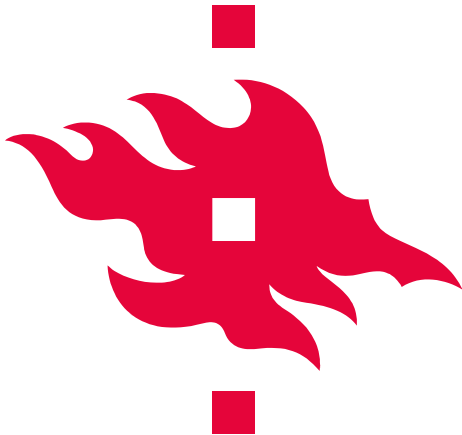# Where are we now?
Current developments in *Positive* Social Sciences related to Text Mining

Krista Lagus

Professor, digital social science

University of Helsinki, Faculty of Social Sciences

Center for Social Data Science (CSDS), co-founder, former director (2019-2021)

Lecturer in:
Perspectives to Digitalizing Society &
Cocreative methods of collaboration in societal challenges

25.3.2022

Realization: The POWER of knowledge and of what we do comes from doing things together

- **Large-scale text analytics as *enablers* of**
    - humanity's peace-making (e.g. Peace Machine – book by Timo Honkela)
    - As showing *ourselves:* who we are in relation to one another / in the context of each other. (e.g. Paths of Wellbeing –article, by Krista Lagus & al, or our Loneliness article that shows the variation of "Us" and the individual responses as part of that variation)
    - As facilitating collaboration rather than other objectives
    - As facilitating our own awareness of us, and of each other, respectfully of each others' privacy needs as well
    - For facilitating "hearing each of us better"
    - For clarifying confusions in communication and understanding better "what do we mean when we say X in different situations"
    - For identifying the positive micro-interactions and enabling it to be noticed & reinforced
- **The uniting, inspiring objective: Creating together a society that is good for all of us**
    - **Part of inventing "Positive social science"**
    - Many concrete projects and activities in this direction are taking place now
    - PhD students and researchers are involved and the community is growing

A currently active research project:
Crowdsourcing an organisation's Code of Conduct (1/2)

- Faculty of Social Sciences arranges for its staff, students and alumni/networks
- **Motivation:** Two cases (reported in the national media Helsingin Sanomat) from students experiencing negative interaction, both from other students and some of the staff (Fall 2021) which led the Faculty to take action to formulate a proposal for a "Code of conduct".
- **Schedule:** Late spring 2022 + data analyses afterwards
- **Platform:** Structured crowdsourcing environment used often by state/communal etc. national organizations, https://otakantaa.fi
- **The participation** to the discussion will not be controlled (to allow easy access, & because it cannot be really controlled, who "belongs" to the entire target group)
- **Assigned moderators** will be involved for the duration of the data collection period (about 2 weeks), to ensure that there is no down-spiraling (in case of risks materialize)
- **Discussions** on the platform will be later **downloaded and analyzed** (statistics, text analytics)
- **Possibility to include a questionnaire** as part of the design of the crowdsourcing data collection process
- **Possibility to design** the platform feel & suitable tools & processes for this small-quantitative-text-data-analysis task

- Intended outcomes

  - **Engagement with the proposal:** People will read, digest and take a stand concerning a proposal for a "Code of Conduct" for the entire community.

  - **Faculty will understand, how the proposal of a Code of Conduct is received.** Is there acceptance/criticism? Can people adopt this one, in their own lives? Are there proposals for improvements? Are there other matters that will come to surface, and need organizational attention?

  - **The entire community will begin to discuss with each other** about how should we act to gether. What is a good way to *be* with one another?

  - **Enough data will be produced** to enable text analytic (quantitative) research on it & obtain results of it

- Possible other outcomes (risks, opportunities)

  - **Internal upheaval in organization** if something even heavier comes to surface

  - **Trolling/intereference by external forces** that are not even part of the community (targeted attemtps at *disrupting* this good-intended process).

  - **Bad publicity (**press, media)

  - The pilot works as an example for national state/communal organizations on opening up their internal discussion, and possibly uncovering the issues that are currently under the rug but affect negatively everyone

  - The analysis tools developed will be transferable to the National Infrastructure (FIN-CLARIAH), to be used by everyone

FIN-CLARIAH:
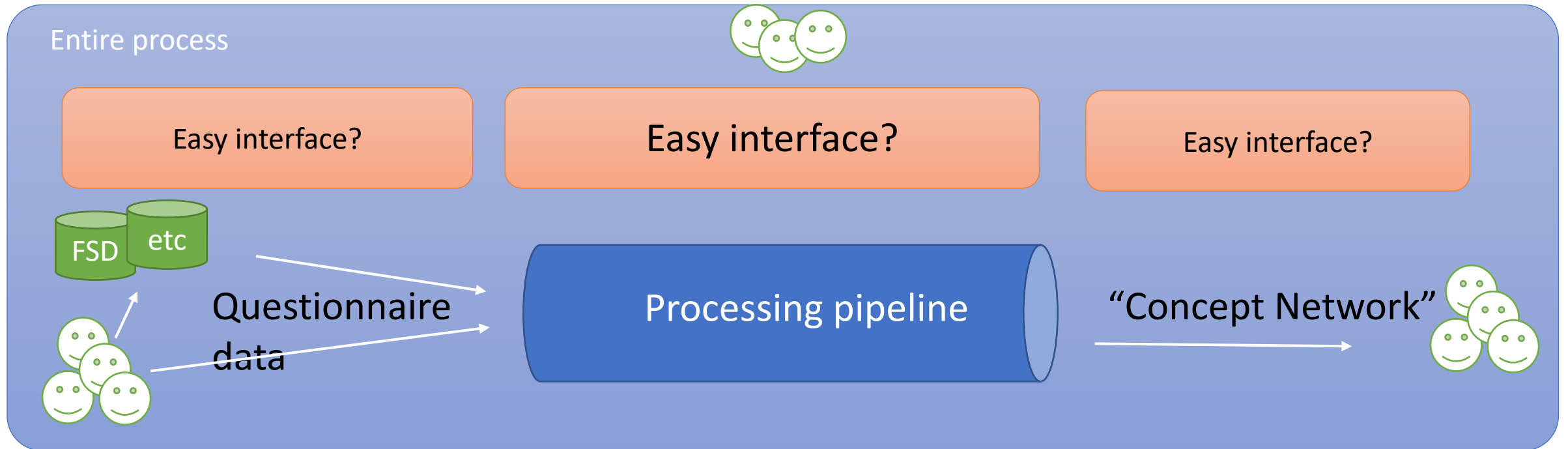# A National infrastructure for social siences and humanities

Will develop **tools and pipelines** that will benefit researchers in the analysis and handling of **big and heterogeneous data** sets

- In development for Social sciences and Humanities research

- Comprised of CLARIN + DARIAH https://www.kielipankki.fi/organization/fin-clariah

- Long-term: 10-year roadmap https://www.kielipankki.fi/organization/roadmap/

- Related to **many data types**: e.g. text, speech, videos, questionnaire data, linked data, gaming data, as long as there is metadata or text associated with it

- Will open **data pipelines from national memory organizations**

- Researchers will be able to **open their** own **data** for *active use*

- Services and data pipelines will be operated at the CSC

- **A two-year funded project** starting now (2022-2023): DARIAH-part in the center

Krista Lagus

Faculty of Social Sciences

WP3.3. Analysis of Questionnaire (e.g. Survey) Open answers

Objective: An interactive processing pipeline with an easy-to use interface, where a researcher can insert their questionnaire data, which contains one or more open questions, and obtain something interesting and valuable in terms of research as an outcome.

LEAD: prof. Krista Lagus, UH CSDS

Entire process

Easy interface?

Easy interface?

Easy interface?

FSD   etc

Questionnaire data

Processing pipeline

"Concept Network"

What is a "Concept network"?

1. **Summarizing analyses concerning central terms in the data** (e.g. sentiment analysis, frequencies of terms etc)
2. **Linking terms and concepts to one another, as well as to the information in other parts of the questionnaire structure (the questions) or the overall data set obtained (**metadata, clusters, statistics of answers)