy spectrum (long left tail). Decrease in entropy also results in increase in average score per participant as seen on the right pane of Figure 15 (correlation between entropy and score is -0.75). This is quite expected, of course: there are not many options to score high, but at least it also shows that we don't have people who only score "medium low" values of 1, 2, 3 -- which would be like saying "I don't really care about CLS".

Figure 15. Distribution of entropy scores (left) and entropy x mean score relationship (right).



There were 6 zero entropy participants - their answers will be discarded in the final phase of the gap analysis, since we are primarily interested in *the gap* and the relative perception of the skill's representation in available training. Uniform answers do not add new information, but might swamp the averages. Entropy also does not show any strong association with career stages, computational experience or CLS involvement and gender, but on average women tend to have slightly higher median entropy than men, while students and professors both have high variability in answers. It might point to the specific visions these groups have, or just reflect the inequality in population sizes. We see that the largest demographic groups (career, involvement, experience) produce the largest dispersion of entropy scores.

It is important to note that, looking at the score distributions (Supplement 6.5) per skill, we see that **participants' agreement on high score skills is more uniform than on low score skills**. In other words, people tend to rate skills high uniformly, but disagreement comes in clusters. This possibly indicates latent controversies and clusters of opinions for many lower-scoring skills. We see two-peaked (on 3 and 5) distributions on **project planning**, **digital forensics**, **simulations and agent-based modeling**, **GIS**, **data preservation** and **software development**. This might reflect distinct groups of opinions and vectors of interest, based on participant's previous exposure to computational skills and training in Digital Humanities. In the end, averages are not telling the whole picture, and CLS INFRA might explore these potentially controversial topics in future.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

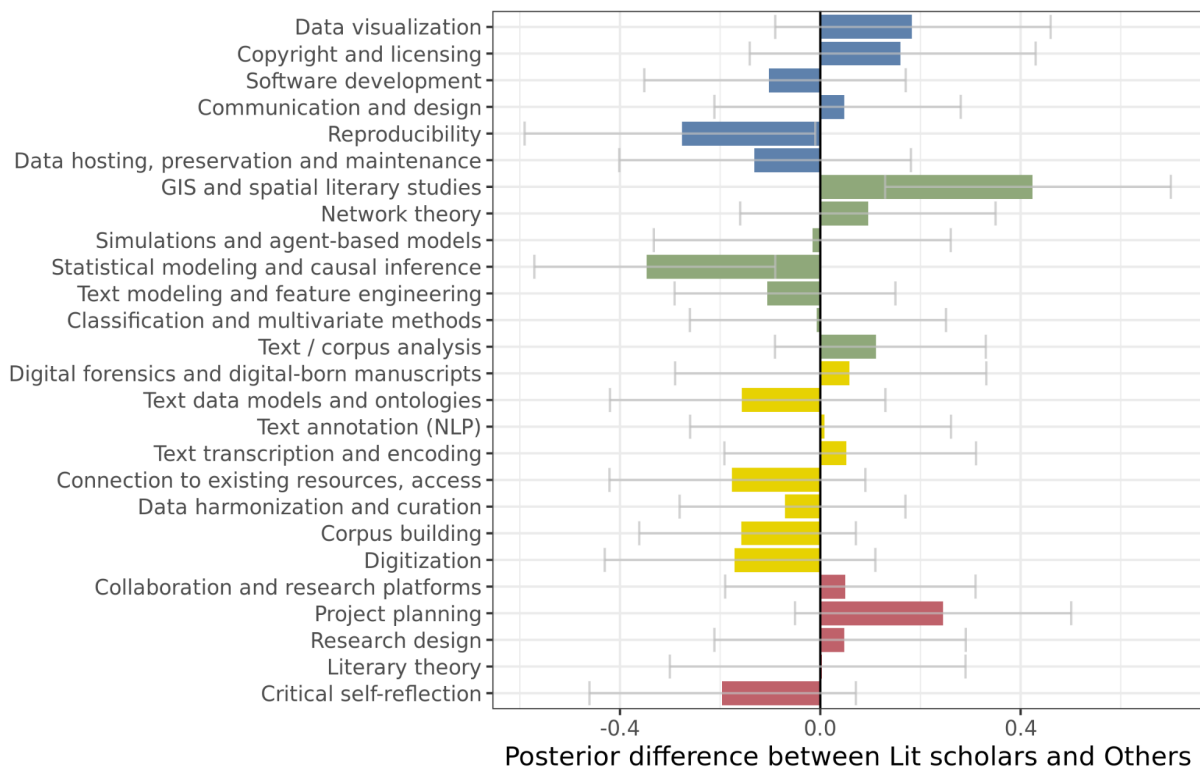
### 4.2.3. What do literary scholars want?

Exactly half of the participants (59) indicated (in one way or another) their relationship to literary studies in their answers. Since this community might be the most immediate target audience for CLS INFRA, it would be important to isolate this group and compare its view on the skills grid with other respondents.

We know already that score differences shaped by the survey are minuscule on the absolute scale, so we need a way to estimate our uncertainty about the found differences. To do that, we fit an ordinal Bayesian regression that predicts score probability by 1) individual skill in a grid and 2) discipline affiliation that was simply split to “Literary scholars” and “Others”. We allow interaction between skills and discipline to be able to model an influence that belonging to a discipline may have on an average skill score.

Figure 16 shows posterior distribution of differences in each skill between average scores of 100 simulated Literary scholars and 100 simulated Others. Error bars show 90% credible intervals and columns indicate an average difference.

*Figure 16. Predicted differences between literary scholars and others. Error bars show 90% credible intervals and columns indicate an average difference.*



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

We see that the many skills exhibit no effect and small differences often lie within wide credible intervals. There are, however, differences that we are more certain about: Literary scholars want **less statistics and more GIS spatial analysis**, compared to another half of the survey. Smaller effects are: preference for **Project planning**, **Data visualization**, **copyright and licensing**, while lower scores for **Reproducibility** and some Collection skills (**Connection to resources**, **Corpus building**, **Digitization**). These are, however, fading trends.

These tendencies could reflect the field's idiosyncrasies: for example, uses of statistical models for literary questions might not be obvious, while maps and geospatial data are easier relatable to literary work. Note that other high scoring **analysis** skills that were directly related to text do not show differences for literary scholars: the general preference for research-oriented cycle remains in this group, at least on paper.

### 4.3. Open questions

#### 4.3.1. Opportunities and challenges for CLS training

68 out of 118 respondents answered the open question "What opportunities/challenges do you see for training and schooling in Computational Literary Studies (if any)?" Although the question specifically asked about training and schooling in CLS, many participants instead responded with general commentary on the state of the field. After manual thematic color-coding (see Supplement 6.3.4.), the data set for the open question consisted of 95 labeled entries divided over 11 topics (table 6).

*Table 6. Color-coded topic labels and entry frequencies in opportunities and challenges responses in order of most to least frequent.*

colour_topic_label	entry_freq
Lack of (centralized) training	19
(In)compatibility CLS	14
Institutionalization	10

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Access	9
Skills to learn	8
Shared vision	8
Textual landscape	7
External constraints	6
(Professional) opportunities	6
Overwhelming	5
Lack of intermediate/advanced training	2

The topic mentioned most frequently by respondents was a **general lack or lack of centrally organized training**. A researcher / postdoc from the United Kingdom, for example, states: “There are very few, and also those that are available (such as DHSI) are short term. Learning technical skills requires long-term investment--I would like to see some training schools that last for several weeks or a few months, perhaps 10 weeks over the summer for example.” And an undergrad from Germany suggests to “Allow people to build a coherent set of skills, with a certain progressive level, by taking several workshops.”

Second comes the topic of the **compatibility between CLS and other disciplines**, especially the traditional Humanities, illustrated by this reply from an Austrian professor: “I see a fundamental challenge in the opposition of some traditional literary scholars towards quantitative/computational methods. Interdisciplinary approaches are seen as a threat to the original discipline and its theories and methods.” Or as an Irish researcher / postdoc encapsulates the sentiment: “overcome the idea that science and Literary Studies are opposite”.

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Third is the concern that institutions are lagging behind in solidifying **CLS theory and methods in core curricula**. A Czech PhD student writes we should “Encourage humanities students to learn the basics of programming during undergrad. Their hands will be very much untied. High schools do not provide this yet.” Several topics are woven together in the response from this Austrian researcher / postdoc: “Challenge: many students and scholars from the field of literature have little knowledge about the opportunities provided by computational literary studies. An introduction to the basics should already be part of university courses, and/or there should be an easy starting point for those interested but not equipped with programming skills.”

Inequalities in **access** to resources for CLS between different communities and languages was also a key issue in the open responses. A researcher / postdoc from the United States phrases it as “Breaking the barriers for access, increase opportunities, breaking barriers for education gaps. [sic]” A colleague from Cyprus illustrates that, “As there is no training in this subject where I live, I started a master's degree in IT with the aim of improving myself in the field of computer literature studies. I formed a research team and set up a website [...] to show our computer analysis literature projects.” Other statements, like this one from a German researcher / postdocs, describe similar tenets that “Analysis of multilingual corpuses/corpuses in other languages than English are underrepresented. Often tools are not adapted enough for these analyses. It would be important to train how to proceed in these cases, how tools can be adapted”.

Other respondents saw the open question as an opportunity to list **desirable skill areas in CLS training**. A Polish post / undergraduate, for example, writes “Learn more about methods, tools,” echoing a statement from a Albanian researcher / postdoc who identifies as an “Opportunity: Learning on how to apply emerging technologies to the study of literary texts.” A Czech PhD candidate makes their personal interpretation of the question explicit by stating “Opportunities (where I think I could use the skills): textual analyses, corpuses of big data, discourse analysis (online discussions related to news media etc.) [sic]”.

Other participants shifted focus away from individual interests and addressed the **shared direction for CLS** as a discipline. A Polish professor sees a challenge in “unknown long-term objectives”, while an Austrian researcher / postdoc ponders: “Maybe making it clear who exactly belongs to this field or where it is situated and what topics fall under it. How does it relate to people who define their field as “Text Mining” more generally for instance? Should they be separated or integrated?”, expressing a need for a clear(er) delineation of CLS as an umbrella term. Several participants also expressed demand for shared or standardized research practice: “Establishment of best practice and research standards,” a German researcher / postdoc writes. A PhD student from the same country similarly answered “Meet the research-specific requirements and conditions”.

The **heterogeneity of the textual landscape** in CLS practice was another recurring theme in the open responses. Participants identified several areas or aspects of computational analysis that they would like to see expanded. For example, an Associate / senior professor from Brazil states “I see opportunities for building corpora for studying Medieval Portuguese, since I am



#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

interested in prosody (rhythm, intonation), considering medieval cantigas.” Others, such as an American professor, describe “The tension between archive (what has survived from the past) and corpus (what is available to study computationally). Findings based on a corpus do not necessarily apply to the archive.”

Several respondents went a more existential route by characterizing the topic of CLS training and schooling as one of **limited resources**: “Time and money,” a professor from Austria states. A Czech PhD student mentions that “Life is short and there is so much to learn”, which can be read as both an opportunity and a pitfall. Some participants also view academia or the activity of building infrastructure itself as a constraint - a Croatian professor warns of the “...danger of devoting a disproportionate amount of resources to infrastructure and theoretical critique of CLS when there is still so relatively little applied CLS practice.”

Our participants also show concern for the intersection between CLS and **career prospects**. Some approach CLS training as a potential stepping stone for professional development, such as an Italian researcher / postdoc who writes that “training in CLS could be the starting point of a wider range of post-doctoral opportunities for PhD candidates that want to work outside the academic world.” Yet other respondents address a perceived gap in specific opportunities to pursue a career in CLS: A Swedish researcher / postdoc writes that “If you are interested in this field, you currently need to work within the much wider (and heterogeneous) field of the digital humanities, which is not always optimal.” This point is reinforced by an Italian colleague who states that “I would like to have more opportunities both for my research and for building an innovative path in the PhD programme I coordinate.”

Another topic arising from our data was that participants feel **intimidated or discouraged** because they miss an entry point or structured path in(to) CLS education. As an American researcher / postdoc describes the experience: “I am overwhelmed with the number of skills in this survey that I feel I need to develop all on top of my regular graduate work. It also seems difficult to learn these skills in ways that are easily applicable to my work but also general enough for people from diverse fields to easily apply them to their work.” An Indian peer describes the learning curve as simply “Immense”, while a German colleague specifies: “Very heterogeneous and large skill set needed; difficult to decide what is really important. In many respects someone who is doing CLS will be using standard solutions, where is it impeding research if the research can only apply these?”

Finally, two participants describe that the **gap in CLS training is qualitative** rather than quantitative, e.g. that there is enough training, but that supply and demand are not mutually aligned. A Polish PhD student states that “It seems to me that there is a lack and need for intermediate courses, there are many opportunities for introductions into various areas you mention, and some for really advanced users, but quite few for people who feel they are in the middle.” A researcher / postdoc from the United States makes a similar observation, although they experience advanced-level training as even more scarce: “Coursework in computational literary studies rarely surpass [sic] the introductory level. There seem to be plenty of introductory

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

courses, for students with no experience, but very few courses for students with programming experience, and fewer yet for advanced students.”

### 4.3.2. Response distribution per career stage

These responses are insightful when close-read, but for broader analysis, the resulting number of entries for each topic is too limited to make unified statements about how the 26 national communities involved in the survey view certain topics in CLS training. However, we found that we can make broader observations when introducing the variable career stage. For instance, all entries labeled **(Professional) opportunities** stem from early career respondents. Distribution per career stage is displayed in table 6; the same data is visualized in Figure 17.

*Table 6. Topic frequency distribution per career stage of survey respondents, including entry ratio.*

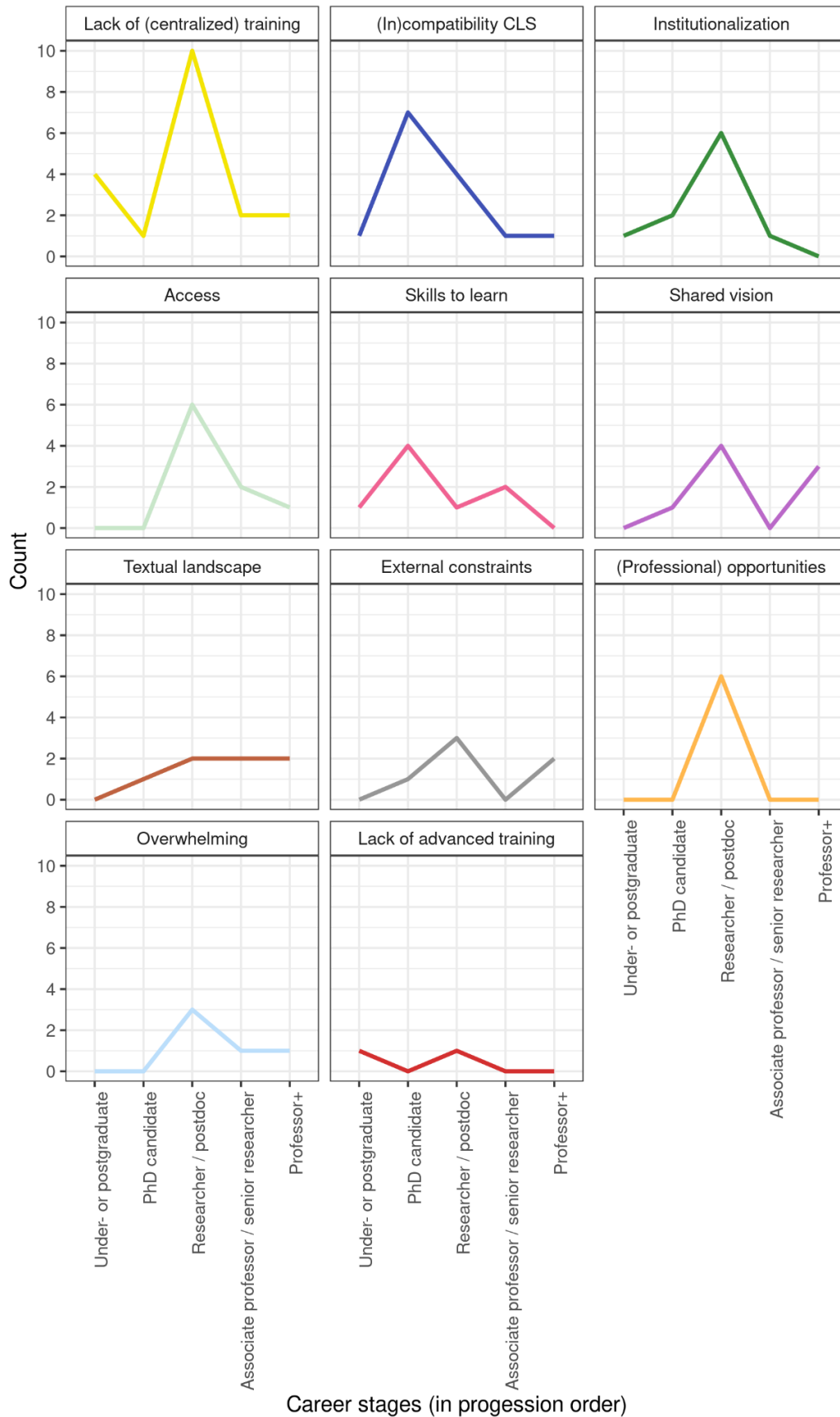
colour_topic_label freq_career_stage	Under- postgraduate	or PhD candidate	Researcher / postdoc	Associate professor / senior researcher	Professor+
Lack of (centralized) training	4	1	10	2	2
(In)compatibility CLS	1	7	4	1	1
Institutionalization	1	2	6	1	0
Access	0	0	6	2	1
Skills to learn	1	4	1	2	0
Shared vision	0	1	4	0	3

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Textual landscape	0	1	2	2	2
External constraints	0	1	3	0	2
(Professional) opportunities	0	0	6	0	0
Overwhelming	0	0	3	1	1
Lack of intermediate/advanced training	1	0	1	0	0
Percentage of total number of entries	8,5%	18,1%	48,9%	11,7%	12,77%

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Figure 17. Visualization of topic frequency distribution per career stage of survey respondents.



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

What can be learned from figure 15 is that researchers and postdocs account for almost half of all the coded entries. That this introduces bias in our data set is evident from the recurring peaks in the middle of the plots for many topics. With this in mind, the topics of plots that *diverge* from that shape warrant closer attention.

It appears that PhD students experience the **perceived disciplinary incompatibility of CLS** most strongly, and that this is also the group that reports wanting to learn the most **specific skills for textual analysis**. PhD students also do not seem to address the **lack of (centralized) training** as frequently in their open responses as the other career categories, as opposed to under/post-graduates themselves. **Job opportunities**, **institutionalization**, and **access to resources** appear to be highly relevant topics for early career researchers. In turn, Professors mention with relative frequency the topics of **external constraints** to learning and teaching research skills, such as time, money, and academic practice. They also raise the issue of formulating a **shared vision** for CLS as a discipline, with centralized research standards.

### 4.3.3. Missed areas in survey

53 out of 118 respondents answered the open question "Are there any topics or areas that we missed in this survey? / Anything else to add?", with 16 participants offering suggestions for topics missed in the survey. These responses were manually color-coded to produce a data set of 16 entries spanning 5 overarching topics (see Supplement 6.3.5.), listed in table 7. Note that there is some thematic overlap in the labels for responses between the two open questions, e.g. the label **Parallel corpora** bears resemblance to the topics **Textual Landscape** and **Access, Application beyond CLS** incorporates elements from the labels **(In)compatibility CLS** and **Shared vision**, and **Application beyond academia** mirrors responses about **(Professional) opportunities** and **institutionalization**.

*Table 7: Topic labels and entry frequencies in missed area responses in order of most to least frequent.*

colour_theme	frequency
Parallel corpora	7
Application beyond CLS	3

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Application beyond academia	2
Deep learning	2
Other	2

Because this open question warranted so few substantive (and often brief, see Supplement 6.3.5.) responses, it is methodologically slippery to use demographic variables such as gender or nationality to interpret them. Rather, we will provide a brief summary for each of the tendencies identified within the responses and highlight an individual response for each topic.

**Parallel corpora** is a phrase borrowed from a respondent to categorize the issue brought up most frequently by respondents: that not all text is equal within current CLS practice. In their answers, participants draw attention to “source texts and their translations,” as a German researcher / postdoc words it, less-researched target languages for literary analysis, (machine translation of) historical text, and multilingual research.

**Application beyond CLS** groups together participants’s suggestions about the impact of CLS on other related (computational) disciplines. One Spanish researcher / postdoc, for example, shares his opinion that “Interdisciplinary studies should be at the forefront, especially in the digital domain. For instance, I am working on a digital edition of a musical source that at the same time requires attention from diachronous linguistics, and there is little opportunity to find colleagues to actually look at this.”

**Application beyond academia** is the term to group responses from participants who want to see more integration between CLS and other non-academic fields, such as high schools and the commercial sector. An Italian professor explains that “We are asked to link our research to the job market, to make joint projects with private companies. I wonder if this aspect can be considered in your approach.”

**Deep learning** was an area covered in the survey under the broader term “machine learning.” Two respondents agreed that it was a missed topic - an Austrian researcher / postdoc kindly points out that “Within machine learning, deep learning is in practice a distinct area.”

**Other** was the label used to group the two last responses without a thematic twin. The missed areas brought up were co-design with users and “matters that relate to the ethics of taking data (such as social media interactions or online book reviews) from the web,” the latter of which was suggested by a professor from India.

## PART V. GAP ANALYSIS

### SUMMARY

- Building on the previous chapters, PART V delivers the gap analysis between demand and supply in CLS training and schooling.
- The methodological steps taken to align demand and supply data are 1) annotation normalization, 2) survey normalization, and 3) bridging annotations and the survey. Survey data was then subdivided by respondent discipline and annotation source to create a more detailed heat map.
- Gap areas identified are research design, certain areas of collection, and analysis. These areas should be the prime suspects for attention in future planning in CLS, both for short-term teaching offers and institutional shifts in academia.
- Yet more so, data do not show many “overrepresented” areas except for text encoding, data modeling, and communication and design, which suggests that diversification and supplementation is a larger concern for CLS training than combating oversaturation.
- Integrated analysis confirms that the skills gap in CLS is indeed more quantitative than qualitative (“the lack is the gap”) and suggests it is potentially generational. Institutionalizing CLS as a recognized discipline and offering structured training would be critically valuable steps taken towards closing that gap.

### 5.1. Methodological steps

We start at constructing a collapsed and averaged “Demand vs. Supply” heatmap that appears in Figure 1 (in Executive Summary): we map all training annotations (hits across skill grid) to one set of values, and survey scores to another.

**STEP 1: Annotation normalization.** Since we had overwhelmingly more absolute grid hits for schools than for courses, to represent them together we:

- 1) derive relative values for each of the sources independently (schools, courses)
- 2) take averages across schools and courses as final set of values
- 3) treat missing grid positions as missing data and not zeros. If either school or a course had a skill registered that was unobserved, its final value would be assumed to be the value from the observed counterpart.

**STEP 2: Survey normalization.** We use average scores of skills as our reference point. We prefer the mean, and not the absolute score, since some skills were left unscored by various participants, and technically some grid entries have non-equal number of respondents. Zero entropy participants that showed no variation in their scoring were excluded from the survey data.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

**STEP 3: Bridging annotations and the survey.** Finally, we transform demand and supply values to 0--1 scale, so that the lowest average proportion in supply, or lowest average score in demand would take the values of 0 (and maximum value would become 1). Another possible approach would be ranking the values and comparing change in ranks, but then the inner difference scales for demand and supply would be destroyed, and it would be critical to maintain them for demand with its many skills scoring in close proximity.

Figure 18. Normalized survey and annotation data plotted side by side as a heatmap (left) and distribution of differences between two dimensions (right).

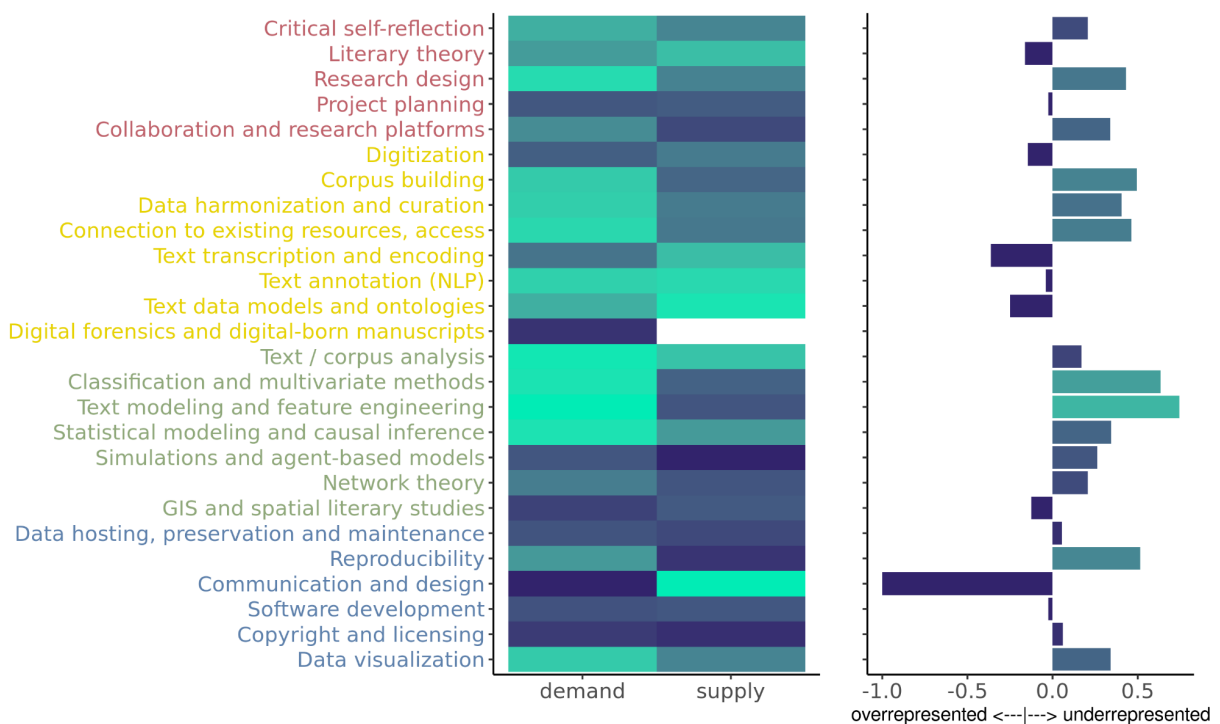


Figure 18 shows a heatmap of demand and supply together with absolute difference between them (survey values minus annotation values). The lighted areas on the left tell the story of the survey: focus is on **research design**, certain areas of **collection** and **analysis**. Incidentally, those lighted up areas are relatively less represented in the teaching offer, as reflected in the positive differences. These areas should be the prime suspects for attention in future planning in CLS, both for short-term teaching offers and institutional shifts in academia.

Noticeably, data do not show too many “overrepresented” areas except for **text encoding**, **data modeling** and, of course, **communication and design**, which scores the biggest possible difference of 1. This is an accident that results mainly from our annotation approach.



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Communication and design skills (writing, publishing HTML, etc.) often come in a bundle with other training (text markup, project planning, networks), and are rarely the main focus of workshops. We do overestimate its presence in supply: this does not render the difference meaningless, of course, but lowers the dramatism of the gap.

These results and the areas of heightened demand show that the CLS community might focus on increasing the presence of certain anticipated opportunities and paths, instead of worrying about overrepresentation.

## 5.2. Final heatmap

The presented above picture of “all in one bowl” has a limited value. It does not normalize supply data by schools that did bring a disproportionate number of annotations (ESU, DHSI, Oxford), and it problematically lumps all complicated survey demographics together, too. The matrix could be further segmented and shaped across a multitude of metadata dimensions. To de-aggregate the figure 18 in the previous part and highlight its problems, we chose to split survey data by discipline - by putting literary studies and digital humanities in the spotlight and contrasting them to each other and remaining respondents. At the same time, courses and school annotations now would be treated separately (as they should have been).

Figure 19. The heatmap across 3 demand dimensions (discipline) and 2 supply dimensions (schools and courses).

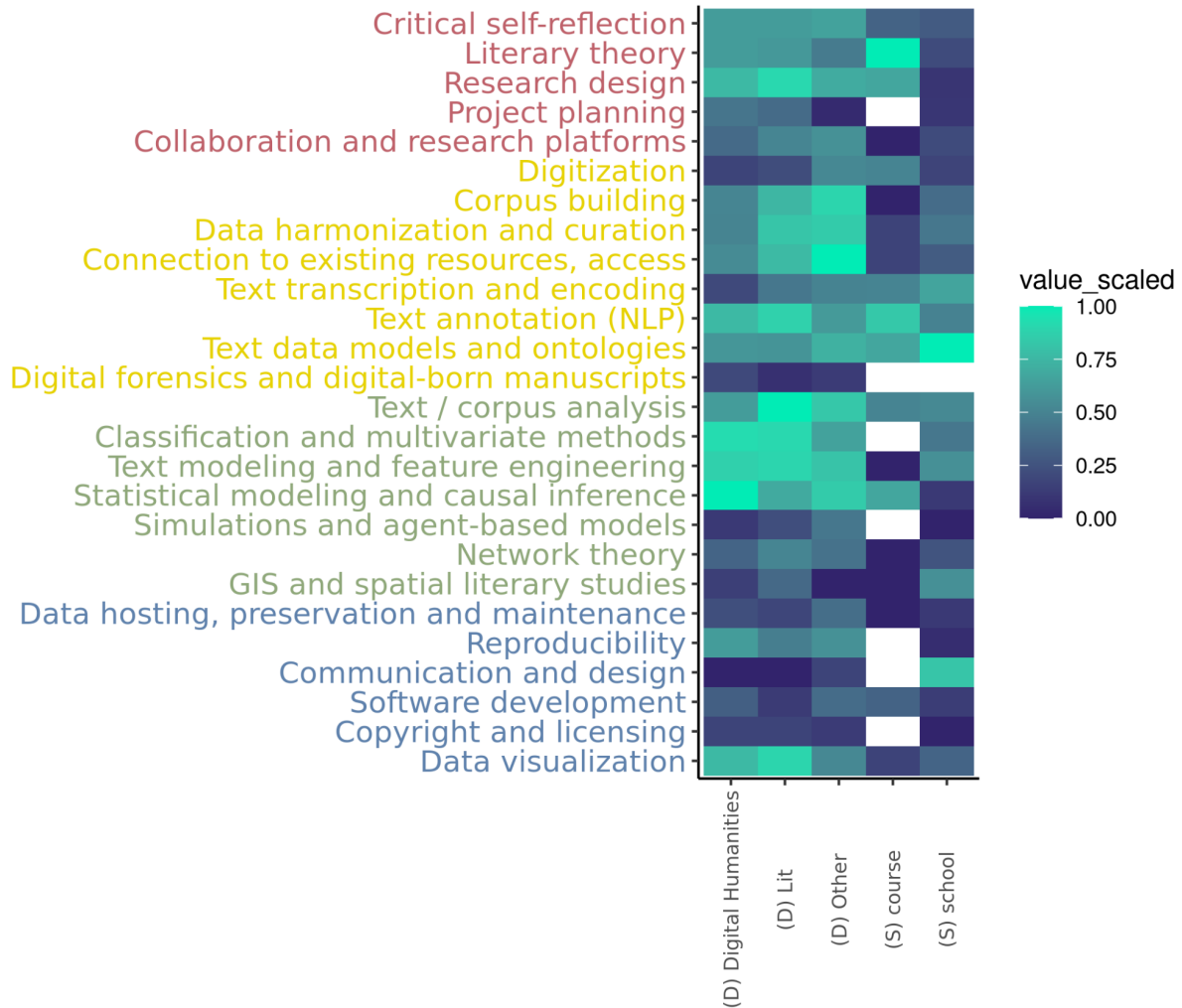


Figure 19 allows to better show areas where different supply sources have mutually distributive relationships (more theory for courses, more delivery and transferable collection skills in schools). Some areas appear to be rarely covered in summer school formats, like statistics and research design.

From the supply perspective several patches of micro-disagreements could be seen with the overall visible consensus. Respondents who identified with Digital Humanities have a darker shade for the large part of collection and theory skills. Literary scholars, as seen previously in section 4.2.3., root less for statistics and more for GIS (although relatively), while Others do supply high values for data acquisition skills.

### 5.3. Conclusion: where approaches intersect

Our quantitative findings have their limitations. Annotation reliability can (and must) always be questioned, and mapping user expectations from the survey's Likert scale onto frequency data of supply might be seen as a questionable decision that makes differences hardly interpretable. In the end, instead of a clear skills gap we found areas of heightened interest for CLS that would be important to explore bilaterally. Specifically, teaching research design, data collection and reuse, advanced text analysis and statistical modeling might become the avenues of increased focus of CLS INFRA and computational literary studies in general.

We will therefore present an *integrated* analysis of the gap in training and schooling for CLS by foregrounding overlaps in our quantitative and qualitative findings. In section 4.2.2. we hypothesized, on the basis of survey respondents who have awarded all skills relatively high scores, that there might be an effect of novelty or saturation at play for newcomers to the discipline. We may have encountered a record of this effect in the response from the American researcher / postdoc who shared frankly: "I am overwhelmed with the number of skills in this survey". Because CLS is a young field, we must remind ourselves not all newcomers are junior researchers, but also researchers who may have been introduced to the field later in their careers. In fact, all responses labeled "overwhelmed" are reported by the career stages "researcher / postdoc" and up.

Here we draw parallels with the consensus among survey participants that starting training early is important through institutionalized and centralized offer, which is propelled forward by disciplinary identification and recognition. These are indeed the issues that hold (relative) weight for our early career researchers and under, who seem to experience the divide between traditional and computational humanities most strongly. This stresses the urgency in building shared schooling infrastructure for the field of CLS, encapsulated by a Polish (under)graduate's answer of "training a new generation of CLS researchers.", which could be interpreted both as a hopeful opportunity and a critical challenge.

Moreover, the predominant opinion among survey respondents was that the availability of training *itself* is an issue for CLS training reinforces that the skills gap is more a quantitative gap in general training, rather than a qualitative gap in oversaturation of selective skill areas.

Lastly, inequalities in access and underrepresentation in the current textual landscape for CLS, such as source language corpora and historic, multilingual, or translated corpora, were among the key themes for both the open question about opportunities / challenges for CLS training and missed areas in the survey. Signals of a key demand for the research community, one that has been a blind spot in our quantitative gap analysis. This exposes the added value of additional, bottom-up qualitative research to supplement theoretically informed top-down models. We stress that however carefully the skills matrix was reviewed and drafted, we do not presume that

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

it is a comprehensive or final way to deal with this important issue. Instead, we hope this heatmap and accumulated data will prove a valuable reference for shaping the future of CLS.

A final note on the part of the researchers. Looking at the male domination of adjacent fields and our own experiences in academia, we suspect that gender bias is perhaps the true hidden inequality in this work. The behavior of the data set along gender lines has continuously attracted our attention, and where possible without overstepping the scope of the report, we have aimed to make these subtle biases explicit. We suspect, however, that there is more to contextualize and interpret about the role of gender in CLS and we consider the topic an important avenue for future work.

## PART VI. SUPPLEMENT

### 6.0. Materials and code

All materials, code, models and anonymized survey data are available at our Gitlab repository: <https://gitlab.clsinfra.io/cls-infra/wp4/t4-1-skills-grid>. Quantitative analysis, modeling and data visualization was done using R language.

### 6.1. References

Bench, S. W., Lench, H. C., Liew, J., Miner, K., & Flores, S. A. (2015). Gender gaps in overestimation of math performance. *Sex Roles: A Journal of Research*, 72(11-12), 536–546. <https://doi.org/10.1007/s11199-015-0486-9>

Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities eScience. *FUTURE GENERATION COMPUTER SYSTEMS*, 29(2), 654-661. <https://doi.org/10.1016/j.future.2011.06.006>

Bleeker, E., Beelen, K., Chambers, S., Koolen, M., Melgar-Estrada, L., & Van Zundert, J. J. (2021, May 31). Persistence, Self-Doubt, and Curiosity: Surveying Code Literacy in Digital Humanities. DH Benelux 2021 #GoesOnline, World Wide Web. Zenodo. <https://doi.org/10.5281/zenodo.4889429>

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016). TaDiRAH: A Case Study in Pragmatic Classification. *Digital Humanities Quarterly*, 010(1).

Bowers, K., Dombrowski, Q., and Risam R. (2021). DSC #12: The DSC and the New Coding Language. *The Data-Sitters Club*, November 2, 2021. <https://datasittersclub.github.io/site/dsc12.html>.

Drude, S., Di Giorgio, S., Ronzino, P., Links, P., van Nispen, A., Verbrugge, K., Degl'Innocenti, E., Oltersdorf, J., Stiller, J., & Spiecker, C. (2016). PARTHENOS D2.1 Report on User Requirements. Zenodo. <https://doi.org/10.5281/zenodo.2204561>

Klein, L. F., curator (2022). Code. *Digital Pedagogy in the Humanities*. Humanities Commons. <https://digitalpedagogy.hcommons.org/keyword/Code/>

#### DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Tasovac, T., Barthauer, R., Buddenbohm, S., Clivaz, C., Ros, S., & Raciti, M. (2018). D7.1 Report about the skills base across existing and new DARIAH communities [Technical Report]. Belgrade Center of Digital Humanities ; DARIAH. <https://hal.archives-ouvertes.fr/hal-01857379>

Unsworth, J. (2000). Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this. In: *Symposium on Humanities Computing: Formal Methods, Experimental Practice*. King's College, London. Vol. 13, pp. 5-00.

Weingart, S.B., Eichmann-Kalwara, N., Lincoln, M., et al. *The Index of Digital Humanities Conferences*. Carnegie Mellon University, 2020. <https://dh-abstracts.library.cmu.edu>. <https://doi.org/10.34666/k1de-j489>

Wissik, T., Edmond, J., Fischer, F., de Jong, F., Stefania Scagliola, S., Scharnhorst, A. , Schmeer, H., Scholger, W. and Leon Wessels, L. (2020). Teaching Digital Humanities Around the World: An Infrastructural Approach to a Community-Driven DH Course Registry. *Library Tribune* 40: 1-27. Preprint available here: <https://hal.archives-ouvertes.fr/hal-02500871/document>

## 6.2. Annotation sources

Table 8. Summer school list and university courses / programs sample used in current teaching annotation. Based on DARIAH [Digital Humanities Course Registry](#) (Wissik 2020).

Schools	Courses (20 sample)
Antwerp Summer School in Digital Humanities, 2022	A Medieval and Early Modern Times with specialization DH (Friedrich-Alexander-Universität Erlangen-Nürnberg)
Digital Humanities at Oxford Summer School, 2011-2016, 2021	BA course: Digital Humanities (Universiteit Antwerpen)
Digital Humanities Summer Institute, 2021	Continuing Education Program "Digital Humanities" (University of Heidelberg)
Digital Methods in the Humanities and Social Sciences (Lausanne University), 2021	Data Clinic (University of Helsinki)
Digital Methods in Humanities and Social Sciences (University of Tartu), 2018-2020	Digital Humanities Laboratory - Laboratorio di Informatica Specialistica (University of Palermo)
European Summer University in Digital Humanities, 2015-2020	Digital Scholarship and Skills Workshop Series (Trinity College Dublin)
Summer School for Literary Studies & Digital Humanities (Leiden University, Netherlands), 2021	Estonian language processing in Python (University of Tartu)
Summer School Neo-Latin Studies and Digital Humanities (University of Bonn), 2021	MA Digital Humanities (ZIM-ACDH)
	MA in Rare Books and Digital Humanities - Université Bourgogne Franche-Comté
	MA Informatica Umanistica (Università degli Studi di Pisa)
	Master Humanidades y Patrimonio Digitales (Universitat Autònoma de Barcelona)
	Master in Digital Humanities (Université de Lausanne)
	Module: Digital Humanities: Concepts and Approaches (Moscow Higher School of Economics)
	PhD level course: Digital Humanities in Literary Studies (Palacký University)
	PhD/MA level course: Basics of Quantitative Data Analysis for the Humanities with R (University of Tartu)
	University Library Digital Humanities Course "Introduction to R & Data for Humanities"
	UT: Bachelor of Arts Sprache, Technologie, Medien (STeM)
	ZIM-ACDH Module: Information Modeling in the

	Humanities
--	------------

## 6.3. Survey response data normalization and encoding

### 6.3.1. Career stage

Career stage data was collected from one of the opening questions that allowed several fixed options, loosely corresponding to academic seniority: from undergraduate, to (post)graduate, PhD candidate, Postdoc, Lecturer/assistant professor, reader/associate professor, to professor. Options “Retired” and possibility to supply your own answer (“Other”) were also included. To process these answers for further work, we tagged these responses to five levels:

1. Students: Under/postgraduate students
2. Doctorate studies: PhD students / PhD candidates
3. Early career: Postdocs / Researchers / Adjuncts / Assistant Professors
4. Mid-level career: Associate professors / Docents / Senior researchers
5. Professors

There were 20 categories in respondent answers to tag, mainly those deviated from proposed answers in researcher jobs, seniority (“senior researcher”) or job positions outside of university-centered academic work (librarians, teachers). In those latter cases we were uncertain with tagging, but assumed a third level of our tagset, for simplicity.

### 6.3.2. Academic background

Raw data to be processed were respondent’s open answers to the question “what is/are your discipline(s)”? Exactly half of the respondents listed more than one discipline in their answer, and up to 7 disciplines were listed per respondent. To process the data for visualization, the different entries within each answer were separated by a new column. The entries remained on the same row so they could remain traceable to their respective respondent. For example, “Digital Humanities and Spanish literature” would be divided into Digital Humanities in column discipline\_1, and Spanish literature in column discipline\_2. After that, the responses were encoded under overarching labels. “Spanish literature” from the previous example was labeled Literature (Spanish), while “Phonology” became Linguistics (Phonology). To further reduce the number of initial labels into workable discipline categories for broader analysis, the label Comparative Literature was also counted as a part of the discipline category Literature, becoming Literature (Comparative).

To calculate the percentages of each discipline category, all its occurrences were counted and



## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

expressed as part of the total number of entries (which was greater than the number of respondents). We preferred this processing method to alternative encoding strategies that would flatten each response to a single discipline because the responses were so strongly multidisciplinary. We also wanted to avoid a hierarchical encoding method where first entries are weighed heavier than their later mentioned counterparts because we are unable to retrieve participant intentions or work under the assumption that one discipline must be more dominant than others. There is, however, some inherent bias in this processing method in that it counts all occurrences as equal: one of the respondents, for example, listed different areas of Translation Studies as multiple disciplines. These in turn were all counted as entries for Translation Studies (Translation Studies (Audiovisual), Translation Studies (Literary)), inflating the percentage of Translation Studies as if multiple respondents reported association with Translation Studies.

### 6.3.3. Geographical span

The responses to the question “In which country are you located?” received 114 responses out of 118 total responses. These responses were manually normalized into 26 categories, each representing a country. We identified 9 CLS INFRA member countries. Normalization was straight-forward and mostly spelling-based: “United States”, “USA” and “US” were merged to become the label “US.” Likewise, “BE” became “Belgium,” “Netherlands” became “The Netherlands”, “SPAIN” became “Spain”, and “The Czech Republic” and “Czechia” became “Czech Republic.”

### 6.3.4. Training challenges / opportunities

The optional open question "What opportunities/challenges do you see for training and schooling in Computational Literary Studies (if any)?" garnered 68 responses (out of 118 responses total), indicating that about 58% of participants opted to write an answer. These 68 responses, averaging ~22 words each, were analyzed and encoded using a bottom-up approach, meaning that they were accumulated, close-read, and then roughly categorized by up to three color-coded labels. All entries, regardless of length, were included. The result was a total number of 94 labeled entries. Each label corresponded with one overarching theme or topic in the responses to this open question, as illustrated in table 9.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

Table 9. Legend to color coding of responses to open question about training challenges / opportunities.

colour_topic_label	clarification	entry_freq
Overwhelming	A clear entry point into CLS material or training	5
Lack of intermediate/advanced training	A qualitative gap in training and schooling	2
Skills to learn	Wanting to learn any skills or a specific skill	8
Lack of (centralized) training	A quantitative gap in training and schooling	19
(In)compatibility CLS	Fundamental challenges of integrating CLS within traditional Humanities or other disciplines	14
Textual landscape	Unexplored avenues of textual scholarship such as historical sources, copyrighted material, or decentralized corpora	7
External constraints	Limited resources, such as time, bureaucracy, or money	6
(Professional) opportunities	CLS as a viable pursuit or career path	6
Institutionalization	Training in CLS being widely or sufficiently solidified in institutional curricula	10
Shared vision	Concern with (the formation of) a communal vision, standards of practice, or identity for the CLS community	8
Access	The topic of distributed opportunities to participate in research and research activities	9

### 6.3.5. Missed areas in survey

The coded answers to the voluntary open question "Are there any topics or areas that we missed in this survey? / Anything else to add?" are displayed in table 10. For this question, responses were fewer and shorter than for the previous open question asking respondents to identify opportunities or challenges for CLS training, with 53 or about 45% of respondents opting to answer. The average answer was ~12 words.

*Table 10. responses to open question about missed areas in survey.*

response_type	frequency	explanation	example
No comment (implicit)	64	Participant refrained from responding	N.A.
No comment (explicit)	23	Participant expressed having no additional suggestions in response	"I found this survey to be clear and complete"
Comment (aspect covered)	14	Participant refers to aspect already covered in survey	"NLP"
Comment (aspect not covered)	16	Participant makes a suggestion about a missed aspect in survey	"A question about target language of the literature studied through CLS"

The fact that 14 comments repeated aspects covered in the survey might point to ineffective communication on the researchers' part about how the different skills in the matrix act as umbrella terms for different kinds of research. It may also indicate a flaw in the survey design, e.g. in the offered amount of information for the participant to process.

For analysis purposes, only the 16 responses (averaging ~16 words each) that covered new aspects were close-read in the same manner as outlined above. Because these responses were shorter, they could all be grouped under one label which produced 16 labeled entries. The five resulting color coded categories are displayed in Table 11.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

*Table 11. Working legend for color coding of responses to open question about missed areas in survey.*

<b>colour_theme</b>	<b>clarification</b>	<b>frequency</b>
Application beyond CLS	Impact of CLS and CLS methods for other disciplines or in an interdisciplinary context	3
Application beyond academia	Linking CLS research to the commercial sector or to pre-university education	2
Deep learning	Deep learning as a distinct area from machine learning	2
Parallel corpora	Multilingual corpora, underrepresented languages, (historical) editions, translations	7
Other	Issues mentioned once, such as user co-creation or social media scraping	2

## 6.4. DARIAH survey consent form

### *Introduction*

CLS INFRA values your privacy and processes your personal data in compliance with the EU General Data Protection Regulation (GDPR).

Your Personal Data is any information related to you. Processing is any operation performed on the data.

According to the Transparency Principle, this document will provide you with information about the processing of your personal data as required by Art. 12, 13, and 14 of the GDPR.

### *Who are we and how can you contact us?*

Contact person: Lisanne van Rossum  
E-mail: [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl);

acting as the data controller within the meaning of the GDPR.

### *For what purpose do we process your data?*

Your data will be processed within the CLS INFRA project (hereinafter: the project).

### *Project description:*

CLS INFRA aims to index the demand for training and schooling in the field of Computational Literary Studies for the purpose of organizing a series of supplementary training schools.

To learn more about the CLS INFRA project, please see <http://clsinfra.io/>.

Please be informed that your data may also be used in different research projects in the domain of Digital Humanities / Computational Literary Studies in accordance with the GDPR.

### *What information about you do we collect and process?*

The following types of information about you are collected and processed within the project unless you decide to opt out of the respective questions:

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

- Gender
- Educational background / title
- Affiliation / professional situation / occupation
- E-mail address

The data are collected directly from you through online survey.

### *Legal basis for the processing of your data:*

Your data is processed on the basis of your consent (Art. 6.1(a) of the GDPR) which you give by accepting this notice.

Exceptionally, where consent is not an appropriate legal basis, your personal data can also be processed on the basis of our legitimate interest in carrying out the project, or further research in the field of Digital Humanities / Computational Literary Studies. Then, the processing is based on Art. 6.1(f) of the GDPR.

### *For how long do we keep your data?*

Your data will be stored for as long as necessary for the fulfillment of the defined research purposes.

### *Will your data be shared with anyone?*

Your personal data will not be shared with or disclosed to anyone outside the CLS INFRA project.

### *Will your data be transferred outside the European Economic Area (EEA)?*

Your data will not be transferred outside the European Economic Area.

### *No profiling or automated decision-making:*

Your data will not be used for profiling or automated decision-making purposes.

### *Your rights with regards to the processing of your data:*

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

The GDPR grants you certain rights with regards to the processing of your personal data. These rights include:

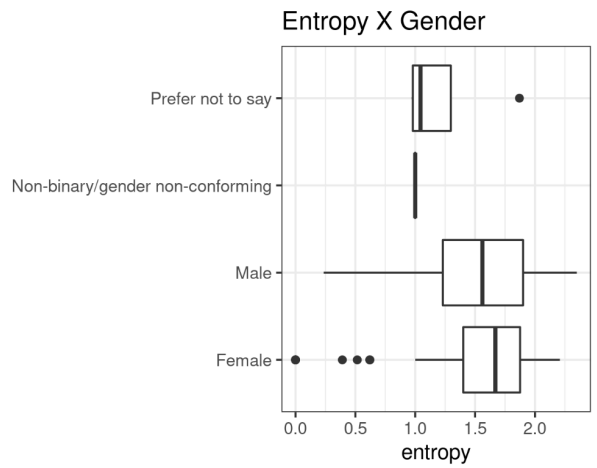
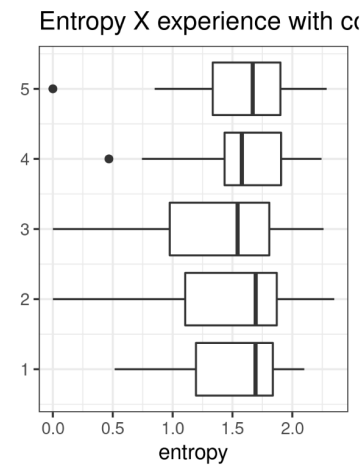
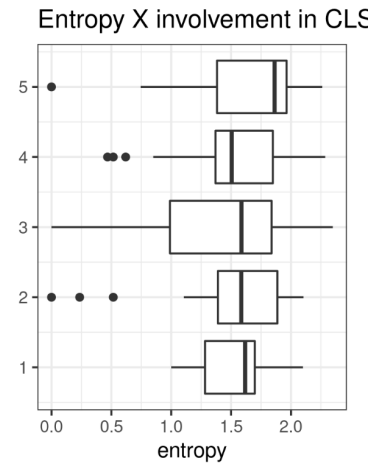
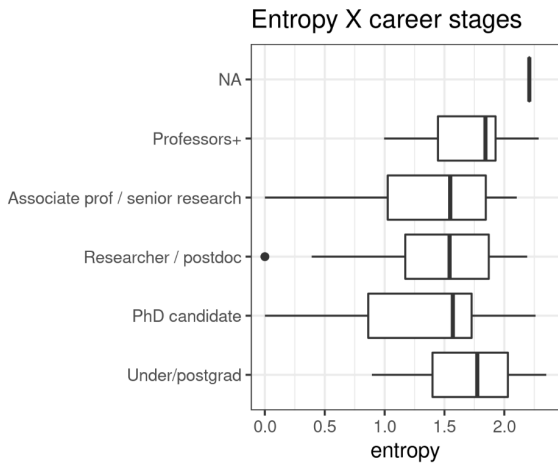
- Access (Art. 15 of the GDPR): you have the right to obtain confirmation as to whether We have your personal data, as well as information about how We process it. You can also request a copy of your personal data, for which We may charge you a reasonable fee based on administrative costs. In order to exercise your right of access, contact us at [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl);
- Rectification (Art. 16 of the GDPR): if your personal data that We process are incomplete or inaccurate, you have the right to request rectification of such data without undue delay. In order to exercise your right to rectification, contact us at [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl);
- Erasure ("right to be forgotten" – Art. 17 of the GDPR): in certain circumstances (e.g. if your data are processed unlawfully or unnecessarily) you may request erasure of your personal data.
- Restriction of processing (Art. 18 of the GDPR): in certain circumstances (e.g. if you contest accuracy of your data that we process or lawfulness of the processing) you may request restriction of processing of your data. Such data will not be erased, but in principle can only be processed with your consent;
- Data portability (Art. 20 of the GDPR): in certain circumstances, you may request transmission of your data to another controller in a structured, commonly used and machine-readable format;
- Right to object (Art. 21 of the GDPR): if you did not consent to the processing, or if it is not necessary to comply with a legal obligation, you may always object to it, in which case We shall no longer process your data.

Moreover, you have the right to:

- Withdraw your consent to the processing of your personal data at any time (Art. 7(3) of the GDPR) by contacting us at [lisanne.van.rossum@huygens.knaw.nl](mailto:lisanne.van.rossum@huygens.knaw.nl). The withdrawal of consent will not affect the lawfulness of processing based on consent before its withdrawal; lodge a complaint with a supervisory authority.

DELIVERABLE 4.1 SKILLS GAP ANALYSIS

## 6.5. Entropy and demographics





DELIVERABLE 4.1 SKILLS GAP ANALYSIS

## 6.6. Distribution of scores per skill



## 6.7. Abbreviation list

API: Application Programming Interface

AŠ: Annotator abbreviation of Artjoms Šeļa

CET: Central European Time

CLS INFRA: Computational Literary Studies Infrastructure (<https://www.clsinfra.io>)

CLS: Computational Literary Studies

CSS: Cascading Style Sheets

DARIAH: Digital Research Infrastructure for the Arts and Humanities (<https://www.dariah.eu>)

DH: Digital Humanities

DHSI: Digital Humanities Summer Institute (<https://dhsi.org>)

DOI: Digital Object Identifier

EC: European Commission

EEA: European Economic Area

ESU (DH): European Summer University (in Digital Humanities) (<https://esu.fdhf.info/>)

EU: European Union

European Union General Data Protection Regulation (GDPR).

FAIR: Four principles of Findability, Accessibility, Interoperability, and Reusability, formulated by Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016).

GIS: Geographical Information System

HTML: HyperText Markup Language

HTR: Handwritten Text Recognition

LaTeX: Markup Language in programming language TeX

LS: Annotator abbreviation of Lisanne M. van Rossum

ML: Machine Learning

## DELIVERABLE 4.1 SKILLS GAP ANALYSIS

NB: Nota bene

NLP: Natural Language Processing

OCR: Optical Character Recognition

Pandoc: Open-source document converter

PhD: Doctor of Philosophy (doctorate)

Postdoc: Post-doctoral researcher

RDF: Resource Description Framework (<https://www.w3.org/TR/rdf-primer/>)

STEM: Science, Technology, Engineering and Mathematics

TaDiRAH: Taxonomy of Digital Research Activities in the Humanities (<http://tadirah.dariah.eu>)

TEI: Text Encoding Initiative (<https://tei-c.org/>)

UK: United Kingdom

US: United States (of America)

VR: Virtual Reality

WP: Work Package

XML: Extensible Markup Language

ZIM-ACDH: Zentrum für Informationsmodellierung (Centre for Information Modelling) -Austrian Centre for Digital Humanities (<https://informationsmodellierung.uni-graz.at/de/>)