*Data Service Infrastructure for the Social Sciences and Humanities (DASISH) brings together all 5 ESFRI research infrastructure initiatives in SSH area. The goal of DASISH is to determine areas of cross-fertilization and synergy in the infrastructure development and to work on concrete joint activities related to data, such as data access, data sharing, data quality, data archiving and legal and ethical aspects.*

# Workflows of text analysis

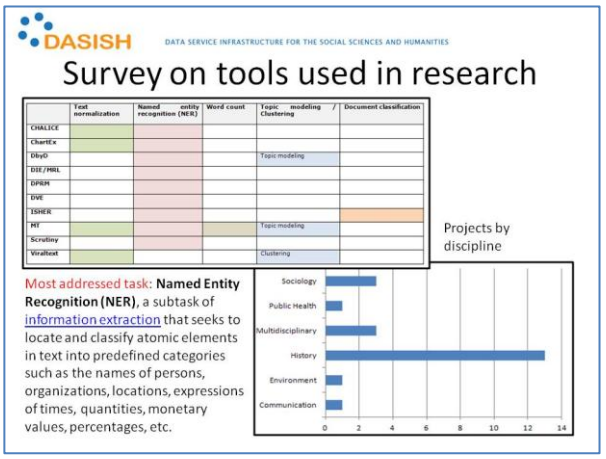
## Background and Description

One of the aims of the DASISH project was to identify typical cross-disciplinary workflows candidate for being dealt by automatic processing chains, study the requirements and implement a number of demonstration cases.

Social Sciences and Humanities (SSH) research has used computers to assist text analysis work since the time of punch cards. The more recent irruption of terms like Digital Humanities, Computational Social Sciences, Culturomics or Big Data Humanities, Arts, and Social Sciences is a further evidence of the interest in text analysis tools in fields such as linguistics, literature, psychology, political science, economics, scientometrics and bibliometrics, sociolinguistics, history, management, education and communication. Although there is a certain variation in how these disciplines refer to what they do ("text analysis", "distant reading", "content analysis", "text mining" or "text analytics"), after some analysis it is clear that they are all referring to the extraction of information from texts with the assistance of software tools.

## Findings & Developments

We first conducted a thorough survey of research papers and project descriptions, with the objective of identifying the kind of software tools that are common to the different SSH disciplines. We have proposed the found common tools as a typical automated e-Research workflow for scholars working with texts. Once identified, DASISH can now offer a discipline-neutral typical workflow, deployed as a web service-based web application, for demonstration purposes. Eventually, this demonstration has been used to ask researchers about requirements for future deployment of tools to support their workflows.



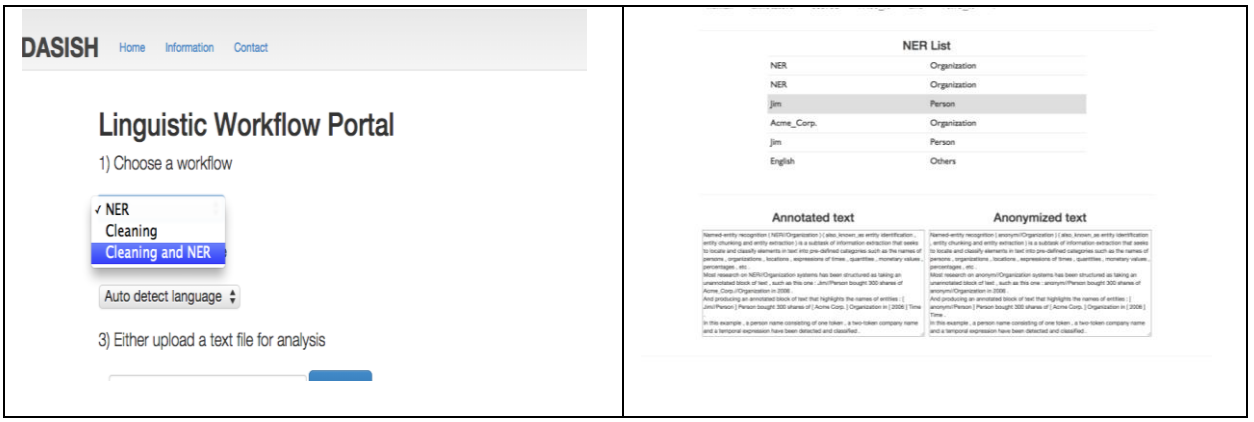

## Cross-disciplinary Workflow

Research in SSH very often involves text analysis to find evidence in terms of particular words that appear in texts. For instance, proper nouns identify entities that can be plotted in maps when they are geographical locations, or counted differently if correspond to male or female for gender related queries. The occurrence and frequency of other types of words can contribute to assess public opinions, to trace events through time, etc. When large quantities of text have to be studied, the use of automatic means becomes a necessity. Research involves the compilation of the texts, conversion to a suitable format and character encoding, cleaning of non linguistic elements, segmentation and tokenization to identify words, and the application of tools that recognize the particular type of sought words.

---

* Deliverable 5.5 of WP5: WORKFLOWS. Available at: http://dasish.eu/publications/projectreports/

# Workflow Demonstrator




Two views of the Demonstrator Web Page at http://dev.dasish.eu:8080/workflows

**Named-entity recognition** (NER) **tools** are able to identify proper nouns and other expressions in text that have a unique reference and classify them into pre-defined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, etc. The raw results are as in the sample below, where Named Entities found in a text are marked with "//TAG". Recognized entities are tagged as follows: person or tag `NP00SP0`, place or tag `NP00G00`, organization or `NP00O00`, and others or `NP00V00`.

```
Toppar listan gör Kapstaden//person i Sydafrika//place som får 8.43 poäng av 10
möjliga . Öriket Maldiverna kvalar in som tvåa med 8.33 poäng medan österrikiska
Zermatt//place knipер tredjeplatsen med 8.29 .
```

Sample of NER tool output for a Swedish text

NER tool input text can be any utf8 unformatted text file. If your text is a pdf, html or rtf file, first use the "Clean" tool or the "Clean+NER" option. The CLEANING workflow accepts PDF, HTML, RTF and flat text in 17 languages. The output is tokenised – taking account of abbreviations – and segmented in sentences, and then realized as flat text or encoded in TEI P5 (Text Encoding Initiative) standard format.

```
<?xml version="1.0" encoding="UTF-8"?> <TEI xmlns="http://www.tei-c.org/ns/1.0"
xmlns:schemaLocation="http://www.tei-c.org/ns/1.0
http://dkclarin.dk/schemas/tei/TEIDKCLARIN_ANNO/xml.xsd">
<teiHeader type="annotation"><fileDesc><titleStmt><title>dasish.txt,Flat text to CBF
converter</title><sponsor>DKCLARIN</sponsor><respStmt><resp>a_annotation</resp><name
>cst.ku.dk<note type="method">flat2cbf</note><date
when="20141126"/></name></respStmt></titleStmt><publicationStmt><distributor>johan@c
stt.dk</distributor><idno type="ctb">20141126-1422-step2</idno><availability
```

Sample of a TEI header as provided by the CLEANING tool

The workflows are not only run without any human intervention, but also automatically assembled from a pool of registered tools when a request is received, using the file format and language of input of output as constraints.

A more detailed description of the Dasish Workflow demonstrator can be found in the Deliverable 5.5 of WP5, available at: http://dasish.eu/publications/projectreports/