Data Service Infrastructure for the Social Sciences and Humanities

EC FP7
Grant Agreement Number: 283646

Deliverable Report

Deliverable: D5.3
Deliverable Name: Workflow Requirements and Application Report
Deadline: M36
Nature: Report

Responsible: Núria Bel (UPF)
Work Package Leader: Daan Broeder (MPI)

Contributing Partners and Editors:

Ben Companjen (DANS)
Bart Jongejan (UCPH)
Georgi Khomeriki (DANS)
Núria Bel (UPF)
Marc Poch (UPF)
Marion Wittenberg (DANS)

SEVENTH FRAMEWORK PROGRAMME

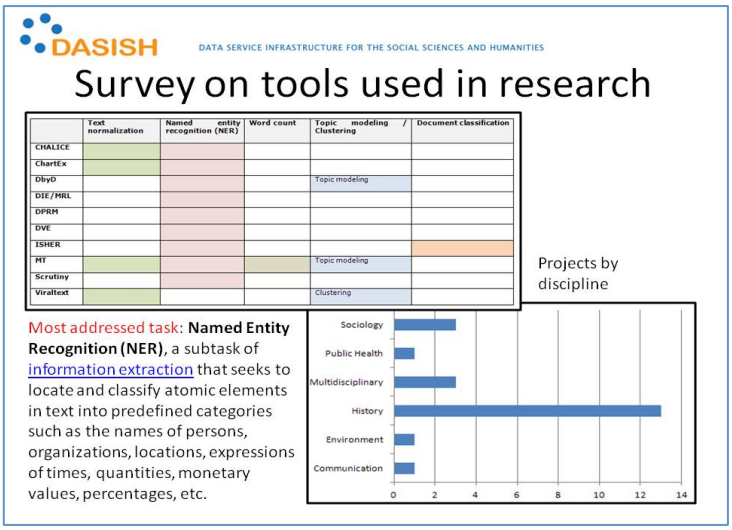
Contents

1. Executive Summary

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.

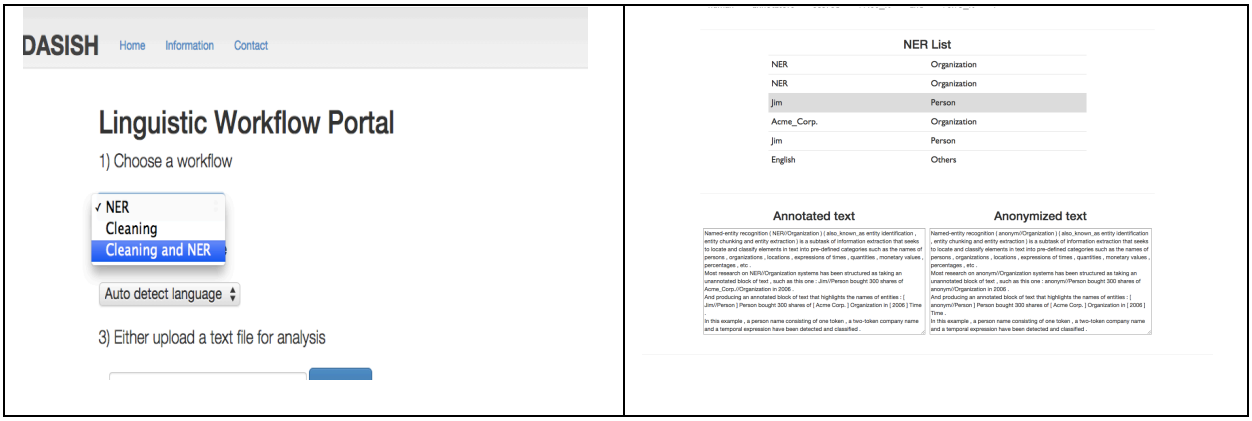
We first conducted a thorough survey of research papers and project descriptions, with the objective of identifying the kind of software tools that are common to the different SSH disciplines. We have proposed the found common tools as a typical automated e-Research workflow for scholars working with 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.



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. This is the origin of the deployed workshop for Named Entity Recognition (NER).

Research also 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. Therefore, WP5.5 has also deployed a workflow that processes text in order to allow processing them.



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.

2. User requirements and use cases

2.1 Use cases survey

In the context of European Research Infrastructures for Humanities and Social Sciences, the DASISH project, Data Service Infrastructure for the Social Science and Humanities, brings together all five ESFRI1 research infrastructure initiatives for the social sciences and humanities (SSH): CLARIN, DARIAH, CESSDA, ESS and SHARE with the aim of identifying areas of cross-fertilization and synergy in the infrastructure development for all five communities.

In this framework, WP5.5 conducted a survey to identify the kind of software tools that are common to different SSH disciplines as to propose them as typical automated e-Research workflows for scholars working with texts. We carried out a survey looking at project descriptions and webs and published papers. Once these tasks were identified, DASISH WP5.5 deployed a number of discipline-neutral typical workflows as web application services for them to be used for training and demonstration purposes.

The survey is annexed to this report. It was conducted by collecting from different information sources and references were found with web searches and, in some cases, from social media channels like Twitter or content curation platforms like Scoop.It. For projects, we browsed the corporate website and grant programs of the principal organizations and initiatives that support Social Sciences and Humanities research: the European Union Seventh Framework Programme (FP7) from the European Commission, German Federal Ministry of Education and Research –BMBF- (Germany), the Joint Information Systems Committee –JISC- (United Kingdom), National Science Foundation –NSF- (United States), Economic and Social Research Council –ESRC-(United Kingdom), National Endowment for the Humanities –NEH- (United States), Social Sciences and Humanities Research Council – SSHRC-(Canada), Digging Into Data Challenge-DIDCH-(international initiative), etc. For publications, we consulted Google Scholar and Scopus databases, review articles, and occasionally Google free searches.

[1] **European Strategy Forum on Research Infrastructures,**
http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=home

The survey, and the annexed report, focused on current practices in computer assisted text analysis in a number of Social Sciences and Humanities research works. We have reviewed 22 project descriptions and 105 publications, we have identified the tasks that motivated researchers for using the different tools and we shortlisted a number of popular tools and how they are used. The analysis of the tasks and tools assisted us as to propose typical automated workflows. We also report on the topics addressed in these researches to show the broad range of their potential application. We focussed in tool combination as we were looking for typical workflows.

The report was organized as follows. In section 2 a number of projects and publications are analysed to identify the use of tools for particular tasks and their combination. We also refer to the general objectives of the research carried out in order to give examples of the actual motivation given the interests of a particular discipline. In section 3, we list and describe the tools that have been mentioned in these works when they were general purpose, that is, not embedded in a particular application meant to give support just to a particular analysis. For instance, many papers referred to "content analysis" methodology which uses quantitative analysis as provided by word frequency counts. For instance, Kirilenko et al. (2010) sum up this view:

"The most frequent words are essentially variables in the further statistical analysis, and the number of these variables should be large enough for a meaningful analysis and, at the same time, be several times smaller than the number of available textual units (Kline 1994). Following (Iker and Harway 1965), the articles were divided into textual units in a range of 1,750–2,000 symbols (about two paragraphs) to improve the case/variable ratio; this increased the US sample to 299 cases and the UK sample to 589 cases. The most frequent words, identified by interactive CATPAC–WORDER procedure, were used to construct a customized dictionary for counting relevant concepts in each textual unit."

In section 4, more information about currently available catalogues of generic tools is supplied. This is to find out the availability of tools for other languages than English. DASISH has to provide tools for the processing of different European languages. Conclusions are at section 5. The analysis of the data showed that indeed there are typical chains of processes that can be performed in general tools and for different languages. The most frequent task is Named Entity Recognition and this is our candidate for proposing a typical workflow dealing with written text.

2.2 User evaluation

To ascertain whether the developed chaining would be a useful service, we interviewed a number of researchers and other professionals. During these interviews the demonstrator was used as a reference to investigate the need for easy access language tools for the interviewee's work and to evaluate the functionality of the demonstrator itself. In the interviews we asked about the kind of work or projects the interviewee is engaged in, the issues for which he or she is using, or could use language tools. We inquired after the need for specific tools and demonstrated our workflow chaining.

We interviewed 5 persons. The background of the interviewees was divers; researchers as well as support staff, within the context of the social sciences, humanities as well as

history. Five interviewees were positive about our demonstrator. One interviewee did not feel the need for such a service, as she doesn't expect more than already is possible with ATLAS-ti (atlasti.com), a commercial tool she uses for her research.

The kind of tools the interviewees would like to see in such a chaining were Named Entity Recognition (NER), Named Entity Disambiguation (NED), Topic modelling and the harmonisation and conversion of dates. Two interviewees expect that NER could be used for anonymisation or pseudonimisation. For this latter use it is very important the tool is very precise. One interviewee mentioned the possibility to share the evaluation results of a specific tool; this would help to estimate the exactness of a specific tool for a similar dataset.

A Web application is seen as very useful. Interviewees would prefer this to a tool that has to be downloaded and used offline. One interviewee made the point that a user web interface is not useful with large datasets, in some cases an API would be more convenient.

Another remark that was made is that the results of the workflow should not be open available, as it is at the moment, but secured.

The summaries of the interviews can be found at the annex II.

3. Implementation

3.1 Selected use cases

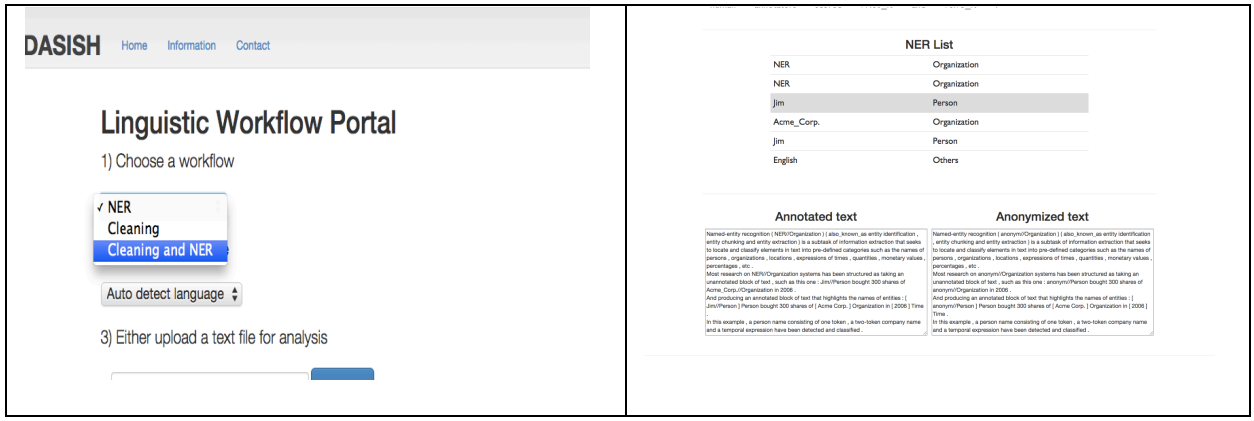
Two use cases have been selected to be used as models: 1) Text cleaning and 2) Named Entity Recognition.
Both use cases have been chosen because they are widely used and needed tasks for the expected DASISH users.

Text cleaning is a basic task before being able to process most textual data. Tools being used to process data require concrete data formats, specific character encodings and other specifications. Data formatting and encoding is one of the main cause for problems for NLP tools. Having a system to clean textual data can a great benefit for DASISH users.
Once the textual data is cleaned it is ready to be processed. There is very varied set of processes that can be of interest for DASISH users that can be applied to cleaned textual data. NER has been chosen because it is widely used in data mining, information extraction, corpus exploitation, etc.

3.1.1 Named Entity Recognition use case

The NER Demonstrator aim is to cover a considerable amount of languages. Typical NER software tools are limited to one or just a few languages therefore more than one NER tool are used.

The demonstrator has a language selector that user can use to specify the input language or make use of the automatic language detection system. When the language is set, manually or automatically, the system can choose the adequate NER tool.



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

To cover a large variety of languages a survey on available NER tools has been carried out. Requirements for selecting a tool were: (i) API or open sourced-free software available; (ii) different languages available to reduce the number of deployments and (iii) quality of the results.

The following NER tools were selected:
- **FreeLing** (English, Catalan, Spanish, Portuguese, Italian, Galician, Asturian, Welsh)
- **Stagger** (Swedish)
- **Clarin.dk NER** (Danish)
- **DBpediaSpotlight** (English, German, Dutch, French, Italian, Russian, Spanish, Portuguese, Hungarian, Turkish)

Table 1 shows the language coverage provided by the selected tools together. The 4 NER tools provide coverage for 15 languages.

Language	FreeLing	Stagger	Clarin.dk NER	Spotlight
Asturian	X			
English	X			
Catalan	X			
Danish			X	
Dutch				X
English				X
French				X
Galician	X			
German				X
Hungarian				X
Italian	X			X
Portuguese	X			X

Russian				X
Turkish				X
Spanish	X			X
Swedish		X		
Welsh	X			

Table 1

3.1.2 Cleaning use case

Cleaning is done with the workflow engine at clarin.dk. The accepted input formats are PDF, HTML and RTF. The output is tokenised and segmented in sentences and can either be realized as flat text or as TEI P5. The workflow performs best for those languages for which lists of abbreviations are made available to the tokeniser/segmenter: Asturian (ast), Catalan (ca), Czech (cs), Welsh (cy), Spanish (es), Galician (gl), Italian (it), Portuguese (pt), Slovene (sl), Norwegian (nb,nn), Danish (da), German (de), English (en), French (fr), Icelandic (is), Dutch (nl), and Russian (ru).

The following tools and utilities are used: pdfminer, html2text, sed, ascii2uni, UCPH's RTFreader, and UCPH's flat text to TEI P5 converter. All text is output as UTF-8. If the input isn't UTF-8, it is automatically converted to UTF-8 from UTF-16 (LE,BE), UTF-32 (LE,BE) or UCS-2 (LE,BE). If the input character set is an 8-bits code, it is assumed to be ISO-8859-1 and converted from that to UTF-8, so this will cause problems for input in one of the other ISO-8859 character sets or an other 8-bit character set altogether. The conversion to UTF8 is done with UCPHs make UTF8 utility.

Because of the great number of workflows (3 input formats, 2 output formats, 17 languages, together 85 different workflows), workflows are not constructed on beforehand, but assembled by the workflow engine itself when a cleaning request is received, based on the full specification of the input (file format, language) and of the output (file format, language). So the workflows are not only run, but also constructed without any human intervention. (Tools are integrated in clarin.dk by specifying their input and output features, together with some boilerplate like the name of the tool and the URL of the webservice that wraps around the tool. The workflow engine does all the rest.

3.1.3 Cleaning and NER workflow

The combined workflow of text cleaning and named entity recognition must accept files or text in various formats, in a user interface that any researcher can use. The results should be presented in a meaningful way to the user. Naturally, the workflow engine must take care of routing cleaned text to the correct NER service.

A web application suits these requirements. The implementation of this web application is described in section 3.2.5.

3.2 Implementation

To implement the selected use cases and other similar scenarios, different tools must be chained. In this case in particular, tools will be deployed and run in different locations; therefore a distributed architecture is required. Web Services (WS) are software systems designed to support interoperable machine-to-machine interaction over a

network. WS can be used to make a tool deployed in a server publicly available to be run by other tools.
Once the needed tools are deployed as WS and can be run remotely by another program is time to combine those tools to create workflows. To summarize, there are two main design decisions to implement the desired workflows: 1) which WS system to use (which protocols, software, etc.) and 2) which workflow system to use to call those WS (which system, software, etc.)

3.2.1 Web Services: SOAP vs REST APIs

SOAP (Simple Object Access Protocol) is a XML-based communication protocol for accessing a web service, created to communicate over HTTP (which is today supported by all internet browsers and services). SOAP is platform and language independent, simple and extensible. SOAP may also be used over HTTPS (which is the same protocol as HTTP at the application level, but uses an encrypted transport protocol underneath) with either simple or mutual authentication.
A SOAP message is an XML document containing:

- an *Envelope* element that identifies the XML document as a SOAP message and constitutes the root element; it contains the *namespace* attribute (which defines the envelope as a SOAP envelope) and the *encodingStyle* attribute (defining the data types used in the document);
- an optional *Header* element, containing application-specific information (like authentication, payment, etc) about the SOAP message;
- a *Body* element, that contains the actual SOAP message (call and response information);
- an optional *Fault* element containing errors and status information; it must be a child of the Body element and it can appear only once in a SOAP message.

Although using SOAP over HTTP allows for easier communication through proxies and firewalls than previous remote execution technology, the technique has the disadvantage of using an application level protocol (HTTP) as a transport protocol (critics have argued that abusing a protocol by using it in a different purpose may conduct in sub-optimal behaviour).
The following table shows a short summary of the advantages and disadvantages of SOAP protocol.

Table 2: Pros and Cons of SOAP

Pros and Cons SOAP	
Pros	**Cons**
Language, platform, and transport independent	Conceptually more difficult, more "heavy-weight" than REST
Designed to handle distributed computing environments	More verbose
Is the prevailing standard for web services, and hence has better support from other standards (WSDL, WS-*) and tooling from vendors	Harder to develop, requires tools
Built-in error handling (faults)	
Extensibility	

REST is an architectural style for distributed hypermedia systems such as the World Wide Web. The term was introduced in 2000 in the doctoral dissertation of Roy Fielding [Fielding 2000], who also participated in the IETF [Internet Engineering Task Force] working groups on URI, HTTP and HTML. The systems which follow REST principles are called RESTful.
In short, the basic REST principles are:

- Application state and functionality are abstracted into resources**;** all types of documents can be used as representations for resources (XML, XHTML, HTML, PNG, ...)

- Every resource is uniquely addressable using a universal syntax for use in hypermedia links (URI –Uniform Resource Identifier)

- All resources share a uniform interface for the transfer of state between client and resource, consisting of a constrained set of well-defined operations (represented by the GET, POST, PUT and DELETE methods) and a constrained set of content types (optionally supporting code on demand);

- The transfer protocol is client-server, stateless, cacheable and layered.

A RESTful web service is a simple web service implemented using HTTP and the principles of REST. Some advantages and disadvantages of REST are listed in the following table.

Table 3: Pros and Cons of REST

Pros and Cons of REST	
Pros	**Cons**
Language and platform independent	Assumes a point-to-point communication model--not usable for distributed computing environment where messages may go through one or more intermediaries
Much simpler to develop than SOAP	Lack of standards support for security, policy, reliable messaging, etc., so services that have more sophisticated requirements are harder to develop ("roll your own")
Small learning curve, less reliance on tools	Tied to the HTTP transport model
Concise, no need for additional messaging layer	
Closer in design and philosophy to the Web	

3.2.2 WORKFLOWS: chaining APIs vs complex Workflow engines

The necessary tools needed to implement the selected scenarios need to be chained in order to obtain the desired results. Workflows are chains of tools, WS, APIs, etc. that represent those chains.

When tools are deployed on a single machine workflows can be represented with “pipes”. Pipes let user chain simple and complex tasks or using the command line. Pipes can also be made within a program (Java, Python, etc.) using some libraries or special calls that allow that program to run the tools and manage inputs and outputs.

There are frameworks designed to integrate and create chains of tools (usually called modules inside the framework). UIMA (https://uima.apache.org/), GATE (https://gate.ac.uk/), etc. let the user develop and integrate modules and chain them to create complex workflows.

On the other hand, when working with tools distributed remotely command line pipes cannot be used. However, workflows can be designed in a program using now other libraries which are specially made for calling remote APIs. There are frameworks specially designed to chain and run web services (Taverna(Hull et al. 2006) editor and engine).

To summarize, in our scenario of NLP tools distributed remotely, there are two options: a) develop a program to chain the tools or b) install and make use of a framework. The first option requires a smaller learning curve, is more flexible (not bounded by a framework) and requires less resources (man hours). On the other hand is less scalable when the number of tools and workflows grow. The framework option requires a considerable amount of resources to install and deploy and a considerable learning curve for developers. It is bounded by the framework specifications and limitations. There are fewer resources left for usability, etc. However, it is a more scalable solution and in some cases already developed modules can be found.

Taking into account the resources in hand and the DASISH end users we have decided to use the first option: developing a usable web interface and a program to orchestrate the workflows behind it. This way, the efforts are put on deploying more tools, more services, covering more languages and usability.

3.2.3 Interoperability

Interoperability is guaranteed at two levels:

1) APIs are deployed using Common Interfaces
2) Data conversion tools are deploy within the APIs

When a tool is deployed as a web service (making its API public) the web service provider is free to do it at will. However, if a few guidelines are followed Interoperability can be achieved more easily. For example, for NER tools deployed we can define a Common Interface (CI) that specifies that mandatory parameters should be *input* is simple plain text and the *language*. If all deployed web service follow this CI the integration of future tools to the system will be much easier.

Different tools use different data formats. For the NER scenario this happens more often for the output data than for input. We can consider that in most cases input will be plain text. On the other hand, NER tools present the results in a varied data formats. To keep interoperability data conversion tools or scripts are deployed within the APIs to guarantee that at least one output is provided with a common data format (some APIs provide the same output in different formats) following a CI specification.

3.2.4 NER APIs

The final list of selected tools accessible via APIs is:

- FreeLing
- Stagger

- clarin.dk NER
- DBpediaSpotlight

Clarin.dk NER and DBpedia Spotlight are already deployed web services and their APIs can be accessed. The demonstrator can simply use Java REST libraries to access these APIs.
On the other hand, FreeLing and Stagger are tools that are made to run locally on a server or laptop. They do not provide a web service or public API by default. To make these tools available for the demonstrator software developed under CLARIN is be used: CLAM (http://proycon.github.io/clam/): Computational Linguistics Application Mediator is software to easily deploy NLP tools as RESTful web services. CLAM makes the necessary call to tools, scripts, data conversion tools, etc. to make the desired API accessible for the DASISH demonstrator.

3.2.5 Web Application

To demonstrate the functionality of the tool chain(s), a web application was created that accepts a text from a user and feeds it to the tool chain based on goal selection by the user and language of the text. The application was designed to be easy to use, by involving the user only for the necessary input and leaving the rest to the application's backend.

The demonstration is available on the public web (via [1]), in source code form (via [2]) and as a Docker image (via [3]). The source code publication and Docker image allow anyone to setup the web application on a private machine or somewhere else on the public Internet.

[1]: http://dev.dasish.eu:8080/workflows/
[2]: https://github.com/DASISH-T55/Workflows
[3]: https://registry.hub.docker.com/u/bencomp/dasish-workflows

The frontend to the workflow engine was developed in Java 7 using the [Wicket] library for web applications. The application by default runs inside the embedded [Jetty] application server.

[Wicket]: http://wicket.apache.org/
[Jetty]: http://www.eclipse.org/jetty/

To build the web application from source and run it, a Java 7 JDK, Maven 3 and Git need to be installed.

The web application is a front end to a lightweight workflow controller. A user can submit text by uploading a file or typing or pasting text into the form. Uploaded files can be in PDF, RTF or HTML format, or plain UTF-8 encoded text. Text input in the form can be plain UTF-8 encoded text or HTML.

The user can select to apply NER, cleaning or both to the submitted text. Optionally, the user can select the language of the text from a list of supported languages. Before the workflow is started, the user needs to enter an email address to receive a notification when the workflow finishes.

After selecting the options and text to process, the workflow engine in the web application selects and executes a workflow. Depending on the selected options and format and language of the input, the workflow steps include file format conversion, language detection or named entity recognition.

The result of the workflow is presented in an HTML page in the following ways:

-The original text is rendered and dynamically overlaid with entity type annotations
-The recognized named entities are listed in a table of word or phrase and attributed entity type
-The text is rendered with inline entity type annotations
-The text is rendered with type annotations, but with entity labels replaced by "anonym".

Following REST principles, the workflow components are stateless. The demonstration web application therefore needs to keep track of the execution state of each submitted job and make sure the file formats of input, intermediate and output files match the required file format of each tool and that the correct protocol is used to communicate with each tool.

Managing connected services

Adding or editing the available services can be easily done by editing the application code, as each service is wrapped within a method consisting of a few HTTP calls. Including other services or updating current services therefore means (copying and) editing a few lines of code. A downside of this approach is that a new build is needed when service information changes.

Protocol usage

The protocol chosen to interact with the workflow components is REST (as discussed above). This means that each component accepts input via an HTTP POST command and responds with an HTTP response code:

- 200 OK: the input was accepted and processed immediately. The result is in the body of the response.
- 201 Created: the input was accepted and a resource was created. The response contains the Location of the result.
- 202 Accepted: the input is accepted, but processing is expected to take longer than the normal response time. The Location header points to a URI where the result will be eventually. The requester (i.e. the workflow engine) will poll the given location to get the result when it is available.
- 404 Not Found: the resource that is the output of the workflow is not yet available, or a resource is requested that does not exist.
- 4xx: the input could not be accepted for processing, for example because the file format cannot be processed by the component.
- 50x Error: the server encountered an error processing the request. The load might be too high, or something else went wrong.

Following these semantics for using HTTP to communicate requests and responses between the workflow engine and each workflow component, adding or replacing components is relatively easy, depending on the requirements for file formats and contents.

Limitations

In the released version of the web application, the following limitations are present:

- when a text is uploaded that was written in an unsupported language, the workflow engine recognises the language but does not communicate this to the user via the web interface.
- when "Cleaning" is picked as the workflow to execute, the result is still put through the applicable NER service.
- when a Danish text is input, the user will receive two emails instead of one, because the Danish tools always send an email too.
- output from various NER services is not presented in a uniform way, because of different tool outputs are kept.

References

D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn. 2006. "Taverna: a tool for building and running workflows of services." Nucleic Acids Research, vol. 34. Web Server issue, pp. 729-732.

Annex I: Computer assisted text analysis: survey on Social Sciences and Humanities current research

Table of contents

Table of tables

Table of figures

1. Introduction

Social Sciences and Humanities research has used computers to assist text analysis work since the time of punch cards in fields such as linguistics, literature, psychology, political science, and communication (Roberts 1997, Popping 2000, for a survey). Nevertheless, the more recent irruption of terms like Digital Humanities (Burdick et al 2012; Liu 2012), Computational Social Sciences (Lazer et al. 2009), Culturomics (Michel et al. 2011; Bohannon 2011) or Big Data Humanities, Arts, and Social Sciences (Parry 2010; King 2011; Leetaru 2012) is the evidence of a recent rise of interest in text analysis tools that seems to be related to the adoption of data-based methods in a broader range of disciplines that include economics, scientometrics and bibliometrics, sociolinguistics, history, public health, management, and education. Although there is a certain variation in how these disciplines refer to what they do ("text analysis", "content analysis", "text mining" or "text analytics", depending on the researcher background), after some analysis it is clear that they are all referring to the extraction of information from texts with the assistance of software tools. The increasing availability of large amounts of digitized texts and recordings (including text, audio, video, etc.) and the interest of their analysis (quite often by means of new visualization displays) has motivated the use of software tools as it became clear the high cost of manually handling and analysing large quantities of data. These tools are based on methods used in Natural Language Processing (NLP), information extraction, text mining, and machine learning (O'Connor et al. 2011; Brier and Hopp 2011; Wiedemann 2013) and are available for a number of languages. In section 4 of this report, we provide information on available registries and catalogues where information about these tools can be found.

Currently, these tools are considered as basic research infrastructure in initiatives such as CLARIN, LanguageGrid and Bamboo that pretend to disseminate their existence as well as to promote their use. In the context of European Research Infrastructures for Humanities and Social Sciences, the DASISH project, **Data Service Infrastructure for the Social Science and Humanities,** brings together all five ESFRI [2] research infrastructure initiatives for the social sciences and humanities (SSH): CLARIN, DARIAH, CESSDA, ESS and SHARE with the aim of identifying areas of cross-fertilization and synergy in the infrastructure development for all five communities. In this framework, the objective of this survey is to identify the kind of software tools that are common to different SCH disciplines as to propose them as typical automated e-Research workflows for scholars working with texts and speech recordings. Once identified, DASISH wants to

[2] **European Strategy Forum on Research Infrastructures,** http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=home

offer a number of discipline-neutral typical workflows as services for training and demonstration purposes.

In the following sections, we report on a survey[3] on current practices in computer assisted text analysis in a number of Social Sciences and Humanities research works. We have reviewed 22 project descriptions and 105 publications, we have identified the tasks that interested to researchers for using the tools and we list a number of popular tools and how they are used. The analysis of the tasks and tools will assist us as to propose typical automated workflows. We also report on the topics addressed in these researches to show the broad range of their potential application. We focussed in tool combination as we were looking for typical workflows. Therefore, in section 2 a number of projects and publications are analysed to identify the use of tools for particular tasks and their combination. We also refer to the general objectives of the research carried out in order to give examples of the actual motivation given the interests of a particular discipline.

In section 3, we list and describe the tools that have been mentioned in these works when they were general purpose, that is, not embedded in a particular application meant to give support just to a particular analysis. For instance, many papers referred to "content analysis" methodology which uses quantitative analysis as provided by word frequency counts. For instance, Kirilenko et al. (2012:507) sum up this view:

> "*The most frequent words are essentially variables in the further statistical analysis, and the number of these variables should be large enough for a meaningful analysis and, at the same time, be several times smaller than the number of available textual units (Kline 1994). Following Iker (Iker and Harway 1965), the articles were divided into textual units in a range of 1,750–2,000 symbols (about two paragraphs) to*
>
> *improve the case/variable ratio; this increased the US sample to 299 cases and the UK sample to 589 cases. The most frequent*

[3] All our information sources were found with web searches and, in some cases, from social media channels like Twitter or content curation platforms like Scoop.It. For projects, we browsed the corporate website and grant programs of the principal organizations and initiatives that support Social Sciences and Humanities research: the European Union Seventh Framework Programme (FP7) from the European Commission, German Federal Ministry of Education and Research –BMBF- (Germany), the Joint Information Systems Committee –JISC- (United Kingdom), National Science Foundation –NSF- (United States), Economic and Social Research Council –ESRC-(United Kingdom), National Endowment for the Humanities –NEH- (United States), Social Sciences and Humanities Research Council – SSHRC-(Canada), Digging Into Data Challenge-DIDCH- (international initiative), etc. For publications, we consulted Google Scholar and Scopus databases, review articles, and occasionally Google free searches.

words, identified by interactive CATPAC–WORDER procedure, were used to construct a customized dictionary for counting relevant concepts in each textual unit."

In section 4, more information about currently available catalogues of generic tools is supplied. This is to find out the availability of tools for other languages than English. DASISH has to provide tools for the processing of different European languages.

Conclusions are at section 5. The analysis of the data showed that indeed there are typical chains of processes that can be performed in general tools and for different languages. The most frequent task is Named Entity Recognition and this is our candidate for proposing a typical workflow dealing with written text.

2. Survey of projects and publications

In this section, we comment on a sample of projects and publications that are reported to have utilised computer assisted text analysis. This survey is not exhaustive. Its aim is to understand what is currently used and how in SCH research.

Projects and publications refer to tools that basically help researchers to identify, annotate, extract and visually localize relevant information. Visualization can be at the same text, with annotation tools that highlight words or sequences of words, or with specific mapping tools. The simplest case is to identify place names in texts and to display them in a map.

2.1 The projects

We have analyzed 22 projects that reported to have used computer assisted text analysis. These projects are (a full description can be found at the Annex I.):

CHALICE (2010-2011), whose aim was "*to provide a historic place-name gazetteer covering a thousand years of history, linked to attestations in old texts and maps*", used Named Entity Recognition (NER) techniques to extract new data from a digitized English Place Names Survey.

ChartEx (2011-in course) by using NLP technologies, NER among others, is working on the extraction of information about places, people and events in medieval charters to know more about domestic life in Middle Age.

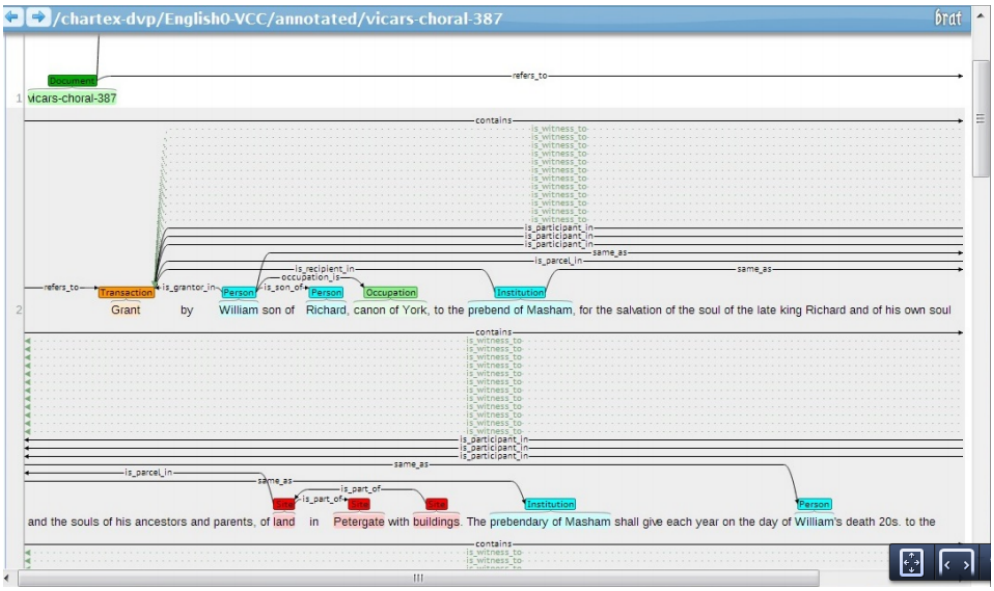


Figure 1. Example of ChartEx NLP processing[4]

DIE/MRL (2010-2011) worked on the exploitation of 18th-century letters to study details about people, places, times and their relationships in big amounts of data. New ways to visualize the data were also proposed.

DbyD aims *"to uncover and represent the argumentative structure of digitized documents".* It is using Topic Models to analyze specific topic areas from data provided by digitized full-texts books, bibliographic databases of journals, and comprehensive reference works. The Sci2 tool[5] "*a modular toolset specifically designed for the study of science*" was employed for implementing several analyses (temporal, geospatial, topical, network) and visualization.

DPRM (2013-2014) is studying the applicability of FOAF (Friend of a Friend) model for describing relationships among Renaissance musicians in a database like a new biographical tool.

DVE (2010-2011), whose aim was "*to identify and track topics about the Greco-Roman world as they appear in more than a million documents produced across thousands of years and in several languages*", used text alignment techniques and proposed scalable NER for the identification of people and places in texts.

[4] http://www.chartex.org/docs/Chartex-Leeds-04072013-HarrisCahill.pdf
[5] https://sci2.cns.iu.edu/user/index.php

ISHER (2012-2013) is researching the application of tools to detect, link, and visualize events, trends, people, organizations, and other entities of interest to social history. Text mining-based rich semantic metadata extraction for collection indexing, clustering and classification is the main focus with the aim of reducing the manual costs currently involved in such activities.

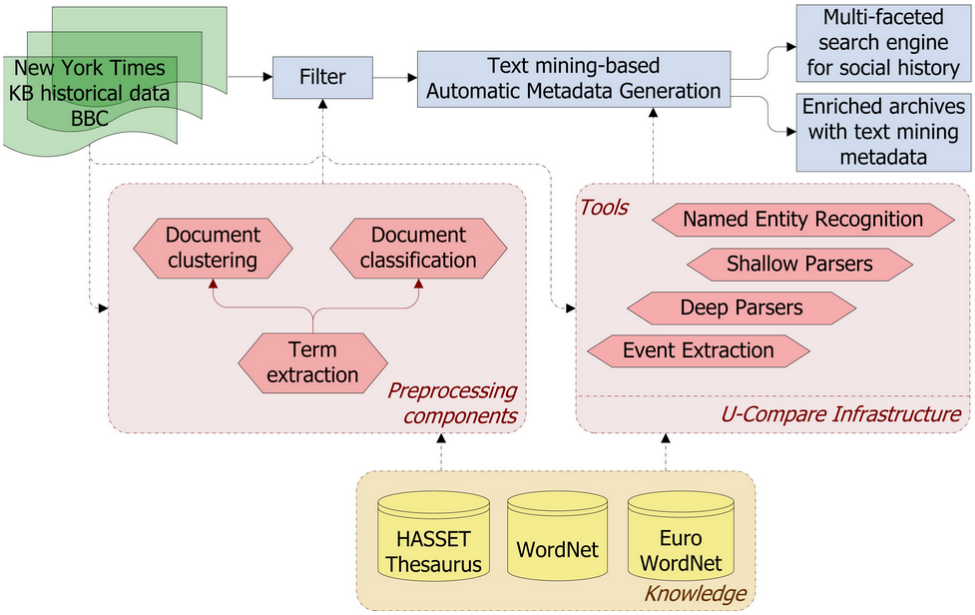


Figure 2. ISHER system architecture[6]

Scrutiny (2009-2010), whose aim was "*to increase the speed and efficiency with which* researchers are able to locate potentially relevant information within large data objects such *as journal articles or full-text datasets*", has developed a Firefox extension, named Scrutiny[7], that scans the web pages selected by individual users and highlights entities that it thinks will interest them.

[6] http://www.nactem.ac.uk/DID-ISHER/ISHERsystemArchitecture.jpg

[7] This tool is also integrated in the tool compilation section of this survey [section 3].

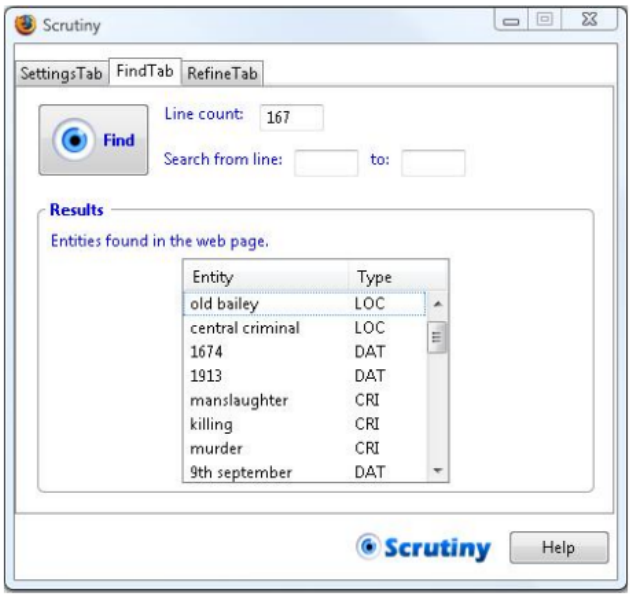


Figure 3. Example of Scrutiny name entity recognition[8]

MT, Mapping Texts (2010-...) "*whose goal has been to develop a series of experimental new models for combining the possibilities of text-mining and geospatial analysis*". It is using topic modeling to identify meaningful language patterns in the analyzed newspaper collection with MALLET[9] toolkit.

Viraltext (2013-2014), whose objective is to trace "*How ideas —literary, political, scientific, economic, religious— circulated in the public sphere and achieve critical force among audiences*". The project is using n-gram indexing to improve the similar fragment search and alignment techniques (as the ones used for Machine Translation) for the detection of reprinted text fragments on Nineteenth-Century American Newspapers. In text analysis process this project proposes *"a simple greedy agglomerative clustering with a "complete link" heuristic. We will experiment with hand-constructed clusters to set the appropriate level of sensitivity, e.g., to join a cluster, any text must align with at least 75% of each of the other texts".*

In order to analyse the data, the projects were sorted by discipline: Environment, History, Public Health, Sociology, and we refer some as Multidisciplinary because they address a combination of disciplines like History of science, social network analysis, cognitive sciences, and digital humanities (CIS project), humanities and sciences (DbyD project), or literary studies and history (SH project).

[8] http://www.hrionline.ac.uk/scrutiny/prototypes/scrutiny_help_19-01-2010.pdf

[9] http://mallet.cs.umass.edu/

As shown in the figure 4, where the distribution of projects by discipline is shown, the discipline that is better represented is History with 13 projects.

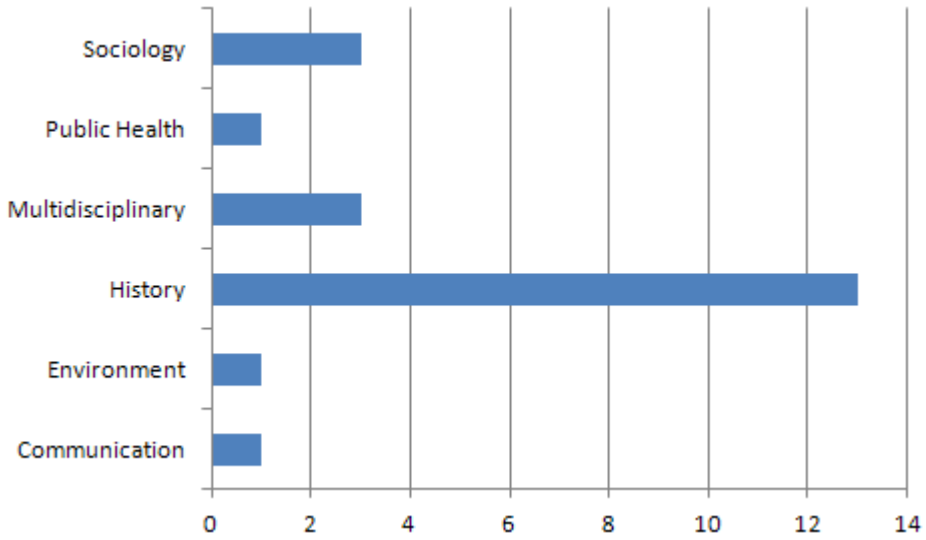


Figure 4. Projects by discipline

As for their temporal location, the oldest one (SciPer project) was developed between 1999 and 2007, while the others are dated from 2009 on. Most of them were funded by governmental institutions. They cover a variety of digital text types: books, journals, historical press, transcribed interviews, law trials, grant proposals, conference proceedings, social media posts, web pages, letters, charters, written communications, and reports.

Concerning the methods used, more than a half of the projects used tools to extract information to analyze language patterns, identify relationships, create content visualization, or geo-referencing. The other projects were concerned with topics related to what is called discourse analysis: metaphor use, topic representation, etc.. In general, they used computational methods describe, explore, search, link, or visualize texts.

Table 1 presents a summary of concrete tasks performed by these tools, as reported by projects, although we could find this information about only ten of them.

	Text normalization	Named entity recognition (NER)	Word count	Topic modeling / Clustering	Document classification
CHALICE					
ChartEx					
DbyD				Topic modeling	
DIE/MRL					
DPRM					
DVE					
ISHER					
MT				Topic modeling	
Scrutiny					
Viraltext				Clustering	

Table 4. Projects by developed tasks

Nine of ten projects performed Named Entity Recognition[10] that is indentifying sequences of words that are names of locations, organizations, persons, etc. Text normalization was carried out in four cases to improve the quality of digitized texts. Text normalization implies handling text encoding issues, identifying non-textual elements such as formulae, etc.

Topic modelling was used in three cases. It is a statistical method that discovers topics mentioned in documents from words occurring together consistently. Word counting was used in one case. The task of counting the occurrence of particular words was reported in one case. Usually is it carried out after a pre-process that includes identifying different inflected word-forms of a particular lemma, in the task called lemmatization, and the identification of the part-of-speech of the chosen lemma. Finally, document classification was utilised only in one occasion.

The tools used for performing these tasks were the following. These tools are further described in section [3].

- *Named entity recognition: Brat* [CharTex project], Stanford NER [MT project], Sci2 [DbyD project], Scrutiny [Scrutiny project]
- *Topic analysis:* MALLET [MT project], Sci2 [DbyD project]

[10] It is interesting to note that the named entity recognition task implies a previously workflow where others tasks are developed like tokenization, sentence splitter, lemmatization, and POS tagging or stemming.

- *Visualization:* Google Maps [MT, DIE/MRL and SH projects], Google Finance[11] time series [MT project], Simile widgets [MT project], Protovis [MT project])
- *Text analysis and visualization:* Voyant tools [DMCI Project], Sci2 [DbyD project], or
- *Text management and storage*: Zotero [DMCI Project].

In Table 2, we list the tools that were used or created to implement these tasks for the seven projects that we could find information about.

Projects	**Tools by tasks**				
	NER	**Topic modeling**	**Visualization**	**Text analysis and visualization**	**Manage and storage text**
ChartEx	Brat				
DbyD	Sci2	Sci2		Sci2	
DIE/MRL			Google Maps		
DMCI				Voyant tools	Zotero
MT	Stanford NER	MALLET	Google Maps, Google Finance,Simile widgets, Protovis		
Scrutiny	Scrutiny				
SH			Google Maps		

Table 5. Project tools by tasks

[11] The Google tools did not consider in compilation tools of this survey because they are generic solutions.

Finally, we want to highlight that six of the projects were concerned with exploring new text content visualization in different project's phases: DbyD, MT, DIE/MRL, DMCI, PBMP, and SH:

DbyD project used visualization for mapping the links between disciplines and sub disciplines.

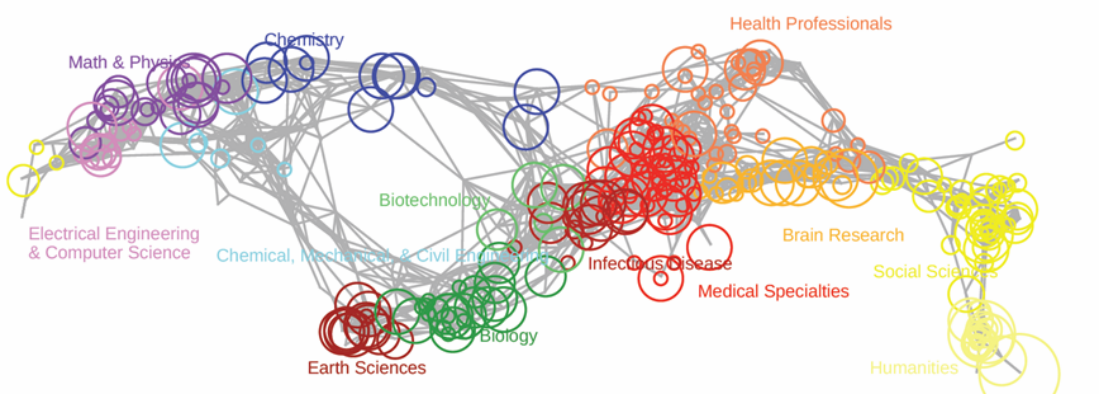


Figure 5. Example map visualization[12]

MT project proposed geospatial visualization to explore language patterns embedded in historical newspapers collection.

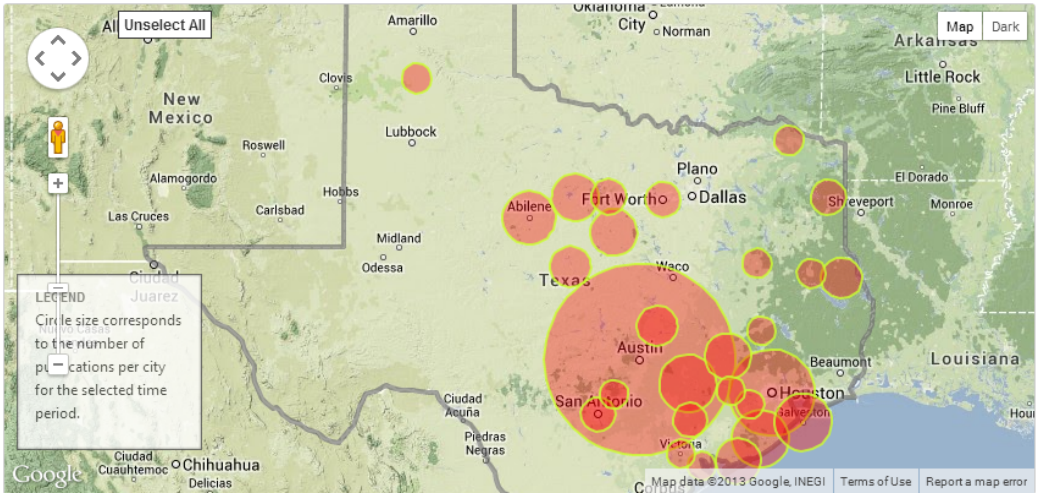


Figure 6. Example Google Map visualization[13]

DIE/MRL project used visualization for exploring and examining information about people, places, times, and relationships on 18th-century letters corpus. This project wanted "*to demonstrate how visual analysis tools can help us to generate new*

[12] http://cns.iu.edu/docs/presentations/2013-allen-digging-montreal.pdf

[13] http://language.mappingtexts.org/

knowledge through methods rooted in humanities scholarship using annotation capabilities and the ability for scholars to insert new data and resolve existing data"[14].

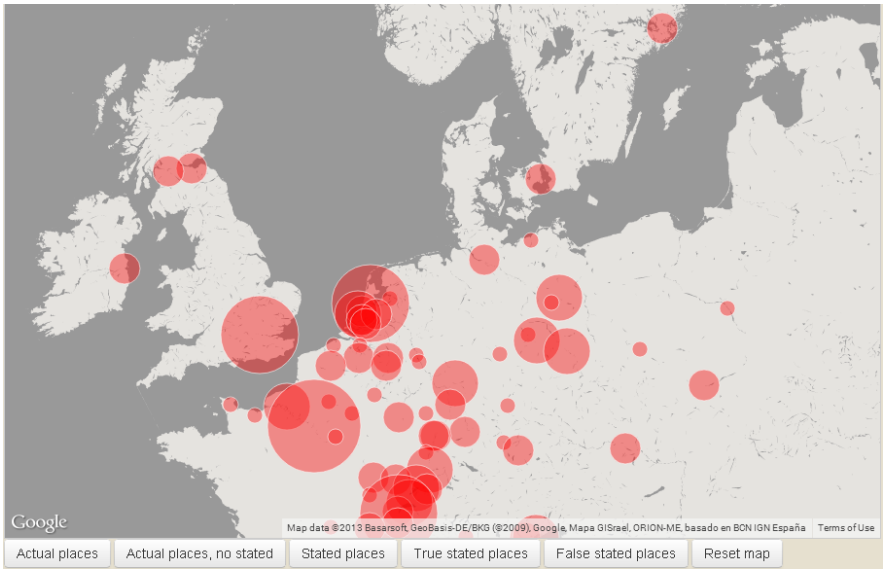


Figure 7. Example Google Map visualization[15]

DMCI project worked on digital research environment, including text analysis tools integrated in the suite called Voyant tools[16]. This suite of tools proposes text visualization through word cloud, frequencies, Kwic index, word distribution, etc.

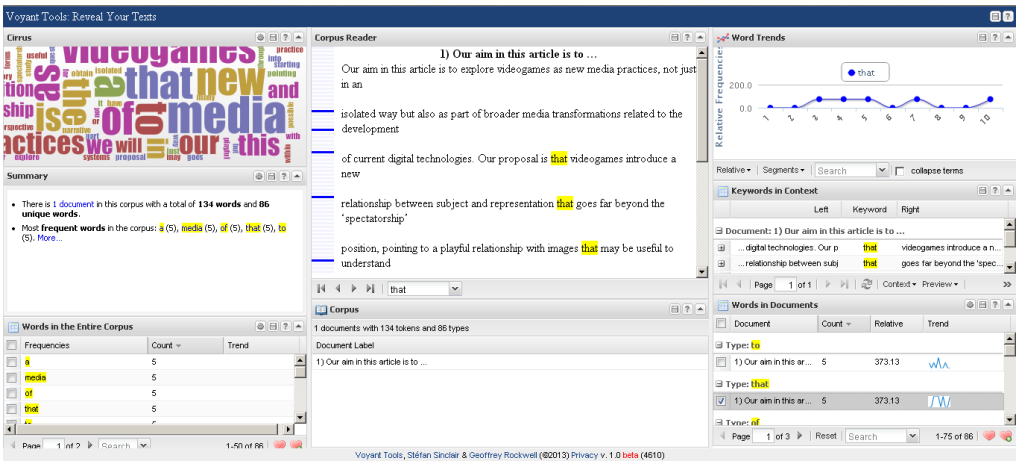


Figure 8. Example Voyant tools visualizations[17]

[14] http://enlightenment.humanitiesnetwork.org/

[15] http://republicofletters.stanford.edu/casestudies/voltairepub.html

[16] Tool suite developed in the framework of the project Heremeneuti.ca: The Rhetoric of Text Analysis http://hermeneuti.ca/

[17] http://voyant-tools.org/

PBMP project has integrated two resources in an online user interface. The first resource is a database of citations and full-text repository about the ancient city of Pompeii. The second resource is a Geographical Information System (GIS) map of the historical city.

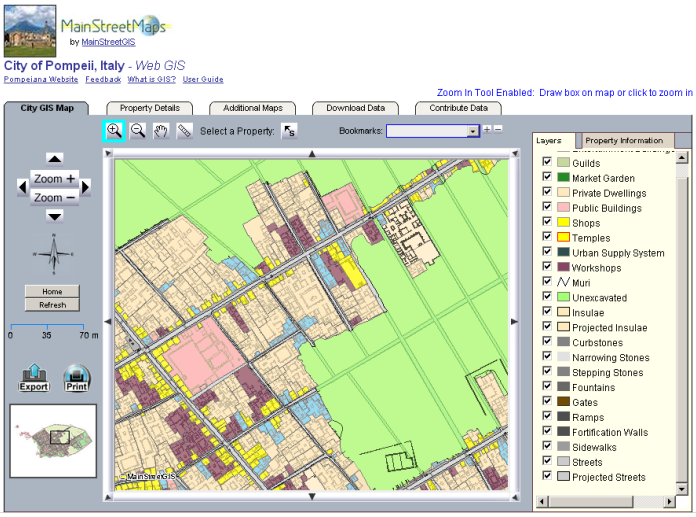


Figure 9. Example GIS Map visualization[18]

SH project considers that an effective way of text analysis in Digital Humanities is the application of Geographical Information Systems (GIS). This project proposed exploratory surveys in literary studies and history.

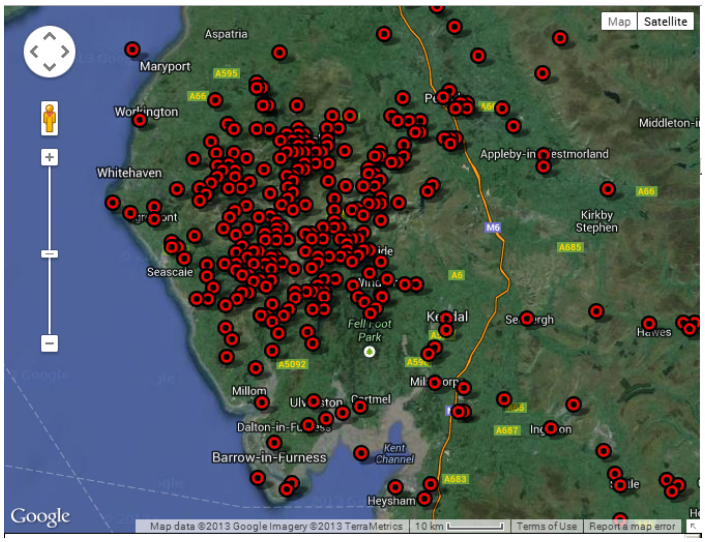


Figure 10. Example Google Map visualization[19]

---

[18] http://www.mainstreetmaps.com/ITALY/Pompeii/

[19] http://www.lancaster.ac.uk/mappingthelakes/v2/lit_file.htm

2.2 Publications

After reviewing the projects, our following step was to review Scopus and Google Scholar databases to search for publications addressing computer assisted text analysis in Social Sciences and Humanities. Also, we reviewed the work of O'Connor et al. (2011) that offer an excellent overview on papers reporting having used computer assisted text analysis in Social Sciences. Our search was restricted by two criteria: dates and keywords. Firstly, we limited our search to papers published from 2009 to the present. Secondly, we used keywords like "textual analysis", "textual data", "quantitative concept analysis", "information mining", "text analytics", "computer-aided text analysis", "automatic text analysis", "computer-assisted text analysis", "quantitative content analysis", "entity recognition", and "entity analysis", all in combination with terms like Social Sciences, Humanities, and names of disciplines (History, Political Science, Public Health, Economics, Psychology, etc.). Our survey is not exhaustive, we found 105 articles, conference papers, and reports referring to the topics mentioned above. We eventually selected 43 works for reporting in this survey. The selection was based on the information provided in papers, because we were interested in what and how was used but also in the availability of the tools for other purposes.

From the surveyed papers, we could see that Social Sciences and Humanities applied computer assisted text analysis for different purposes depending on the disciplines. This is an overview by disciplines:

In **Economics**, a popular trend is to perform sentiment analysis, that it to extract information from text with text mining techniques to determine the positive or negative attitude of a writer with respect to some topic. The application was with respect to different economical concepts like "stock market" (Gilbert 2009; Bollen et al. 2010), "financial volatility" (Kogan 2009), or "consumer opinions" (Archak 2011; Netzer 2012). Some studies worked on statistical prediction methods for economic behaviour (Askitas and Zimmerman 2009), or box-office revenues for movies (Asur and Humberman 2010; Joshi and others 2010). The used texts were Twitter posts, corporate 10-K reports, critics or user opinions; all of them were internet texts.

Another application, Bollen and others (2010:1), researched "*whether measurements of collective mood states, derived from large-scale Twitter feeds, are correlated to the value of the Dow Jones Industrial Average (DJIA) over time*". The authors employed the

Google-Profile of Mood States algorithm and the OpinionFinder[20] tool to explore sentiment analysis.

In **Psychology**, text analysis software tools are mostly used for studies about person's mental and affective state as manifested in their language. In this field, one of the most used tools is LIWC[21], a commercial text analysis software that calculates the degree to which people use different categories of words across a wide array of texts -emails, speeches, poems, Twitter posts, etc.- (Tausczik and Pennebaker 2009; Golder et al. 2011).

In the field of **Library and Information Science**, most frequently Scientometrics or Bibliometrics surveys analyze influential topics (Gerrish and Blei 2010; Ramage 2011) or paper's citations (Yogatama 2011; Bethard and Jurafsky 2010), but recently other topics seem to rise interest, for instance Morillo et al. (2013) explore a semi-automatic method to identify institutional addresses from bibliographic databases to produce bibliometric indicators. All of these works applied word counting and statistical packages to analyse data from digitized collection of texts.

In **Sociolinguistics**, tags used in social media have opened new possibilities for analyzing linguistic uses. Eisenstein et al. (2010 and 2011) used Twitter geographic tags for studying geographic linguistic variation. They used a Twitter tokenizer named TwittMotif[22] "*to preserve emoticons and blocks of punctuation and other symbols as tokens*".

In **Public Health** domain, some studies looked at how queries and social media posts help detecting and measuring issues like seasonal epidemics (Ginsberg 2009; Culotta 2010; Paul 2011). MacEachren et al. (2011a and 2011b) proposed Twitter mapping tools for implementing potential applications for crisis management. These authors developed a geovisual analytics application (SensePlace2[23]) that extracts place-time-attribute information from Twitter. This application loads Twitter texts, processes the text to extract particular information (including NER with ANNIE tool), makes a georeferenced analysis, and finally indexes the texts.

[20] http://mpqa.cs.pitt.edu/opinionfinder/

[21] h http://www.liwc.net/

[22] http://tweetmotif.com/about

[23] http://www.geovista.psu.edu/SensePlace2/

Berchialla et al. (2010) and Hamon and Grabar (2013) report on different information extraction use cases. The first use case works on Public Health Surveillance on child injuries from newspapers clipping. They tested and discussed three information extraction techniques: NER approach (with ANNIE tool), a rule based system (with VisualText tool), and a machine learning method based on Support Vector Machines. The second use case designs a system to extract occurrences of medication names and related information from narrative clinical documents. These authors use NLP and text mining tools to extract relevant information. The tools are integrated in a web environment named Ogmios platform "that facilitates communication between them, making the management of linguistic and semantic annotations easier" (Hamon and Grabar 2013: 553). In the processing phase, this project applied linguistic and semantic annotation with different modules for NER, word and sentence segmentation, term and semantic tagging (with Genia tool), and term extraction (with Lingua-YaTeA tool).

History research focuses on the exploitation of historical corpora. Horton et al. (2009) propose a Naïve Bayesian classifier trained on labelled articles from the *Encyclopédie* of Denis Diderot and Jean le Rond d'Alembert to determine class membership for the remaining unclassified articles. Later, they applied this model to the *Journal de Trévoux, or Mémoires pour l'Histoire des Sciences & des Beaux-Arts*, another corpus from 18th century, with the aim of increasing Literature 18th century understanding.

Yang et al. (2011) addressed the identification of topics in Texas historical newspapers over a given period (1829 to 2008). This research used two tools: the Mallet toolkit to conduct topic modelling and the Stanford NER to extract information.

Finally, it is worth to mention the white papers of Cohen et al. (2011) and Torget et al. (2011) that report on the results of DCMI and MT projects. Both projects used text mining tools but with different objectives. On the one hand, DCMI project attempted to connect existing digital resources and tools like digitized *Old Bailey Proceedings*, *Zotero* (to store and manage texts), and Voyant toolset (for analysis and visualization) to help historian's daily work. On the other hand, MT project was oriented towards combining text mining with geospatial mapping to study historical newspapers. In this project, the researchers focused "*on three popular metrics among humanities scholar for studying language patterns*" (Cohen et al. 2011: 25): word counts, named entity recognition (with Stanford NER), and topic modeling (with MALLET). The project also worked with Protovis and Simile applications to explore the visualization.

In the field of **Political Science** is also frequent the application of opinion mining techniques to perform sentiment analysis. O'Connor et al. (2010) studied consumer confidence and political opinion (2008-2009) and its manifestation in social media by means of sentiment word frequency counting in Twitter messages. They opted for using a deterministic approach based on prior linguistic knowledge, counting instances of positive and negative sentiment words in the context of a topic keyword. The positive and negative words were defined by the subjectivity lexicon in the OpinionFinder system.

Black et al. (2012) work on the analysis of opinions as stated in tweets. They explore their potential value for assessing people political behaviour by using text mining tools to extract semantic content (people, places, topics and opinions). They proposed a text analysis workflow (with the UIMA platform [24]) that includes sentence splitting, tokenization, PoS tagging, lookup of place names, dictionary lookup and, rule-based phrasal analysis (with the Cafetière platform). Also, it worked with TermMine tool integrated in Cafetière platform for term extraction.

Also, Stephens-Davidowitz (2013) studied how much racial animus influence voting in U.S. presidential election, comparing Barack Obama's 2008/2012 vote shares with John Kerry's 2004 vote share. This study used the percent of Google search queries that include racially charged language in same periods as a data source. In the same way, Metaxas et al. (2011) explored the power of electoral predictions using social media data (Twitter) in several Senate races of the two recent US congressional elections.

Klüver (2009, 2011) studied the influence of interest groups in the European Union. In these works, the author analyzed a number of documents using the Wordfish and the Wordscore programs that developed statistical models of word count. Suzuki (2010) analyzed foreign-policy texts from Japanese prime ministers, the Diet addresses, since 1948 to 2008. He worked using multivariate analysis and machine learning methods, and introduced linguistic knowledge for analysis. During the corpus workbench, Suzuki employed two tools for linguistic analysis: ChaSen for morphological analysis and CaboCha for dependency analysis.

Quinn et al. (2010) and Grimmer (2010) applied a topic modeling analysis to study Senator's speeches and press releases with unsupervised (Bayesian inference) or supervised learning techniques respectively.

[24] http://uima-framework.sourceforge.net/

In **Literature and Linguistics**, computer text analysis is used to address a larger variety of subjects. Some examples are the following.

Argamon et al. (2009) applied text mining techniques to detect language patterns in large corpora to find linguistic evidence of gender, race, and nationality distinctiveness in Black Drama texts since 1950 to 2006. The text mining task was developed with a full text analysis free system called PhiloMine[25] that offers means for data gathering from textual databases, feature extraction from texts and Machine Learning-based text classification.

Elson et al. (2010) extracted social networks from nineteenth-century British novels and serials. They used NER analysis (with Stanford NER tool) to identify characters in a literary text and to detect the existence of social networks.

Bamman and Crane (2011:1) worked on "*a method for automatically identifying word sense variation in a dated collection of [digitized] historical books*" by training and using WSD classifiers in parallel corpora in Latin and English. In word alignment phase they used MGIZA++ program.

Oelke et al. (2012:35) analyzed a subset of the Swedish Literature Bank "*focusing on the extraction of persons' names, their gender and their normalized, linked form, including mentions of theistic beings (e.g., Gods' names and mythological figures), and examined their appearance over the course of [13] novels*". Extraction was done with a slightly adapted NER system [26] for Swedish language. Also, this work proposed a visual exploration of data with techniques such as network representation, summary plots, or literature fingerprints.

van Dalen-Oskam (2012: 360) proposed "to analyze the stylistic functions of name usage in literary texts, wanting to be able to compare the usage and functions of names across texts, oeuvres, genres, time periods, and cultures or languages" using computational assisted text analysis techniques. The author implemented a use case to recognize personal and geographic names based on a Dutch novel "Boven is het stil" (2006) written by Gerbrand Bakker and its English translation "The Twin" (2009), translated by David Colmer. In this study, the author used Stanford NER tool to identify entities and a script for token counting.

[25] http://code.google.com/p/philomine/

[26] The tool name could not be identified in reviewed articles.

There is a number of authors that are applying geospatial analysis to literary texts. Gregory and Cooper (2011:90) explored "*how such digital technology [GIS] might be used to map out literary articulations of geographically located dwelling and spatial mobilities*". Note that they propose to use a GIS tool for critical interpretation rather than a simple spatial visualization. The authors implemented a study related to GIS analysis of letters and diaries by T. Gray[27] and S. Coleridge[28] [29]. Corpora preprocess involved four phases: text digitization, place names identification and tagging (by manual typing), place names georeferencing by coordinates, and GIS conversion of resulting information.

Gregory and Hardie (2011) presented a study that uses corpus methods to link corpus textual content with a geo-referenced database in a GIS system. This survey used "*part-of-speech tagging to extract instances of proper nouns from a corpus, and a gazetteer to limit these instances to those representing place–names, a database of the places mentioned in a corpus can be created, visualized, and analyzed using GIS technology*" (Gregory and Hardie 2011: 297). The corpus was POS tagged with CLAWS software, and USAS tool was used to develop semantic analysis.

This overview of current practices gives an idea about the actual use of software tools to perform SCH focussed studies. Tools are used to process text and to extract information useful for further processing and analysis. It is noticeable that most of the surveyed works use pre-processing tools to clean, normalize and segment texts with the objective of having lists of words occurring in studied texts and most commonly also a word counting for statistical purposes. Quite a number of these works go beyond and extract particular information from texts, most frequently names of persons, locations, organizations, etc. This is the task of the Named Entity Recognition tools. NER is the most frequent task performed as far as our survey reveals (see also Table 1). In the next section we review the actual tools used and their characteristics.

[27] Gray T. 1971.Correspondence of Thomas Gray Toynbee P and Whibley L. ed, rev Starr H W. vol III. Clarendon Press, Oxford.

[28] Coleridge S T. 1956. Collected letters of Samuel Taylor Coleridge Griggs E L. ed vol II. Clarendon Press, Oxford.

[29] Coleridge S T. 1957. The notebooks of Samuel Taylor Coleridge Coburn K. ed vol I. Routledge & Kegan Paul, London.

3. Tools

In general terms, tools for text analysis can be classified into three main classes: tools for annotation, tools for visualization and tools for extracting information. Obviously, these classes can be related. For instance, text annotation allows for a new visualization display and also to extract useful information summing up facts in annotated documents.

Today there are several tools to assist text analysis available for the researchers both commercial and open source. In this section, we only include the tools mentioned by analyzed projects (Table 5) and publications (Table 6), and in the next section we provide details about catalogues and registries of tools. The main goal of reviewing the tools that have actually been used in projects and research work is to assess most addressed tasks as evidence about current interests among researchers of different disciplines. This first overview confirms that the most addressed task is Named Entity Recognition performed after some processing of texts and whose output is mostly used for data visualization.

The tools mentioned in revised projects were: Brat, MALLET, Sci2, Scrutiny, Simile, Stanford NER, Protovis, Voyant tools, and Zotero.

Tool name	**Project name**	**Discipline**	**Task performed**	**Type**
Brat	**CharTex**	History	Annotation (NER)	Used
MALLET	**MT**	History	NER, topic modeling, etc.	Used
Sci2	**DbyD**	Multidisciplinary	NER, topical analysis, etc. / Visualization	Used
Scrutiny	**Scrutiny**	Multidisciplinary	NER	Created
Simile	**MT**	History	Visualization	Used
Stanford NER	**MT**	History	NER	Used
Protovis	**MT**	History	Visualization	Used
Voyant tools	**DMCI**	History	Word count /	Used

			Visualization	
Zotero	**DMCI**	History	Store and manage texts	Used

Table 6. Referenced tools by Project

Brat[30] is a web application to create "*annotations for named entity recognition and binary relations for simple relational information extraction tasks, among others*". It is an open source project developed collaboratively by researchers from National Centre for Text Mining (NacTeM) and University of Manchester (United Kingdom), Aizawa laboratory at University of Tokyo, and Research Center for Knowledge Media and Content Science of National Institute of Informatics (NII) of Japan.

MALLET [31] "*is a Java-based package for statistical natural language processing, document classification, clustering, topic modelling, information extraction, and other machine learning application to text*". This tool suite was developed by Andrew MacCallum (University of Massachusetts Amherst) with contributions from graduate students and staff at University of Massachusetts Amherst, and also contributions from University of Pennsylvania researchers.

The **Science of Science**[32] **(Sci2)** Tool is "*a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of scholarly datasets at the micro (individual), meso (local), and macro (global) levels*". This initiative is supported in part by the Cyberinfrastructure for Network Science Center and the School of Library and Information Science at Indiana University, the National Science Foundation, and the James S. McDonnell Foundation.

Scrutiny [33] is a Firefox extension developed at Scrutiny project. This tool "*will be developed using natural language processing, including `named entity recognition´ based on a Bayesian learning methodology*." Its primary purpose is to increase the speed and efficiency with which HE and non-HE researchers are able to locate potentially relevant information within large data objects such as journal articles or full-text dataset. The Scrutiny project was a collaborative effort between the Old Bailey and Central

[30] http://brat.nlplab.org/

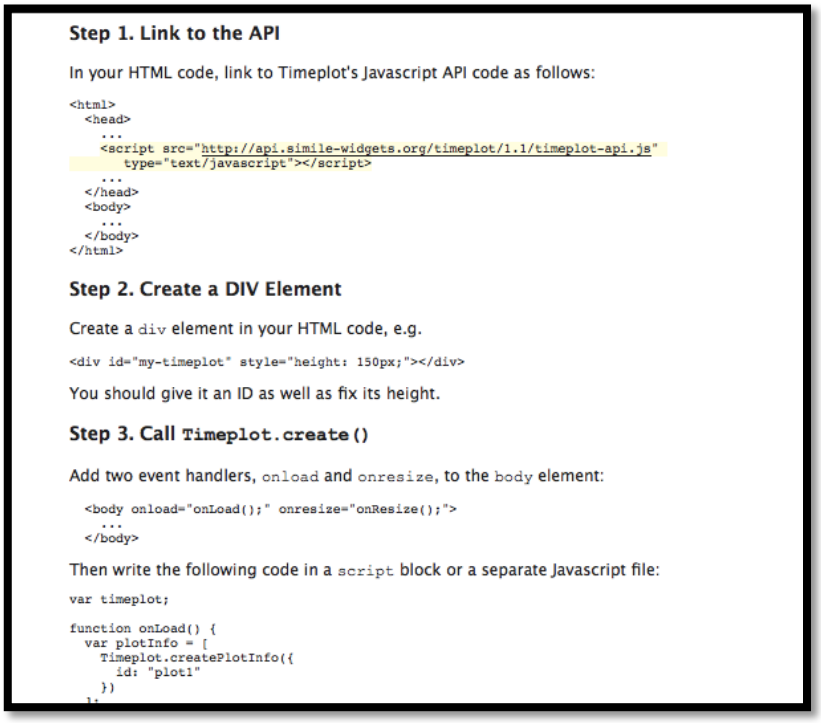
[31] http://mallet.cs.umass.edu/

[32] https://sci2.cns.iu.edu/user/index.php

[33] http://www.hrionline.ac.uk/scrutiny/

Criminal Courts and Plebeian Lives projects, the Humanities Research Institute, and PlayGen (a serious game company).

SIMILE Widgets[34] are a collection of open-source data visualization web widgets. It is composed by four components: Exhibit, Timeline, Timeplot, and Runway. Exhibit supplies the code to create web pages with text search and filter functionalities, with interactive maps, timelines, and other visualizations like flags, bubbles, etc. Timeline widget implements interactive timelines. Timeplot is a DHTML-based AJAXy web application for plotting time series and overlay time-based events over them. Runway is a Flash widget to display images in an interactive visualization similar to the Cover Flow of Apple iTunes. The SIMILE widgets collection is an open source "spin-off" from the SIMILE project developed by MIT Libraries and MIT CSAIL.



Step 1. Link to the API

In your HTML code, link to Timeplot's Javascript API code as follows:

```
<html>
  <head>
    ...
    <script src="http://api.simile-widgets.org/timeplot/1.1/timeplot-api.js"
       type="text/javascript"></script>
    ...
  </head>
  <body>
    ...
  </body>
</html>
```

Step 2. Create a DIV Element

Create a `div` element in your HTML code, e.g.

```
<div id="my-timeplot" style="height: 150px;"></div>
```

You should give it an ID as well as fix its height.

Step 3. Call `Timeplot.create()`

Add two event handlers, `onload` and `onresize`, to the `body` element:

```
<body onload="onLoad();" onresize="onResize();">
  ...
</body>
```

Then write the following code in a `script` block or a separate Javascript file:

```
var timeplot;

function onLoad() {
  var plotInfo = [
    Timeplot.createPlotInfo({
      id: "plot1"
    })
  ];
```

Figure 11. Example of guidelines for using SIMILE widgets

Stanford NER[35] is a Java application to name entity recognition. This tool recognizes entities as person, organization, or location for English, but it also is available for other languages. Stanford NER is a CRF Classifier that "*provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models*". The application is supported by the Stanford Natural Language Processing Group.

[34] http://www.simile-widgets.org/

[35] http://nlp.stanford.edu/software/CRF-NER.shtml

Protovis[36] is a graphical toolkit for visualization developed at the Stanford Visualization Group. This tools suite "*defines marks through dynamic properties that encode data, allowing inheritance, scales and layouts to simplify construction*." It offers a broad gamma of visualizations such as traditional graphics (area, bar cand colums charts, scatterplots, etc.), custom views (bullet charts, sparklines, Anderson's Flowers, etc.), interaction views (parallel coordinates, pan and zoom, tooltips, etc.), hierarchies (dendograms, treemaps, circle packing, etc.), among others.

Voyant tools[37] (also known as Voyeur tools) is a web-based text analysis environment developed in the Hermeneuti.ca[38] project. This project was a collaborative effort of Stéfan Sinclair (McGill University) and Geoffrey Rockwell (University of Alberta). These tools were used to perform the study of frequency and distribution of data and the visualization in different formats (tab separated values, Kwic index, word cloud, etc.) by embedded tools.

Zotero[39] is a tool to collect, organize, cite, and share research sources. It was produced by Roy Rosenzweig Center for History and New Media at George Mason University. Zotero is well know like bibliographic manager but at DMCI project was employed like research environment to storage and manage texts.

It is noticeable that only in two cases, projects provided web-based applications to operate the tools: Brat and Voyant tools. The rest require, in different degrees, downloading, installing the tools. In these cases, there are different degrees of user familiarity with software programming, but in most cases, using the tool implies some kind of command line-based participation of the user. We will comment more on that later.

The tools[40] or platforms used or created in revised articles were: ANNIE, CaboCha, Cafetière web platform, ChaSen, CLAWS, Genia, Lingua-YaTeA, LIWC, MALLET,

[36] http://mbostock.github.io/protovis/

[37] http://voyant-tools.org/

[38]Hermeneuti.ca – The Rhetoric of Text Analysis: http://hermeneuti.ca/voyeur

[39] http://www.zotero.org/

[40] Only the used/created tools did not describe before they will describe in this section.

MGIZA++, OBAPI, OGMIOS NLP Platform, OpinionFinder, PhiloMine, Protovis, SensePlaces2, Simile, Stanford NER, TermMine, TweetMotif, USAS, VisualText, Voyant tools, Wordfish, Wordscore, and Zotero.

	Tool name	Publications	Acronym project	Discipline	Task performed	Type
1	**ANNIE**	Berchialla et al.	--	Public Health	NER	Used
2	**ANNIE**	MacEachren et al.	--	Public Health	NER	Used
3	**CaboCha**	Suzuki	--	Political Science	Dependency analysis	Used
4	**Cafetière web platform**	Black et al.	--	Political Science	NER, sentiment analysis	Used
5	**ChaSen**	Suzuki	--	Political Science	Morphological analysis	Used
6	**CLAWS**	Gregory and Hardie	--	Literature	POS tagging	Used
7	**Genia**	Hamon and Grabar	--	Public Health	lemmatisation, POS tagging	Used
8	**Lingua-YaTeA**	Hamon and Grabar	--	Public Health	Term extraction	Used
9	**LIWC**	Tausczik and Pennebaker	--	Psychology	Morphological analysis, Word count, sentiment analysis, etc	Used
10	**LIWC**	Golder et al.	--	Psychology	Morphological analysis, Word count, sentiment analysis, etc	Used
11	**MALLET**	Yang and others	Mapping Historical Texts	History	Topic modeling	used
12	**MALLET**	Torget et al.	MT	History	Topic modeling	Used
13	**MGIZA++**	Bamman and Crane	--	Literature	Word alignment	Used
14	**OBAPI**	Cohen et al.	DMCI	History	Search	Created
15	**OGMIOS NLP Platform**	Hamon and Grabar	--	Public Health	Lemmatisation, POS tagging	Used
16	**OpinionFinder**	Bollen and others	--	Economics	Sentiment analysis	Used

17	**OpinionFinder**	O'Connor et al.	--	Political Science	Sentiment analysis	Used
18	**PhiloMine**	Argamon and others	--	Literature	Document classification	Used
19	**Protovis**	Torget et al.	MT	History	Visualization	Used
20	**SensePlaces2**	MacEachren et al.	--	Public Health	Load corpus, NER, georeferenced analysis, index	Created
21	**Simile**	Torget et al.	MT	History	Visualization	Used
22	**Stanford NER**	Yang and others	Mapping Historical Texts	History	NER	Used
23	**Stanford NER**	Torget et al.	MT	History	NER	Used
24	**Stanford NER**	Elson et al.	--	Literature	NER	Used
25	**Stanford NER**	Van Dalen-Oskam	--	Literature	NER	Used
26	**TermMine**	Black et al.	--	Political Science	Terminology extraction	Used
27	**TweetMotif**	Eisenstein and others	--	Sociolinguistics	NER Twitter	Used
28	**USAS**	Gregory and Hardie	--	Literature	Semantic analysis	Used
29	**VisualText**	Berchialla et al.	--	Public Health	Automatic Extraction Rules Generation	Used
30	**Voyant tools**	Cohen et al.	DMCI	History	Word count / Visualization	Used
31	**Wordfish**	Klüver	--	Political Science	Word count / Statistical analysis	Used
32	**Wordscore**	Klüver	--	Political Science	Word count / Statistical analysis	Used
33	**Zotero**	Cohen et al.	DMCI	History	Store and manage texts	Used

Table 7. Referenced tools in publications

ANNIE (A Nearly-New Information Extraction system) [41] is an information extraction system distributed with GATE, an open source solution for text processing developed by the University of Sheffield. ANNIE system has the following modules:

[41] http://gate.ac.uk/sale/tao/splitch6.html#chap:annie

tokenizer, gazetteer, sentence splitter, RegEx sentence splitter, part of speech tagger, semantic tagger, orthographic coreferencer (OrthoMatcher), and pronominal coreferencer.

CaboCha[42] is a dependency/syntactic parser based on machine learning for Japanese language.

Cafetière platform[43] is an application for text mining. Its first public version allows to upload plain text files to the server and to carry out: named entity recognition (on the basis of dictionary lookup and rule-based), term extraction (using embedded tool TermMine), and sentiment analysis. It was developed by the National Centre for Text Mining at University of Manchester.

ChaSen[44] is a morphological analyzer for Japanese language. It was developed at Computational Linguistics Laboratory of the Graduate School of Information Science at Nara Institute of Science and Technology (Japan).

CLAWS[45] is a part of speech (POS) tagger for English language. It was developed by UCREL (University Centre for Computer Corpus Research on Language) of Lancaster University.

Genia[46] is a tool that allows part of speech tagging, shallow parsing, and named entity recognition for biomedical texts. This solution was implemented in GENIA Project developed at Tsujii Laboratory of University of Tokyo.

Lingua-YaTeA[47] is an extension (in Perl) "*to extract terms from a corpus and to provide a syntactic analysis in a head-modifier format*". It was developed by researchers of Université Paris 13.

[42] https://code.google.com/p/cabocha/

[43] http://www.nactem.ac.uk/newsitem.php?item=159

[44] http://sourceforge.jp/projects/chasen-legacy/

[45] http://ucrel.lancs.ac.uk/claws/

[46] http://www.nactem.ac.uk/GENIA/tagger/

[47] https://metacpan.org/pod/Lingua::YaTeA

LIWC (Linguistic Inquiry and Word Count)[48] is a text analysis software that "*calculates the degree to which people use different categories of words across a wide array of texts -emails, speeches, poems, Twitter posts, etc*.-" It was developed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis.

MGIZA++[49] a word alignment application based on GIZA++. It was extended to multi-threading, resumes and incremental training.

Old Bailey API (OBAPI)[50] is a facility to work directly with texts from Old Bailey Proceedings. The OBAPI demonstrator allows to build a faceted query and export text to Voyant tools, or directly to work with the text throughout the API. This API was developed in the "*Datamining with Criminal Intent*" project.

Ogmios NLP Platform[51] is a platform composed by several modules to develop NLP tasks such as named entity tagging, word and sentence segmentation, POS tagging, lemmatization, term tagging, syntactic parsing, semantic tagging, and anaphora resolution. This platform was building during the ALVIS[52] project.

OpinionFinder[53] is a system for processing documents and automatically identifying subjective sentences and aspects as opinion agents, direct subjective expressions and speech events, and sentiment expressions. This tool was developed at the University of Pittsburgh, Cornell University, and the University of Utah.

PhiloMine[54] is an extension of PhiloLogic tool developed by the ARTFL Project and the Digital Library Development Center (DLDC) at the University of Chicago. PhiloMine develops several machine learning, text mining, and document clustering tasks.

SensePlaces2[55] is an application to geovisual analytics that analyzes place-time-attribute information from Twitter posts and supports crisis management.

---

[48] http://www.liwc.net/
[49] http://sourceforge.net/projects/mgizapp/

[50] http://www.oldbaileyonline.org/static/API.jsp

[51] http://search.cpan.org/~thhamon/Alvis-NLPPlatform-0.6/bin/ogmios-nlp-server

[52] http://cordis.europa.eu/ist/kct/alvis_synopsis.htm

[53] http://mpqa.cs.pitt.edu/opinionfinder/

[54] https://code.google.com/p/philomine/

[55] http://www.geovista.psu.edu/SensePlace2/

TermMine[56] is a tool for terminology extraction that integrates a domain-independent method for automatic term recognition (ATR) and acronym recognition (AcroMine). It was developed by the National Centre for Text Mining at University of Manchester.

TwitterMotif[57] is an application that summarizes Twitter information. Given a word or phrase, this application finds related tweets and groups them by statistically unlikely phrases that co-occur. Currently this application is down.

USAS (UCREL Semantic Analysis System)[58] is a framework for automatic semantic analysis of texts. It was developed by UCREL (University Centre for Computer Corpus Research on Language) of Lancaster University.

VisualText[59] is an integrated development environment for implementing information extraction and natural language processing systems, and text analyzers for several tasks (for example shallow extraction, intelligent web crawlers, categorization, text mining, etc.). It is developed by Text Analysis International.

Wordfish[60] is an application to extract political positions from texts documents. It works with word frequencies to place documents onto a single dimension. It was written in the R statistical languages and developed by researchers of Research Center SFB "Political Economy of Reforms" of the University of Mannheim and Department of Political Science of the University of Houston.

Wordscore[61] is a solution for extracting dimensional information from political texts using computer assisted content analysis. Currently there are two implementations: a command line version for Stata and a Java graphical version. Wordscore is part of an ongoing project by researchers of Trinity College Dublin, New York University, and University of Mannheim.

---

[56] http://www.nactem.ac.uk/software/termine/

[57] http://tweetmotif.com/about. Currently this application is down.

[58] http://ucrel.lancs.ac.uk/usas/

[59] http://www.textanalysis.com/

[60] http://www.wordfish.org/

[61] http://www.tcd.ie/Political_Science/wordscores/

4. Where to find information about tools and required resources

Others ways that the researchers have noticed about texts analysis tools were several online registries developed in the last years. Projects such as CLARIN, METANET4U, DARIAH-EU, Bamboo, PANACEA, TextGrid, or TAPoR 2.0 have been involved in the creation of tools, environments, and online registries with different but complementary purposes. In the following pages we will present a brief description about these initiatives.

Project	**Funding entity**	**Date**	**Research orientation**	**Field of Knowledge**	**Name registry**
CLARIN, Common Language Resources and Technologies http://www.clarin.eu/	European Union	2007-2011	Digital infraestructure	Humanities and Social Sciences	CLARIN Language Resource facet browser http://catalog.clarin.eu/vlo/
DARIA-EU, Digital Research Infrastructure for the Arts and Humanities http://www.dariah.eu/	European Union	2008-2011	Digital infraestructure	Humanities and Social Sciences	
Bamboo http://www.projectbamboo.org/	Andrew W. Mellon Foundation	2008-2011	Digital infraestructure	Digital Humanities	Bamboo DiRT http://dirt.projectbamboo.org/
PANACEA, Platform for the Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human	European Union	2010-2012	Tools and resources factory	Human Language Technologies	Panacea Registry http://registry.elda.org/

Language Technologies http://www.panacea-lr.eu/					
METANET4U, Enhancing the European Linguistic Infrastructure http://metanet4u.eu/	European Union	2011-2013		Human Language Technologies and Resources	Metashare Platform http://metashare.elda.org/
TextGrid, Virtual Research Environment for the Humanities http://www.textgrid.de/en/home/	German Federal Ministry of Education and Research	2006-2012 And 2012-2015	Digital infraestructure	Humanities	TextGrid Lab
TaPOR 2.0, Text Analysis Portal for Research http://www.tapor.ca/	Canadian Institute for Research Computing in the Arts, University of Alberta, Social Science and Humanities Research Council, Canada Foundation for Innovation		Digital infraestructure, Community	Arts, Humanities, Social Sciences	TaPOR 2.0, http://www.tapor.ca/

Table 8. Projects and catalogues related to texts analysis tools

The aim of **CLARIN** project was "*to facilitate access to collections of linguistic data (texts, multimedia recordings, dictionaries, etc.) and make possible the use in the net of analysis and exploitation tools of these data based on language technologies, specially*

for the research in Humanities and Social Sciences". At de beginning this project was funding by the European Union, but today CLARIN is one of the Research Infrastructures selected for the European Research Infrastructures Roadmap by the European Strategy Forum on Research Infrastructures (ESFRI) and consequently is funded by the participating countries. CLARIN ERIC maintains a Virtual Language Observatory to help the exploration of language resources and technologies around the world. Part of this observatory is the CLARIN Language Resource facet browser, an online tools and resources catalogue that currently displays 313,124 entries.

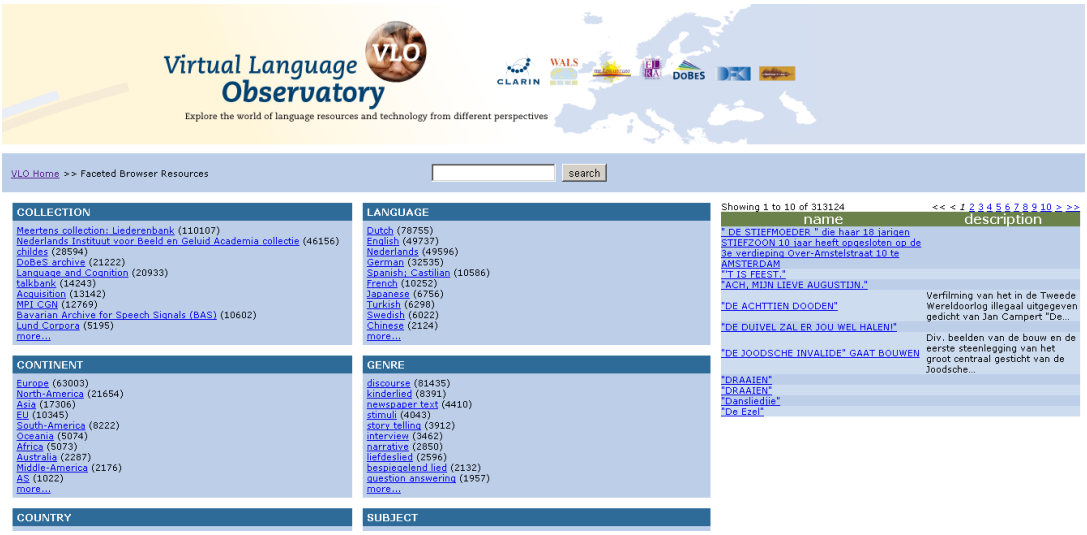


Figure 12. CLARIN Language Resource facet browser

The CLARIN Language Resource facet browser interface allows search by the collection, continent, country, organisation, data provider, tag, language, genre, subject, resource type, and national project.

Tool examples from this collection are:

- *Victor*,[62] "*a web page cleaning tool*";
- *LX-NER*,[63] "*a Named Entity Recognizer for Portuguese*"; or
- *Apertium Old Catalan morphological analyzer*[64].

62 http://catalog.clarin.eu/vlo/?wicket:bookmarkablePage=:eu.clarin.cmdi.vlo.pages.ShowResultPage&q=Victor&fq=resourceType:Application+/+Tool&docId=http://lrt.clarin.eu/node/3733

63 http://catalog.clarin.eu/vlo/?wicket:bookmarkablePage=:eu.clarin.cmdi.vlo.pages.ShowResultPage&q=LX-NER&fq=resourceType:Application+/+Tool&docId=http://lrt.clarin.eu/node/3659

64 http://catalog.clarin.eu/vlo/?wicket:bookmarkablePage=:eu.clarin.cmdi.vlo.pages.ShowResultPage&q=apertium&docId=http://lrt.clarin.eu/node/3597

METANET4U project had as key objective "*the implementation of a pan-European digital platform (META-SHARE) that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing in the different European languages*". Today, the Meta-Share platform hosts 1,684 language entries, 185 of them are tools and web services.

Figure 13. META-SHARE platform website

Resources and tools can be searched by language, resource type, media type, availability, license, restrictions of use, validated, linguality type, modality type, MIME type, conformance to standards/best practices, and domain.

The 185 tools included in this platform correspond to different subtypes. There are 104 tools, 40 web services, 9 NLP development environments, 5 platforms, and 23 corresponding to other classes (they gathered under "other" category).

Examples of some of Meta-Share tools and services are:

- *IULA tokenizer Web Service*[65] "*that splits a file in plain text format and UTF-8 encoded into units (tokens)*";
- *CollTerm* [66] , "*that collects collocation and term candidates from large representative corpus*"; or

[65] http://metashare.elda.org/repository/browse/iula-tokenizer-web-service/d32d941892c211e28763000c291ecfc8fb3aa57df0ce4a83b2362e2111c1ce36/

[66] http://metashare.elda.org/repository/browse/collocation-and-term-extractor/a89c02f4663d11e28a985ef2e4e6c59e76428bf02e394229a70428f25a839f75/

- *Web Content Extractor*[67], "*a tool for web pages content extraction for building web corpora*."

The goal of **DARIAH-EU** was "*to enhance and support digitally-enabled research and teaching across the humanities and arts*". The DARIAH-EU infrastructure will interconnect tools, information, people, and methodologies for researching, exploring, and supporting research across disciplines from humanities, social sciences, and arts. In DARIAH-EU, its Virtual Competency Centre (VCC) e-Infrastructure is working on a digital environment to share data and tools created by Digital Humanities community. In October 2012, DARIAH-EU presented an application at the European Research Infrastructure Consortium (ERIC).

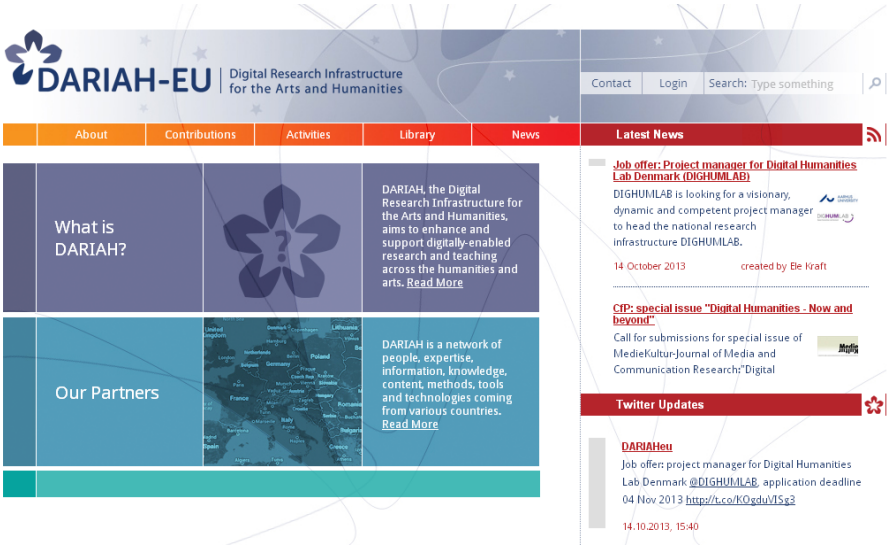


Figure 14. DARIAH-EU project website

The aim of **Project Bamboo** was to promote a digital infrastructure that supports digital humanities research through the development of shared technology services. The project resultant infrastructure is a registry of digital research tools for scholarly use, the Bamboo DiRT. The tools collected at Bamboo DiRT are categorizing in 29[68] classes. Also is possible marking the content with tags.

[67] http://metashare.elda.org/repository/browse/web-content-extractor/9e14ee4a663d11e28a985ef2e4e6c59e51a55e76bd4b47f39338db609624ff54/

[68] Analyze data, Analyze texts, Author an interactive work, Blog, Brainstorm/generate ideas, Build and share collections, Collect data, Communicate with colleagues, Conduct linguistic research, Convert/manipulate files, Create a mashup, Edit images, Find research materials, Make a dynamic map, Make a screencast, Manage bibliographic information, Manage tasks, Network with other researchers, Organize research materials, Publish and share information, Search visually, Share bookmarks, Stay current with research, Take notes/annotate resources, Transcribe handwritten or spoken texts, Use an iPad, Visualize data, Write a paper, and Write collaboratively.

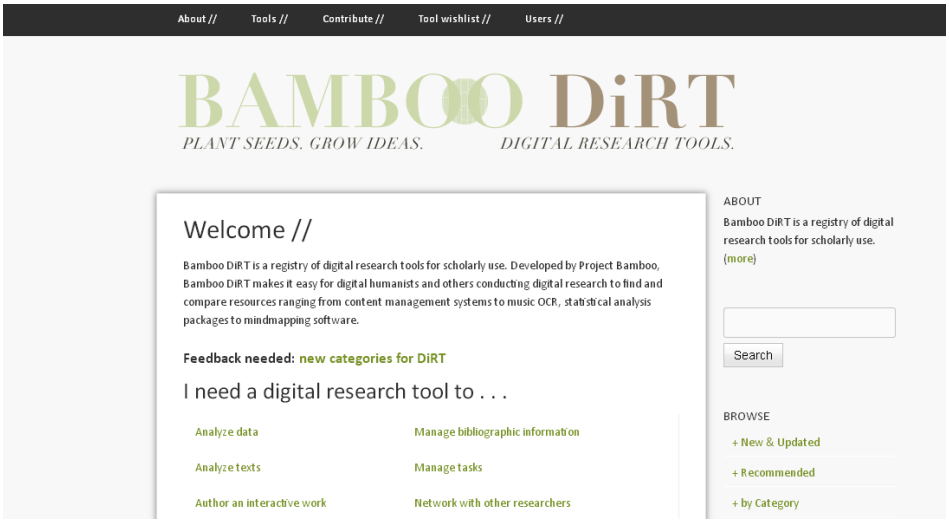


Figure 15. Bamboo DiRT website

These classes are presented in a broad formulation to encompass a broad spectrum of activities or tasks developed by different disciplines in the framework of the Digital Humanities. Some examples of these Bamboo DiRT tools are:

- *Textalyser*[69], "*an online text analysis tool that provides detailed statistics of your text, including features like the analysis of words groups, finding out keyword density, analysing the prominence of word or expressions"*;
- *MALLET*[70], "*a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text*"; or
- *Voyant tools*[71], "*web-based text analysis environment where users can apply a wide variety of tools to any text they import*".

PANACEA project's goal was to build a language resources factory for the automation of all the stages involved in the acquisition, production, updating, validation, and maintenance of linguistic technologies and resources. Across the duration of the project, created linguistic technologies and resources were disseminating on its website project and ad hoc registries to improve their later utilization. One of them, exclusively dedicated to tools and web services is the PANACEA registry with 162 registered entries.

[69] http://textalyser.net/

[70] http://mallet.cs.umass.edu/index.php

[71] http://voyant-tools.org/

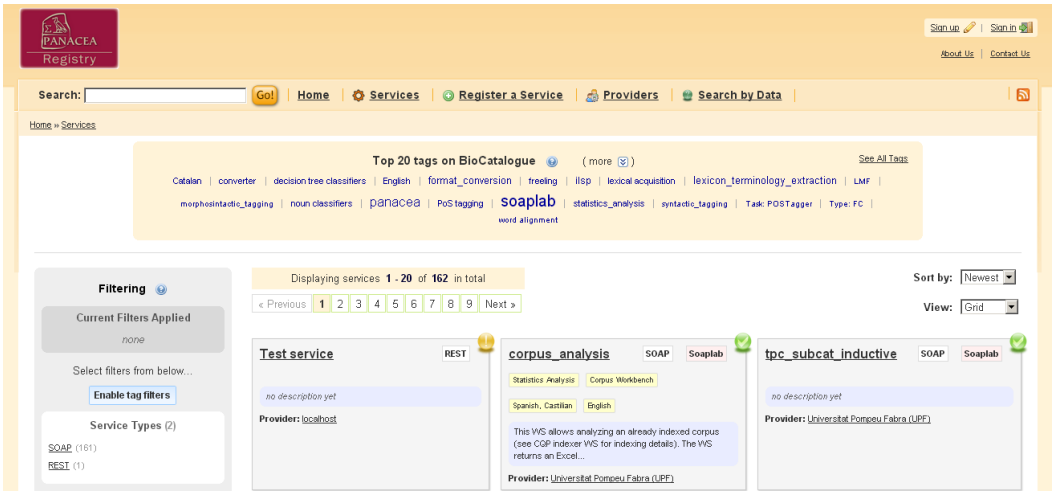


Figure 16. PANACEA registry website

The tools and services are categorized by development platform (Soap/Rest), language, subject categories, providers, and country. Also is possible marking the content with tags.

Examples of tools included in PANACEA Registry are:

- *Twitter NLP Web Service*[72], "it is a fast and robust Java-based tokenizer and part-of-speech tagger for Twitter. This web is based on the Twitter NLP tool developed by Noah's ARK group (Noah Smith's research group at the Language Technologies Institute, School of Computer Science, Carnegie Mellon University)";
- *Naive Bayes classifier Web Service*[73], *"that performs traditional Naive Bayes classification of instances given in a Weka file"*; or
- *Anonymizer Web Service*[74], *"that substitutes proper nouns with tags"*.

TextGrid project, since its start in 2006, has as goal to establish a virtual research environment for offering tools and services for the creation, analysis, editing, and publication of texts and images to Humanities research community. This project has two main components: the TextGrid Laboratory (TextGrid Lab) and the TextGrid Repository (TextGrid Rep). The TextGrid Lab includes tools and services that are designed to help the work of text-based disciplines like philology, linguistics,

[72] http://registry.elda.org/services/261

[73] http://registry.elda.org/services/229

[74] http://registry.elda.org/services/252

musicology, and art history. On the other hand, TextGrid Rep supports the storage and re-utilization of research data across their preservation. This project is funding for the period from June 2012 to May 2015 by the German Federal Ministry of Education and Research (BMBF).

The TextGrid Lab includes Eclipse-based interactive tools such as the XML editor, the Text-Image Link Editor (German "Text-Bild-Link-Editor", TBLE), the Text-Text Link Editor, and interfaces to dictionaries and reference material.

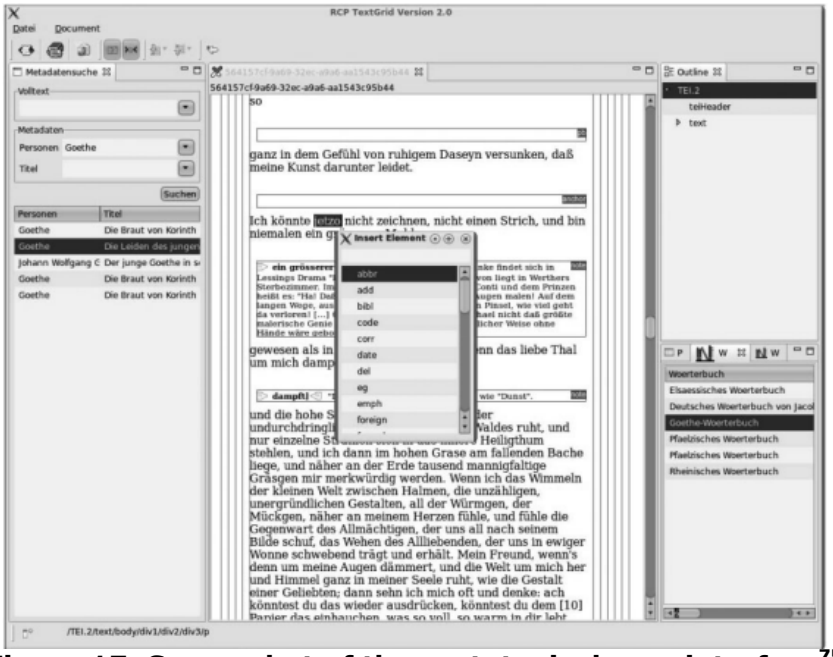


Figure 17. Screenshot of the prototypical user interface[75]

The XML editor allows switching between a technical view (with tags and attributes) and a more simple structural view. The Text-Image Link Editor aims to link text sequences with image sections with the purpose of creating files that contain text elements and topographic descriptions. The Text-Text Link Editor handles links between arbitrary fragments of XML document. On the other hand, the TextGrid Lab comprises Web Service-based tools that allow tasks such as collation, lemmatization, sorting, tokenization, etc.

[75] From left to right: Query interface, XML editor, interface to the Trier dictionary network (Wörterbuchnetz, http://www.woerterbuchnetz.de). (Zielinski et al. 2009)

The **TAPoR 2.0** project is a redesign of the previous version of TAPoR project (version 1.0)[76]. This second version of the Project has as goal a portal implementation to discover tools employed in text analysis and retrieval. The project's results are a registry for discovery digital research tools and a community oriented to Humanities scholars, students and others interested in computed assisted textual research.

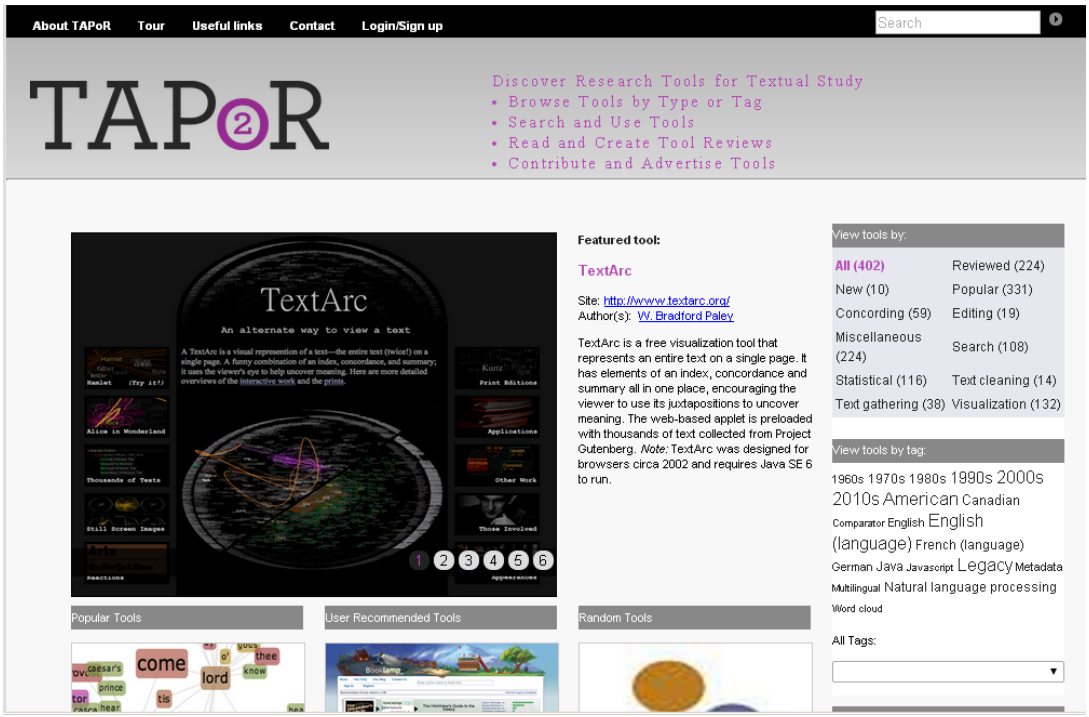


Figure 18. TAPoR 2.0 project and registry website[77]

The tools collected at the registry are categorizing in nine classes: Concording, Editing, Miscellaneous, New, Popular, Reviewed, Search, Statistical, Text cleaning, Text gathering, and Visualization. Also is possible marking the content with tags.

Examples of some of TAPoR 2.0 digital tools are:

- *Extract Text - XML (TAPoRware)[78], "This tool extracts text found within specific tags in an XML document. It is part of the TAPoRware toolset; an HTML version is also available*";

[76] TAPoR 1.0 is available at http://portal.tapor.ca/portal/portal. It had like goals supporting research into text representation, text analysis tool development, text analysis techniques and theory, and the access and usability of electronic text environments.

[77] http://www.tapor.ca/

[78] http://taporware.ualberta.ca/~taporware/xmlTools/xmlquery.shtml?

- *Collate: Interactive Collation of Large Textual Traditions*[79], it is *"a program designed for scholars concerned with the difficulties of medieval vernacular traditions. It aimed to help scholars with the preparation of critical editions, and could collate up to a hundred texts. Collate was also capable of handling marked up text. It was available for Macintosh only, and could be purchased directly from its creator";*
- *EURAC: Double Tree*[80]*, "It is a free, open source Java application providing a visualization component for supporting exploratory corpus analysis. It focuses particularly on analyzing concordances, and can also represent a KWIC for a single word by collapsing the contexts into a double sided tree. Each side of the tree can be expanded independently to browse the results. ce* that substitutes proper nouns with tags".

[79] http://www.hd.uib.no/humdata/2-91/robin.htm

[80] http://www.eurac.edu/en/research/institutes/multilingualism/Projects/LInfoVis/DoubleTree.html

5. Conclusions

1) One of the most frequent tasks is NER, it is language dependent but most of the European languages do have means to perform that task. Besides, there are several tools to handle old-languages, which is also an area of application

2) Topic modelling is a very popular tool nowadays, relatively easy to handle and gives workable information. Recently, we discovered some open tools which could be included in the demonstrator (https://code.google.com/p/topic-modeling-tool/).

3) Data Centers are requested to provide data in standards, along the ones used by tool providers.

4) It is important to provide tools for different languages to fit researcher's requirements.

References

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews." *Management Science* 57 (8): 1485–1509. doi:10.1287/mnsc.1110.1370.

Argamon, Shlomo, Russell Horton, Mark Olsen, and Sterling Stuart Stein. 2009. "Gender, Race, and Nationality in BlackDrama, 1850-2000: Mining Differences in Language Use in Authors and Their Characters." *Digital Humanities Quarterly* 3 (2).

Askitas, Nikos, and Klaus F Zimmermann. 2009. "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* 55 (2): 107–20.

Asur, Sitaram, and Bernardo A Huberman. 2010. "Predicting the Future with Social Media." In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 1:492–99. IEEE.

Bamman, David, and Gregory Crane. 2011. "Measuring Historical Word Sense Variation." In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 1–10. ACM.

Berchialla, Paola, Cecilia Scarinzi, Silvia Snidero, Yousif Rahim, and Dario Gregori. 2012. "Information Extraction Approaches to Unconventional Data Sources for 'Injury Surveillance System': The Case of Newspapers Clippings." *Journal of Medical Systems* 36 (2): 475–81. PH1. doi:10.1007/s10916-010-9492-1.

Bethard, Steven, and Dan Jurafsky. 2010. "Who Should I Cite: Learning Literature Search Models from Citation Behavior." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 609–18. ACM.

Black, William, Rob Procter, Steven Gray, and Sophia Ananiadou. 2012. "A Data and Analysis Resource for an Experiment in Text Mining a Collection of Micro-Blogs on a Political Topic." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet U\ugur Do\ugan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, 2083–88. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/1056_Paper.pdf.

Bohannon, John. 2011. "Google Books, Wikipedia, and the Future of Culturomics." *Science* 331 (6014): 135. doi:10.1126/science.331.6014.135.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2 (1): 1–8.

Bos, Wilfried, and Christian Tarnai. 1999. "Content Analysis in Empirical Social Research." *International Journal of Educational Research* 31 (8): 659 – 671. doi:http://dx.doi.org/10.1016/S0883-0355(99)00032-4.

Brier, Alan. 2011. "Computer Assisted Text Analysis in the Social Sciences." *Quality and Quantity* 45 (1): 103–28. /z-wcorg/.

Burdick, A., J. Drucker, P. Lunenfeld, T. Presner, and J. Schnapp. 2012. *Digital_Humanities*. Mit Press. http://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf.

Cohen, Dan, Frederick Gibbs, Tim Hitchcock, Geoffrey Rockwell, Jorg Sander, Robert Shoemaker, Stefan Sinclair, et al. 2011. "Data Mining with Criminal Intent. Final White Paper." http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf.

Culotta, Aron. 2010. "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages." In *Proceedings of the First Workshop on Social Media Analytics*, 115–22. ACM.

Eisenstein, Jacob, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. "A Latent Variable Model for Geographic Lexical Variation." In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–87. Association for Computational Linguistics.

Eisenstein, Jacob, Noah A Smith, and Eric P Xing. 2011. "Discovering Sociolinguistic Associations with Structured Sparsity." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 1365–74. Association for Computational Linguistics.

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138–47. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1858681.1858696.

Gerrish, Sean, and David M Blei. 2010. "A Language-Based Approach to Measuring Scholarly Impact." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 375–82.

Gilbert, Eric, and Karrie Karahalios. 2010. "Widespread Worry and the Stock Market." In , 59–65.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (7232): 1012–14.

Golder, Scott A., and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333 (6051): 1878–81. doi:10.1126/science.1202775.

Gregory, Ian N., and David Cooper. 2011. "Mapping the English Lake District: A Literary GIS." *Transactions of the Institute of British Geographers* 36 (1): 89–108. PHILIT_GIS2. doi:10.1111/j.1475-5661.2010.00405.x.

Gregory, Ian N., and Andrew Hardie. 2011. "Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems." *Literary and Linguistic Computing* 26 (3): 297–314. LIN5. doi:http://dx.doi.org/10.1093/llc/fqr022.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35. doi:10.1093/pan/mpp034.

Guerin-Pace, France. 1998. "Textual Statistics. An Exploratory Tool for the Social Sciences." *Population: An English Selection* 10 (1): pp. 73–95.

Hamon, Thierry, and Natalia Grabar. 2010. "Linguistic Approach for Identification of Medication Names and Related Information in Clinical Narratives." *Journal of the American Medical Informatics Association* 17 (5): 549–54. PH2. doi:10.1136/jamia.2010.004036.

Horton, Russell, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer. 2009. "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie." *Digital Humanities Quarterly* 3 (2).

Joshi, Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A Smith. 2010. "Movie Reviews and Revenues: An Experiment in Text Regression." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 293–96. Association for Computational Linguistics.

King, Gary. 2011. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331 (6018): 719–21. doi:10.1126/science.1197872.

Kirilenko, Andrei, Svetlana Stepchenkova, Rebecca Romsdahl, and Kristine Mattis. 2012. "Computer-Assisted Analysis of Public Discourse: A Case Study of the Precautionary Principle in the US and UK Press." *Quality and Quantity* 46 (2): 501–22. JMMC4. /z-wcorg/. doi:10.1007/s11135-010-9383-z.

Kirschenbaum, M. G. 2009. "The Remaking of Reading: Data Mining and the Digital Humanities." In *Nat. Sci. Found. Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation. Http://www. Cs. Umbc. Edu/\textbackslash hillol/NGDM07/abstracts/talks/MKirschenbaum. Pdf. Accessed*. Vol. 29.

Klüver, Heike. 2009. "Measuring Interest Group Influence Using Quantitative Text Analysis." *European Union Politics* 10 (4): 535–49. PS2. doi:10.1177/1465116509346782.

———. 2011. "The Contextual Nature of Lobbying: Explaining Lobbying Success in the European Union." *European Union Politics* 12 (4): 483–506. PS3. doi:10.1177/1465116511413163.

Kogan, Shimon, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. "Predicting Risk from Financial Reports with Regression." In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–80. Association for Computational Linguistics.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23. doi:10.1126/science.1167742.

Lebart, L., A. Salem, and L. Berry. 1998. *Exploring Textual Data*. Text, Speech and Language Technology. Springer.

Leetaru, Kalev H. 2012. "A Big Data Approach to the Humanities, Arts, and Social Sciences: Wikipedia's View of the World through Supercomputing." *Research Trends* 30 (september). http://www.researchtrends.com/issue-30-september-2012/a-big-data-approach-to-the-humanities-arts-and-social-sciences-summary/.

Liu, Alan. 2012. "The State of the Digital Humanities: A Report and a Critique." *Arts and Humanities in Higher Education* 11 (1-2): 8–41. doi:10.1177/1474022211427364.

MacEachren, A.M., A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. 2011. "SensePlace2: GeoTwitter Analytics Support for Situational Awareness." In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, 181–90. doi:10.1109/VAST.2011.6102456.

MacEachren, Alan M, Anthony C Robinson, Anuj Jaiswal, Scott Pezanowski, Alexander Savelyev, Justine Blanford, and Prasenjit Mitra. 2011. "Geo-Twitter Analytics: Applications in Crisis Management." In *Proceedings, 25th International Cartographic Conference, Paris, France*. http://nzdis.org/projects/projects/berlin/repository/revisions/61/entry/trunk/MastersDocs/Papers/Filer_Papers/MacEachren_ICC_2011.pdf.

Metaxas, Panagiotis Takis, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. "How (not) to Predict Elections." In *Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 165–71. IEEE.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. CULT1. doi:10.1126/science.1199644.

Morillo, Fernanda, Javier Aparicio, Borja González-Albo, and Luz Moreno. 2013. "Towards the Automation of Address Identification." *Scientometrics* 94 (1): 207–24. LIS1. doi:10.1007/s11192-012-0733-6.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. 2012. "Mine Your Own Business: Market-Structure Surveillance Through Text Mining." *Marketing Science* 31 (3): 521–43. doi:10.1287/mksc.1120.0713.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Washington, D.C.: Association for the Advancement of Artificial Intelligence. http://brenocon.com/oconnor_balasubramanyan_routledge_smith.icwsm2010.tweets_to_polls.pdf.

O'Connor, Brendan, David Bamman, and Noah A Smith. 2011. "Computational Text Analysis for Social Science: Model Assumptions and Complexity." *Public Health* 41 (42): 43.

Oelke, Daniela, Dimitrios Kokkinakis, and Mats Malm. 2012. "Advanced Visual Analytics Methods for Literature Analysis." In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 35–44. LaTeCH '12. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2390357.2390364.

Organisation for Economic Co-operation and Development. 2002. *Frascati Manual 2002 Proposed Standard Practice for Surveys on Research and Experimental Development*. Paris: OECD Publishing. http://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2002_9789264199040-en.

Parry, Marc. 2010. "The Humanities Go Google." *The Chronicle Higher Education*, May 28. http://chronicle.com/article/The-Humanities-Go-Google/65713/.

Paul, Michael J, and Mark Dredze. 2011. "You Are What You Tweet: Analyzing Twitter for Public Health." In *ICWSM*.

Pauwels, Teun. 2011. "Measuring Populism: A Quantitative Text Analysis of Party Literature in Belgium." *Journal of Elections, Public Opinion & Parties* 21 (1): 97–119. /z-wcorg/. doi:10.1080/17457289.2011.539483.

Popping, R. 2000. *Computer-Assisted Text Analysis*. New Technologies for Social Research Series. SAGE Publications.

Presner, Todd. 2009. "The Digital Humanities Manifesto 2.0." *Digital Humanities*. http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): pp. 209–228.

Ramage, Daniel, Christopher D Manning, and Susan Dumais. 2011. "Partially Labeled Topic Models for Interpretable Text Mining." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 457–65. ACM.

Roberts, Carl W., ed. 1997. *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcript*. Mahwah, NJ: Lawrence Erlbaum.

Stephens-Davidowitz, Seth. 2013. "The Cost of Racial Animus on a Black Presidential Candidate: Using Google Search Data to Find What Surveys Miss." *SSRN Journal 2012: 1* 55.

Strijbos, Jan-Willem, Rob L. Martens, Frans J. Prins, and Wim M. G. Jochems. 2006. "Content Analysis: What Are They Talking About?" *Computers & Education* 46 (1): 29 – 48. doi:http://dx.doi.org/10.1016/j.compedu.2005.04.002.

Suzuki, Takafumi. 2011. "Investigating Macroscopic Transitions in Japanese Foreign Policy Using Quantitative Text Analysis." *International Relations of the Asia-Pacific* 11 (3): 461–90. PS4. doi:10.1093/irap/lcr001.

Svensson, Patrik. 2010. "The Landscape of Digital Humanities." *Digital Humanities Quarterly* 4 (1). http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Tausczik, Yla R, and James W Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54.

Teichert, Thorsten, Patrick Mairif, Katja Schöntag, and Heye, Gerhard. 2011. "Co-Word Analysis for Assessing Consumer Associations: A Case Study in Market Research." *Text, Speech and Language Technology* 45: 115–24. doi:10.1007/978-94-007-1757-2_10.

Torget, Andrew J, Rada Mihalcea, Jon Christensen, and Geoff McGhee. 2011. "Mapping Texts: Combining Text-Mining and Geo-Visualization To Unlock The Research Potential of Historical Newspapers. A White Paper for the National Endowment for the Humanities." HIST4. http://mappingtexts.stanford.edu/whitepaper/MappingTexts_WhitePaper.pdf.

Van Dalen-Oskam, Karina. 2013. "Names in Novels: An Experiment in Computational Stylistics." *Literary and Linguistic Computing* 28 (2): 359–70. PHILIT4. doi:http://dx.doi.org/10.1093/llc/fqs007.

Wiedemann, Gregor. 2013. "Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14 (2). http://www.qualitative-research.net/index.php/fqs/article/view/1949.

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea. 2011. "Topic Modeling on Historical Newspapers." In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–104. LaTeCH '11. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2107636.2107649.

Yogatama, Dani, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. "Predicting a Scientific Community's Response to an Article." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 594–604. Association for Computational Linguistics.

Zielinski, Andrea, Wolfgang Pempe, Peter Gietz, Martin Haase, Stefan Funk, and Christian Simon. 2009. "TEI Documents in the Grid." *Literary and Linguistic Computing* 24 (3): 267–79. doi:10.1093/llc/fqp016.

Webgraphy

American Historical Association. 2012. "Digital History Abounds: A Roundup of Recent NEH Grant Projects." 2013. http://blog.historians.org/2013/07/digital-history-abounds-a-roundup-of-recent-neh-grant-projects/.

Columbia University, University of Brighton, University of Leiden, University of Toronto, University of York, and University of Washington. 2013. "ChartEx: Charter Excavator." Accessed September 23. http://www.chartex.org/.

Cordell, Ryan. 2013. "Infectious Texts: Mapping Viral Networks in Nineteenth-Century Newspapers." Accessed September 20. http://www.viraltexts.org/.

Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). 2012a. "Changing Climates." http://cass.lancs.ac.uk/?page_id=79.

———. 2012b. "Distressed Communities: Perception and Reality." http://cass.lancs.ac.uk/?page_id=88.

———. 2012c. "Hate Speech." http://cass.lancs.ac.uk/?page_id=83.

———. 2012d. "Religion, Citizenship and Integration." http://cass.lancs.ac.uk/?page_id=86.

———. 2012e. "Understanding Corporate Communications." http://cass.lancs.ac.uk/?page_id=90.

Indiana University Bloomington. 2010. "Cascades, Islands, or Streams? Time, Topic, and Scholarly Activities in Humanities and Social Science Research." 2013. http://did.ils.indiana.edu/.

Indiana University, University of Dundee, University of East London, and University of London. 2013. "Digging By Debating." Accessed September 23. http://diggingbydebating.org/.

Kings College London. Centre for e-Research. 2011. "The Digitisation of England's Place Names | Where Do You Think You Are?" http://englishplacenames.cerch.kcl.ac.uk/.

Lancaster University. 2012. "Spatial Humanities." http://www.lancaster.ac.uk/spatialhum/.

———. 2013. "Metaphor in End of Life Care (MELC) Project." http://ucrel.lancs.ac.uk/melc/.

National Endowment for the Humanities. 2013. "ODH Project Director Q&A: Ryan Cordell." Accessed September 20. http://www.neh.gov/divisions/odh/featured-project/odh-project-director-qa-ryan-cordell.

National Endowment for the Humanities, Andrew J. Torget, Rada Mihalcea, Jon Christensen, and Geoff McGhee. 2013. "MappingTexts_WhitePaper.pdf." Accessed September 20. http://mappingtexts.org/whitepaper/MappingTexts_WhitePaper.pdf.

National Endowment of the Humanities. 2013a. "NEH Grant Details: Digital Prosopography for Renaissance Musicians: Discovery of Social and Professional Networks." https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51636-13.

———. 2013b. "NEH Grant Details: Uncovering Reprinting Networks in Nineteenth-Century American Newspapers." https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51728-13.

Pompeii Bibliography and Mapping Project. 2013. "Pompeii Bibliography and Mapping Project." http://digitalhumanities.umass.edu/pbmp/?page_id=13.

Romppel, Matthias. 2013. "Software for Content Analysis and Text Analysis: Qualitative Analysis." http://www.content-analysis.de/software/qualitative-analysis.

Roy Rosenzweig Center for History and New Media. 2013. "Criminal Intent." Accessed September 23. http://criminalintent.org/.

Sinclair, Stéfan, and Geoffrey Rockwell. 2014. "Voyant Tools: Reveal Your Texts." http://voyant-tools.org/.

Social Science Consulting. 2011. "Text Analysis." 2013. http://www.textanalysis.info/.

Stanford University.Humanities Center. 2013. "Digging Into the Enlightenment." Accessed September 23. http://enlightenment.humanitiesnetwork.org/.

TAPoR Team. University of Alberta. 2013. "TAPoR 2.0: Text Analysis Portal for Research." http://tapor.ca/.

TextGrid. 2012. "TextGrid: Digital Editing – Research – Archiving." http://www.textgrid.de/en/.
Tufts University. 2013. "Dynamic Variorum Editions." Accessed September 23. http://sites.tufts.edu/dynamicvariorum/.
University of Hertfordshire, University of London, and University of Sheffield. 2013. "Connected Histories." http://www.connectedhistories.org/.
University of Leeds. Centre for the History and Philosophy of Science. 2013. "SciPer Project." http://www.leeds.ac.uk/arts/homepage/426/sciper_project/.
University of Leeds. School of Philosophy. 2013. "Science in the 19th Century Periodical." Accessed September 23. http://www.sciper.org/.
University of Manchester. National Centre for Text Mining (NaCTeM). 2013. "DID - ISHER Project." http://www.nactem.ac.uk/DID-ISHER/.
University of North Texas, and Stanford University. 2011. "Mapping Texts." http://mappingtexts.org/.
University of Sheffield. Humanities Research Institute (HRI),. 2009. "Scrutiny." http://www.hrionline.ac.uk/scrutiny/.

Annex. List of project descriptions

ENVIRONMENT

Project name	**Changing Climates**
Project acronym	**CHC**
Web site	http://cass.lancs.ac.uk/?page_id=79
Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)
Dates	2012-
Description[81]	A project about a corpus-based research of discourses around global warming, energy, and mobilities in Brazil and Britain. It will analyze talk, text, and social activities in two different economies with opposite positions about climate change: widely accepted (Brazil) and with public sphere skepticism (Britain).
Research orientation	Discourse analysis

HISTORY

Project name	**Digital Prosopography for Renaissance Musicians: Discovery of Social and Professional Networks**
Project acronym	**DPRM**
Web site	

[81] Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). Changing Climates [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://cass.lancs.ac.uk/?page_id=79

	https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51636-13
Funding entity	The National Endowment of the Humanities Office of Digital Humanities (NEH-ODH)
Partners	Johns Hopkins University
Dates	2013-2014
Description[82]	This project is exploring the applicability of FOAF (Friend of a Friend) for describing relationships between Renaissance musicians in a new biographical tool. FOAF is a semantic web resource, a machine-readable ontology describing persons, their activities and their relations. The project aims "*to know how to extract biographical and relational data automatically from textual corpus using natural language processing technologies, creating a model applicable to other fields of knowledge and time periods*".
Tools used/created	Without specification
Research orientation	Information extraction Named Entity Recognition (NER) Linked Open Data

Project name	**Uncovering Reprinting Networks in 19th Century American Newspapers [aka Infectious Texts]**

[82] National Endowment of the Humanities. NEH Grant details: Digital Prosopography for Renaissance Musicians: Discovery of Social and Professional Networks [Internet]. 2013 [cited 2013 Sep 20]. Available from: https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51636-13

Project acronym	**Viraltext**
Web site	https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51728-13 http://www.viraltexts.org/
Funding entity	National Endowment for the Humanities (NEH)
Partners	Northeastern University
Dates	2013-2014
Description[83]	The project aims to improve the search on a Nineteenth-Century American Newspapers corpus using and developing data mining tools. It is working spaces-efficient n-gram indexing to identify candidate newspapers and then exploits local models of alignment to identify reprinted fragments unknown a priori.
Tools used/created	Without specification
Research orientation	Information extraction Tool development

Project name	**Integrated Social History Environment for Research – Digging into Social Unrest**
Project acronym	**ISHER**

[83] Cordell R. Infectious Texts: Mapping Viral Networks in Nineteenth-Century Newspapers[Internet]. [cited 2013 Sep 20]. Available from: http://www.viraltexts.org/

Web site	http://www.nactem.ac.uk/DID-ISHER/
Funding entity	Digging Into Data Challenge
Partners	The University of Manchester, Radboud University Nijmegen, University of Illinois at Urbana-Champaign
Dates	2012-2013
Description[84]	This project is researching the development of text mining tools for automated analysis of news archives. These tools will provide "*an integrated framework to detect, link, and visualize events, trends, people, organizations, and other entities of interest to social historians*".
Tools used/created	Without specification
Research orientation	Information extraction - Tool development Named Entity Recognition (NER) Visualization

[84] University of Manchester. National Centre for Text Mining (NaCTeM). DID - ISHER Project [Internet]. 2013 [cited 2013 Sep 23]. Available from: http://www.nactem.ac.uk/DID-ISHER/

Project name	**ChartEx**
Project acronym	**ChartEx**
Web site	http://www.chartex.org/
Funding entity	Digging Into Data Challenge [supported by NSF, NEH, JISC and SSHRC/CRSHC]
Partners	University of Washington, IMLS, Leiden University, NWO, University of York, AHRC/ESRC/JISC, University of Toronto, SSHRC, University of Brighton, Columbia University
Dates	2011-?
Description[85]	The aim of this project "*is developing new ways of exploring the full text content of digital historical records*". The project is working with medieval charters from the 12th to the 16th centuries. The new ChartEx tools will apply "*Natural Language Processing and Data Mining automatically to extract information about places, people and events in their lives from the charters and find new relationships among them*".
Tools used/created	Brat (text annotation, NER)
Research orientation	Information extraction - Tool development Named Entity Recognition (NER)

[85] Columbia University, University of Brighton, University of Leiden, University of Toronto, University of York, University of Washington. ChartEx: Charter Excavator [Internet]. [cited 2013 Sep 23]. Available from: http://www.chartex.org/

Project name	**Connecting Historical Authorities with Linked data, Indices Contexts and Entities**
Project acronym	**CHALICE**
Web site	http://chalice.blogs.edina.ac.uk/
Funding entity	Joint Information Systems Committee (JISC)
Partners	Kings College London. Centre for e-Research, Queen's University Belfast. Centre for Data Digitisation and Analysis, University of Edinburgh. School of Informatics. Language Technology Group
Dates	2010-2011
Description[86]	The project created a Linked Data gazetteer for United Kingdom historic place names and linked it to other sources of place names reference information (e.g. geonames.org). This project worked Named Entity Recognition techniques to extract data from select digitized volumes of the English Place Name Survey.
Tools used/created	Without specification
Research orientation	Information extraction – Linked data Named Entity Recognition (NER)

[86] Kings College London. Centre for e-Research. The Digitisation of England's Place Names | Where do you think you are? [Internet]. 2011 [cited 2013 Sep 23]. Available from: http://englishplacenames.cerch.kcl.ac.uk/

Project name	**Data Mining with Criminal Intent**
Project acronym	**DMCI**
Web site	http://criminalintent.org/
Funding entity	National Endowment for the Humanities (NEH)
Partners	George Mason University, NEH; University of Hertfordshire, JISC; University of Alberta, SSHRC; The National Archives (United Kingdom), McMaster University, the Open University, Amherst College, University of Sheffield, Trent University, and the University of Western Ontario
Dates	2010-2011
Description[87]	"*This project explored the tools and infrastructures that will make possible the "ordinary working historian", currently with little use of digital techniques, with the support of text mining and visualization*". The DMCI project created a digital research environment that allowed the user to query through the Proceedings of the Old Bailey Online, save and manage results sets in a Zotero account, and send the selected texts and results for analysis and visualization (to tools like Voyant).
Tools used/created	Zotero / Voyant tools
Research orientation	Information extraction Visualization

[87] Roy Rosenzweig Center for History and New Media. Criminal Intent [Internet]. [cited 2013 Sep 23]. Available from: http://criminalintent.org/

Project name	**Digging Into the Enlightenment: Mapping the Republic of Letters**
Project acronym	**DIE/MRL**
Web site	http://enlightenment.humanitiesnetwork.org/
Funding entity	National Endowment for the Humanities (NEH), Joint Information Systems Committee (JISC), National Science Foundation (NSF)
Partners	Stanford University, University of Oklahoma, Oxford University
Dates	2010-2011
Description[88]	This collaborative project analyzed *"the degree to which the effects of the Enlightenment can be observed in the letters of people of various occupations on a corpus of 53,000 18th-century letters named Electronic Enlightenment (EE). The hypothesis of the project proposed a new perspective about the practice of interpretative research in the humanities. This view aimed to integrate innovative visualization and annotation techniques into interactive tools for exploring and analyzing information about people, places, times, and relationships into the "Republic of Letters"."*
Tools used/created	Google Maps
Research orientation	Information extraction – Tool development Named Entity Recognition (NER)

[88] Stanford University.Humanities Center. Digging Into the Enlightenment: Mapping the Reppublic of Letters [Internet]. [cited 2013 Sep 23]. Available from: http://enlightenment.humanitiesnetwork.org/

	Visualization

Project name	**Dynamic Variorum Editions**
Project acronym	**DVE**
Web site	http://sites.tufts.edu/dynamicvariorum/
Funding entity	Digging Into Data Challenge [supported by NSF, NEH, JISC and SSHRC/CRSHC]
Partners	Tufts University, Imperial College London and Mount Allison University
Dates	2010-2011
Description[89]	The project's goal was to identify and track topics about the Greco-Roman world as they appear in multilingual public document collections (Internet Archive, JSTOR, HathiTrust, etc.). This project aimed to create an environment to generate "dynamic variorum" editions of texts based on a services infrastructure. The project worked mining primary and secondary digital sources to locate where people and places from the Greco-Roman world are discussed, which Greek and Latin works are cited, and what kinds of things are said about the people, places, and texts of the Greco-Roman world over time.
Tools used/created	Without specification
Research orientation	Information extraction – Creation of digital research environment

[89] Tufts University. Dynamic Variorum Editions [Internet]. [cited 2013 Sep 23]. Available from: http://sites.tufts.edu/dynamicvariorum/

	Named Entity Recognition (NER)

Project name	**Mapping Texts**
Project acronym	**MT**
Web site	http://mappingtexts.org/
Funding entity	
Partners	University of North Texas, Stanford University
Dates	2010-
Description[90]	The project aimed to develop "*new ways of discovering and analyzing language patterns embedded in historical newspaper databases. Its approach was to combine text mining and geospatial visualization methods to explore massive collections of electronic texts*."
Tools used/created	MALLET (topic modeling), Stanford NER (named entity recognition), Google Maps, Google Finance time series, Simile widgets, and Protovis (visualization)
Research orientation	Information extraction - NER Visualization

[90] University of North Texas, Stanford University. Mapping Texts [Internet]. 2011 [cited 2013 Sep 20]. Available from: http://mappingtexts.org/

Project name	**Connected Histories: Sources for Building British History, 1500-1900**
Project acronym	**CH**
Web site	http://www.connectedhistories.org/
Funding entity	Joint Information Systems Committee (JISC)
Partners	University of Hertfordshire, University of London. Institute of Historical Research, University of Sheffield. Humanities Research Institute (HRI)
Dates	2009-2011
Description[91]	The project supplied *"an integrated search environment for querying distributed electronic content about early modern and 19th century British history"*. The project provided a search engine that did not search these resources directly, but it searched indexes that it has created from the full content.
Tools used/created	Without specification
Research orientation	Creation of digital search environment Federated search

Project name	**Scrutiny: A Firefox Extension for Entity Recognition within Research Data**

[91] University of Hertfordshire, University of London, University of Sheffield. Connected Histories [Internet]. 2013 [cited 2013 Sep 23]. Available from: http://www.connectedhistories.org/

Project acronym	**Scrutiny**
Web site	http://www.hrionline.ac.uk/scrutiny/
Funding entity	Joint Information Systems Committee (JISC)
Partners	University of Hertfordshire, University of Sheffield. Humanities Research Institute (HRI), and PlayGen Limited
Dates	2009-2010
Description[92]	The project developed a Firefox extension, named Scrutiny. This extension scans "*web pages selected by individual users and highlight entities that it thinks will interest them*. *The users are be able to train Scrutiny to identify relevant entities to their field of study both by using pre-defined, subject specific entity recognition files*, *and by refining Scrutiny's understanding of their personal interests through an iterative process"* (accepting/discarding Scrutiny's suggestions).
Tools used/created	Scrutiny (NER)
Research orientation	Information extraction - Tool development Named Entity Recognition (NER)

[92] University of Sheffield. Humanities Research Institute (HRI),. Scrutiny [Internet]. 2009 [cited 2013 Sep 23]. Available from: http://www.hrionline.ac.uk/scrutiny/

Project name	**Science in the Nineteenth-Century Periodical**
Project acronym	**SciPer**
Web site	http://www.sciper.org
Funding entity	Art and Humanities Research Council (AHRC), The Leverhulme Trust, and Modern Humanities Research Association (MHRA)
Partners	University of Leeds. Centre for the History and Philosophy of Science, University of Sheffield. Centre for Nineteenth-Century Studies
Dates	1999-2007
Description[93]	The project's goal was to identify and analyze the representation of science, technology and medicine, as well as the inter-penetration of science and literature, in the general periodical press between 1800 and 1900 in Britain. This project addressed the reception of scientific ideas in the general press, but also examined the creation of non-specialist forms of scientific discourse within a periodical format, and the ways in which they interact with the miscellany of other types of articles found in nineteenth-century periodicals.
Tools used/created	Without specification
Research orientation	Discourse analysis

[93] University of Leeds. Centre for the History and Philosophy of Science. SciPer project [Internet]. 2013 [cited 2013 Sep 23]. Available from: http://www.leeds.ac.uk/arts/homepage/426/sciper_project/

Project name	**The Pompeii Bibliography and Mapping Resource**
Project acronym	**PBMP**
Web site	http://digitalhumanities.umass.edu/pbmp/?page_id=13
Funding entity	National Endowment for the Humanities (NEH), University of Massachusetts Amherst, UMass Digital Humanities Initiative
Partners	University of Massachusetts Amherst
Dates	
Description[94]	The project is working to create an integrated framework that binds two resources. The first resource is a database of citation and full-text repository about the ancient city of Pompeii. The second one is a Geographical Information System (GIS) map of that city. The integrated online interface planned will allow the user to explore the bibliographic database and repository via the GIS map, or to find places in a search of the database or repository and display them in the GIS map.
Tools used/created	Digital environment
Research orientation	Information extraction – Creation of digital research environment Geo-referenced analysis

PUBLIC HEALTH

[94] Pompeii Bibliography and Mapping Project. Pompeii Bibliography and Mapping Project [Internet]. 2013 [cited 2013 Sep 20]. Available from: http://digitalhumanities.umass.edu/pbmp/?page_id=13

Project name	**Metaphor in End-of-Life Care [CASS Affiliated Projects]**
Project acronym	**MELC**
Web site	http://ucrel.lancs.ac.uk/melc/
Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)
Partners	• European Association for Palliative Care • International Association for Research and Applying Metaphor • International Observatory on End of Life Care • Lancaster University. Faculty of Health & Medicine PhD in Palliative Care • Lancaster University. Management School MA in Hospice Leadership
Dates	2007-
Description[95]	A corpus-based study of the metaphors used to talk about end-of-life care by patients nearing the end of life, unpaid family caregivers and health professionals. The project is studying interviews and online forum data in order to investigate how metaphors may help or hinder successful communication among these different groups. The project's goal is using the findings to improve the quality of communication at the end of life.
Tools used/created	Without specification
Research orientation	Discourse analysis

SOCIOLOGY

[95] Lancaster University. Metaphor in end of life care (MELC) project [Internet]. 2013 [cited 2013 Sep 23]. Available from: http://ucrel.lancs.ac.uk/melc/

Project name	**Hate Speech**
Project acronym	**HS**
Web site	http://cass.lancs.ac.uk/?page_id=83
Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)
Dates	2012-
Description[96]	A project that will take a more measured approach than that evident in the press from a linguistic perspective by looking at the use of hate speech. The project's aim will explore to what extent it can evidently determine linguistic triggers for prosecution, and hence a linguistic warrant for action.
Tools used/created	Without specification
Research orientation	Discourse analysis Sentiment analysis

[96] Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). Hate Speech [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://cass.lancs.ac.uk/?page_id=83

Project name	**Religion, Citizenship and Integration**
Project acronym	**RCI**
Web site	http://cass.lancs.ac.uk/?page_id=86
Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)
Dates	2012-
Description[97]	A survey of 100 interviews with immigrants, building on a Home Office funded project entitled 'What Works' which looks at the role of religion in the lives of well-integrated immigrants. The project's aim will be improve understanding integration if does work or not.
Tools used/created	Without specification
Research orientation	Discourse analysis

Project name	**Distressed Communities: Perception and Reality**
Project acronym	**DCPR**

[97] Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). Religion, Citizenship and Integration [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://cass.lancs.ac.uk/?page_id=86

Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)
Dates	2012-
Description[98]	This project will apply thematic geo-referencing to a broad-coverage press corpus, mapping popularly-perceived associations between social issues (poverty, deprivation, immigration, etc.) and different communities, regions or localities to investigate how far attitudes match the reality found in statistical datasets. This survey is covering United Kingdom press corpus and statistical data.
Tools used/created	Without specification
Research orientation	Discourse analysis Geo-referenced analysis

COMMUNICATION

Project name	**Understanding Corporate Communications**
Project acronym	**UCC**
Web site	http://cass.lancs.ac.uk/?page_id=90
Funding entity	Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS)

[98] Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). Distressed Communities: Perception and Reality [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://cass.lancs.ac.uk/?page_id=88

Dates	2012-
Description[99]	A survey about a comprehensive analysis of the form, content and impact of communications between large, publicly traded corporations and their key stakeholder groups. This project is studying the following three key aspects of corporate governance: a) compliance with governance requirements and recommendations; b) executive remuneration; and c) senior management turnover.
Tools used/created	Without specification
Research orientation	Discourse analysis

MULTIDISCIPLINARY

Project name	**Cascades, islands, or streams? Time, topic, and scholarly activities in humanities and social science research**
Project acronym	**CIS**
Web site	http://did.ils.indiana.edu/
Funding entity	Digging Into Data Challenge [supported by NSF, NEH, JISC and SSHRC/CRSHC]
Partners	University of Wolverhampton, Indiana University Bloomington, NSF Université de Montréal, SSHRC/AHRC, ESRC

[99] Economic and Social Research Council. Centre for Corpus Approaches to Social Science (ESRC. CASS). Understanding Corporate Communications [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://cass.lancs.ac.uk/?page_id=90

Dates	2012-?
Description[100]	The project's aim is *"to create tools for analysis of topic lifecycles across heterogeneous corpora"*. This project will study the development of topics in history of science, social network analysis, cognitive science, and digital humanities. It will combine data from formal sources (dissertations, conference proceedings, journal articles, and grant proposals), and informal communication channels (listservs, blogs, and twitter) with the motivation to provide a more holistic view on scientific communication.
Tools used/created	Without specification
Research orientation	Information extraction Tool development

Project name	**Digging by Debating: Linking massive datasets to specific arguments**
Project acronym	**DbyD**
Web site	http://www.jisc.ac.uk/whatwedo/programmes/digitisation/diggingintodata/digbydebate.aspx
Funding entity	Digging into Data Challenge [through the National Endowment for the Humanities (NEH), Joint Information Systems Committee (JISC), Economic and Social Research Council (ESRC), and Arts and Humanities Research Board (AHRB)]
Partners	Indiana University, University of Dundee, University of East London, and University of London

[100] Indiana University Bloomington. Cascades, Islands, or Streams? Time, Topic, and Scholarly Activities in Humanities and Social Science Research [Internet]. 2010 [cited 2013 Sep 23]. Available from: http://did.ils.indiana.edu/

Dates	2012-
Description[101]	The project aims *"to uncover and represent the argumentative structure of digitized documents. Users will be able to see the semantic landscape of books and articles, to zoom into specific topic areas, and to use cutting-edge interpretive techniques to perform linguistic analyses of the raw text. Arguments and debates expressed in these texts can be connected to and can serve to anchor online discussions that form a part of the Argument Web, an emerging environment".* The project is working on four level of analysis: Macro (Sci/Phil Maps) to Micro (detailed arguments). It is proposing visualizing contact points between philosophy and the sciences (they starting with philosophy and psychology, topic modeling to identify rich content in a chosen topic, identify and map key arguments by novel analysis framework for propositions and arguments, and sentence modeling to get back to HathiTrust materials.
Tools used/created	SCI2 (topic analysis, NER)
Research orientation	Information extraction – Topic - NER Platform development (InterDebates)

Project name	**Spatial Humanities: texts, geographic information systems and places**
Project acronym	**SH**
Web site	http://www.lancs.ac.uk/spatialhum/
Funding entity	European Research Council (ERC)

[101] Indiana University, University of Dundee, University of East London, University of London. Digging By Debating [Internet]. [cited 2013 Sep 23]. Available from: http://diggingbydebating.org/

Partners	Lancaster University
Dates	2012-2016
Description[102]	This project is aiming to create a gradual change in the way that place; space and geography are considerate in the study of humanities. It is developing and applying methodologies to allow unstructured texts - including books, newspapers and official reports - to be analyzed in a manner that stresses space, place and mapping. The project is working currently to: a) Develop appropriate techniques for the analysis of textual sources within a GIS environment; b) Analyze qualitative sources in the spatial humanities: the English Lake District, c) Bridge the quantitative and qualitative divide: Health and society in nineteenth and twentieth century England & Wales; and d) Develop the skills-base in the spatial humanities.
Tools used/created	Google Maps
Research orientation	Information extraction Geo-referenced analysis

[102] Lancaster University. Spatial Humanities [Internet]. 2012 [cited 2013 Sep 23]. Available from: http://www.lancaster.ac.uk/spatialhum/

Annex II: User evaluation of the demonstrator

To ascertain whether the chaining would be a useful service, we interviewed a couple of researchers and other professionals. During these interviews the demonstrator was used as a reference to investigate the need for language tools for the interviewee's work and to evaluate the functionality of the demonstrator itself. In the interviews we asked about the kind of work or projects the interviewee is engaged in, the issues for which he or she is using, or could use language tools. We inquired after the need for specific tools and demonstrated our workflow chaining.

We interviewed 5 persons. The background of the interviewees was divers; researchers as well as support staff, within the context of the social sciences as well as history. Three interviewees were positive about our demonstrator. One interviewee did not feel the need for such a service, as she doesn't expect more than already is possible with ATLAS-ti, a commercial tool she uses for her research.

The kind of tools the interviewees would like to see in such a chaining were Named Entity Recognition (NER), Named Entity Disambiguation (NED), Topic modelling and the harmonisation and conversion of dates. Two interviewees expect that NER could be used for anonymisation or pseudonimisation. For this latter use it is very important the tool is very precise. One interviewee mentioned the possibility to share the evaluation results of a specific tool; this would help to estimate the exactness of a specific tool for a similar dataset.

A Web application is seen as very useful. Interviewees would prefer this to a tool that has to be downloaded and used offline. One interviewee made the point that a user web interface is not useful with large datasets, in some cases an API would be more convenient.

Another remark that was made is that the results of the workflow should not be open available, as it is at the moment, but secured.

In the following section you find the summaries of the interviews (one of them in Spanish, but we produced a Summary in English).

Interviews

Interview 1

Background information interviewee
Education: Phd informatics (social science informatics)
Career: assistant professor
Gender: Female
Age: between 30 and 40

Projects
Interviewee was engaged in the PoliMedia project and is now involved in the Talk of Europe Project (ToE).

The PoliMedia project links the minutes of the debates in the Dutch Parliament (Dutch Hansard) to the databases of historical newspapers and ANP radio bulletins to allow cross-media analysis of coverage in a uniform search interface.
For each fragment from a single speaker in a debate, relevant information has been extracted (the name of the speaker, the date, important terms from its content and important terms from the description of the complete debate). This information was then combined to create a query with which the archives of newspapers, radio bulletins and television programmes are being searched. Media items that corresponded to this query were retrieved and a link was created between the speech and the media item, creating a Semantic Web of Dutch Hansard and media coverage.

ToE is a similar project for Europe. It makes the data of the plenary sessions of the European Union available as linked open data, enriched with biographical and political information on the speakers. The main source for the semantic dataset is the collection with the full transcripts of the plenary meetings. The EU publishes these in all current official languages of the EU on its website. Represented online are the date and title of each agenda item, the name of the speakers and the transcript of every speech, including all translations. To become useful for scholars in the Humanities and the Social Sciences, the data created by Talk of Europe requires tools for exploration and analysis. ToE organises creative camps to bring together developers and academic researchers, with the goal of making inventive use of the dataset, exploiting web and natural language processing techniques to add new knowledge and functionality to the dataset. The goal is to develop proof-of-concept tools that can be applied in scholarly research in the political sciences and humanities.

Problems they want to solve with language tools
Recognition of names
Harmonisation of dates
Topic modelling

How do they solve them currently?

The texts of the plenary sessions are available in HTML, scraped of the Parliament website. Named Entity Recognition (NER) is applied, however the quality of Named Entity Disambiguation (NED), (linking to the controlled vocabulary) was too low for proper usage.

Very important is topic modelling. Ideally you would like to compare the content of the speeches with speeches with a similar content from another period, for example a comparison of speeches about the environment in the fifties with today speeches. Therefore you need topics, which characterize the content of the speech. For this purpose, they have used an unsupervised Latent Dirichlet Allocation (LDA) with MALLET. The topics were linked to news archives by using them as search term.

Drawbacks of their current work procedures
MALLET works fine for producing links, however the search terms still remain words and are not genuine topics by which texts from different periods can be compared. The topics were discarded after linking because they had little more significance. The result of this kind of topic modelling is ambiguous, because the produced topics are basically just words and not genuine topics.

In Polimedia the source texts were already in XML and the entities were already annotated, so NER/NED were not necessary. In texts that were OCRed there are many conversion errors, in many times even misspellings of names of politicians. In those cases links to external information isn't possible.

How would these be improved?
Useful would be a possibility to share training data and annotated texts. Interviewee mentioned a research project about conflict analysis (who is in conflict with whom) by using entity co-referencing tools. At the moment the existing tools they use are not always clear about the quality of their results. Ideally you would like to get information about precision and recall.

With NER, NED, topic modelling, sentiment analysis, it is difficult to determine how correct the results are with a particular dataset. If it would be possible to share the evaluation results of a specific tool, this would help to estimate the exactness of a specific tool for a similar dataset. However, this isn't always possible to republish the texts, for example with newspapers. You can publish the links, but the whole text is not available due to copyrights and licences.

Would a chaining of tools be of use for their project? (Reference to our demonstrator)
Interviewee sees the idea of the chaining of tools as very worthwhile. A Web application would be very useful. Interviewee would prefer it to a tool that she has to download and use offline. However a user web interface is not useful with large dataset, in some cases an Application Programming Interface (API) would be more convenient. The same applies for the output; it should be possible to pipe these into other tools (for Polimedia and ToE in XML or RDF format).

Which other tools should be chained?
Another useful tool would be a universal date conversion tool, which converts all kinds of dates to real machine actionable dates.

Interview 2

Background information interviewee
Education: MA in Dutch Language and Culture
Career: Medior Datamanager
Gender: Female
Age: between 30 and 40

Activities
Interviewee is datamanager at a national data archive. The archive promotes sustained access to digital research data for the social sciences and the humanities. For this purpose, it has developed an online archiving system, which can be used by researchers to archive and reuse data in a sustained manner.

After a researcher has uploaded his data and documentation, the datamanager checks the dataset on several aspects: are the data files in an appropriate format, is there enough documentation available, is the dataset comprehensible for other users, is there no disclosure possible of private sensitive information. The latter aspect is in particular important for social science quantitative as well as qualitative data. Information like names, addresses, telephone numbers, and social security number should be removed, other information like postal codes and dates of birth should be recoded.

Problems they would like to solve with language tools
The datamangers have to check the data and metadata for above-mentioned information. For quantitative data in particular the string variables have to be checked for disclosure risks. For large datasets, with a lot of string variables, this can be very time-consuming. At the moment this work is done by hand. For qualitative data (Interviews) the transcripts and the metadata can contain names. In some cases, this information has to be removed. This work is also done manually. It would be great if tools could support these workflows.

Another workflow that would be very useful is topic modelling. The depositor of the data has to label the dataset with keywords. In some cases this is done very sparingly. For some datasets a tool could be useful to extract keywords out the documentation automatically.

Drawbacks of their current work procedures
The current workflow is very time consuming, tools to minimise the amount of work would be very useful.

Would a chaining of tools be of use for their project? (Reference to our demonstrator)
Such a tool could be very useful for the anonymisation of metadata and data. However interviewee states that you have to be sure that the tool is very precise, so no privacy sensitive data is left, otherwise this causes legal implications. The tool needs to recognise even very rare names. She expects that there remains human checking necessary. In the case of anonymisation, the workflow should be secured, and not be open as it is at the moment.

Which other tools should be chained?

Another useful functionality would be the conversion of dates.

Interview 3

Background information interviewee
Education: PhD in history
Career: researcher / postdoc
Gender: female
Age: between 30 and 40

Description of the projects
Interviewee is employed as a researcher at an institute for historical research and documentation on war and contemporary society. She is involved in the European Holocaust Research Infrastructure project (EHRI) for which she is project leader on identification and description of sources relating to the Holocaust (not restricted to a specific place or language). In her previous job she did research into Jewish Antwerp and Antwerp Diaspora during World War 2. The core activity of her work as a historian is to analyse sources, to link them with each other. Within the EHRI project she is also involved in documenting sources with metadata and keywords.

For her PhD-thesis and during her post-doc research the interviewee analysed historical sources. She did quantitative analysis to get a socio-economic overview of the settlement and occupations of Jews, together with a qualitative analysis of relationships. For this research she looked (manual) at data from primary sources, sometimes analysis of existing databanks or created her own standardised datasets in Excel format. Text sources are available in various formats: paper, microfilm, scanned images (not OCRed, often combinations of pre-printed forms, typed text and handwritten text) and photos. One of the outputs of her research was a visual representation overview of results on a map.

In addition to the above-mentioned variety of text sources, within the EHRI-project she also has to deal with different kinds of catalogues with records of archival descriptions: card indexes, stencils, word-files, databases or otherwise. She is team leader of the project team whose task is to identify sources and to enter the already existing associated archival descriptions into the EHRI metadata catalogue without changing them (plus on-site surveys and EHRI-authored descriptions in English). The archival descriptions exist in a variety of languages, including Russian, Ukrainian and Polish. Some sources have multiple descriptions (usually in an original language and an English translation).

Issues to tackle
There are several reasons why automatic linking of multi-lingual descriptions is error-prone; names are sometimes translated or transliterated, often there are various combinations of transliterations and translations in various languages and alphabets.
For example, some alphabets have characters that have no equivalent or more than one equivalent in the Latin script, sometimes different translations or transliterations have conflicting political connotations. Not only geographical places names change over time, people can take on other names as well.

For the linking of sources from various origins, by person name, place name or topic, the above-mentioned issues cause problems. Authority lists would help, but these have to be made. For her previous research interviewee made translation tables herself, because without these the link between names couldn't be recognised.

Within the EHRI project the choice has been made to incorporate the current official name in the thesaurus. The project members, who have to enter the descriptions manually, know from which lists they have to choose the names. These lists were made within another EHRI work package. The thesauri vary in richness: authority lists of individuals and organisations with names and descriptions, authority lists of camps and ghettos with names and translations, and a hierarchical keyword list. Thesauri serve multiple purposes: clarification and assistance for non-English speakers, prevention of misspellings / typos, a guidance for documentalists for preventing them to use their own terms.

The catalogues with archival descriptions of the institutes have no link to this central EHRI thesaurus. A simple experiment linking subject terms from descriptions to the EHRI thesaurus was performed at a late stage in the project, but the result had not enough quality to be used to apply the method to all descriptions.

Drawbacks of their current work procedures
In general, it is very time consuming (which is not available) for researchers to create authority lists and often these private lists are of low quality and not meant to share. If a generally applicable list were available, for example of all geographic names of Europe including historical names, this would be very beneficial for researchers. Interviewee has searched for such lists and has found one for occupations (a division of employment sectors, but from a contemporary perspective). In publications the choice of such a list has to be underpinned. Drawback of current practice in this research area is the fact that each researcher has his own way of documenting, this makes communication difficult.

Another issue is that researchers do not always get easy access to the sources, due to the privacy sensitivity of the content of the collection and hence cannot freely share their databases created with the material.

Possible improvements interviewee mentioned:
General improvement: more standardized procedures for source description
Secure virtual research environment in which researchers can get access to sensitive data for their research. The data remain within the secure environment. Interviewee isn't certain if all historians would accept such a service.

Knowledge of and experiences with language tools.
At university she learned to work (manually) with card indexes, there was no computer training. During her PhD research she learned to use databases for her research. Thereafter she learned how to use Geographic Information System software (GIS). Interviewee states that the use of appropriate tools makes it easier to combine quantitative as well as qualitative analysis within research.

User requirements for tools in comparison to the workflow tool.

For demonstration purposes of the workflow tool an English description about Germany was analysed. The results are not flawless: not all named entities were recognized, or recognised as the right type. Some other terms were recognised as named entity while they are not.

To use such a tool for anonymisation of person names would be very useful. In sources about the Holocaust and the Second World War, such as criminal records, person names should be anonymised. Preferably there would be an option to choose which names should be anonymised.

Another preferable option would be pseudonymisation, for people as well as organisations and geographic locations. Pseudonymisation gives another name (pseudonym) to the chosen named entities. Such a tool would be very useful for EHRI in the context of the for the inventory of criminal files, as a part has to be anonymised and another part may be released (e.g. Hitler, Eichmann and Himmler need no anonymisation or pseudonyms, whereas someone who was a suspect but was not convicted or cleared of charges could have a right to privacy).
When names can be released, linking them to thesauri would be very useful (a task related to Named Entity Disambiguation).

Interviewee 4

Background information interviewee
Education: Master cultural anthropology with minor social science informatics
Career: previous jobs researcher and lecturer, now owner private company in training and research
Gender: Female
Age: between 50 and 60

Activities
Interviewee is involved in various activities in the field of qualitative social sciences:

- Empirical data collection using methods including qualitative in-depth interviews, focus groups, (participant) observation, photography and the consulting and gathering of secondary sources (such as reports, correspondence, journals, etc., as well as online sources).
- Qualitative data analysis with the help of specialist analysis software such as ATLAS.ti.
- Methodological advice on the organisation of research projects.
- Organisation of courses and training related to qualitative research methods, including data collection, data analysis and the use of specific qualitative research software, such as ATLAS-ti.

Current use of software tools
Interviewee considers ATLAS-ti as the *Rolls Royce* under the qualitative data research and analysis software, other applications are NVivo, MAX QDA and QDA Miner. All of them are commercial products.

However she considers software applications only as an aid to structure her data, the analysis of the material is only possible by human work; reading and re-reading the transcripts of interviews (hermeneutic circle). Qualitative research would like to look at the world through the eyes of the respondent, to discover his of her way of reasoning. This is the only way to do justice to the respondent. No research is objective; it is always influenced by the experiences, values and pre-suppositions of the researcher. By triangulation, the use of several analysis strategies, or cross verification from two or more sources, you can ward the subjectivity. This strategy has to be done by the same researcher, the interviewee doesn't believe in inter-coder reliability. This latter is a strategy conducted by quantitative researchers. Only when you start your research, you can ask a colleague to look at your coding's to discover blind spots and bias.

Tools flatten the content, reducing the information. At most very smart software can add some value, but only as an aid. Previously, before she used the computer, she had five paper copies of her transcripts. With cutting and putting the pieces together she made her work material. Now, with the computer, you can endless code the material, try new inspirations and thoughts and sort the material in different ways. This isn't possible without a computer.

What does the tools do?
Atlas-ti makes it possible to look for (combinations of) words in text, such as transcripts. It also makes list of words with frequencies. You can code fragments.
Anonymisation of transcriptions is done by the transcriber him or herself, while making the transcript. The transcription manual provides instruction how to do this.
QDA Miner can produce keywords in context, which isn't possible with ATLAS-ti.

Atlas-Ti helped her to do several kinds of analysis, she hadn't thought herself. For example with ATLAS-ti you can make network views. She used this feature for the analysis of the mentioning of causal relations by respondents within a set of interviews. This gave her extra information she wouldn't have discovered without ATLAS-ti.

Experiences with software tools
Interviewee was trained in computer skills during her training at University. At that time there wasn't yet any software for qualitative research. All the software was focused on quantitative methods. Interviewee considers herself as an early adopter.
She notices a shift in attitude with the new generation of students. They are grown up with computers; they would like to search immediately.

Would a chaining of tools be of use for her research?
She has no need for other language tools, as she doesn't expect more than already is possible with ATLAS-ti. So, a chaining of tools wouldn't be of any use for her research.

Interviewee notices presentations on conferences about tools, which doe the same as the before mentioned big 4 software tools. She wonders why those people not cooperate with the developers of those (commercial) products.

Background information interviewee

Education: PhD in Language Sciences and Applied Linguistics
Career: assistant professor and researcher
Gender: Female
Age: between 30 and 40

Summary of the interview.
She is assistant professor. She is teaching undergraduate classes and carrying out research in Linguistics. Her research interests include: automatic summarization, terminology extraction, and automatic detection of neologisms. Currently, she is working on automatic detection of terminological neologisms in specialized texts (Medicine, Economy, Law, etc.) and in methods for the assessment of their degree of neologicity. She is interested in tools that help her in these tasks. Computer skills: advanced Internet user, familiar with different programming languages and tools for linguistic work. Her experience with DASISH NER tool was not satisfactory, the tool did not work. But she is interested in using this type of tools as well others for Natural Language Processing. She appreciates the availability of these tools in the form of web application.
(The interview was conducted in Spanish, we summarize the main statements).

Projects (neologism detection)

Proyecto sobre detección de neología terminológica. Dependiendo de la investigación que esté llevando a cabo, debo identificar entidades nombradas o secuencias de unidades concretas. Por ejemplo, en el caso de la extracción de terminología, busco secuencias de categorías gramaticales que suelen evidenciar términos poliléxicos, como Nombre + Adjetivo (ej. "cáncer óseo") o Nombre + Preposición + Nombre (ej. "cáncer de mama").

Problems they want to solve with language tools (methods to the automatic detection of neologisms)

Por ejemplo, una de mis líneas de investigación es el análisis del discurso. Aunque existen herramientas para el análisis automático de diversos niveles de la lengua (morfológico, sintáctico, léxico, etc.), existen aún muy pocas herramientas relacionadas con el análisis automático de las relaciones y estructuras discursivas existentes en los textos. También son escasas las herramientas que permiten detectar automáticamente neologismos, por ejemplo.

How do they solve them currently? (She uses different tools, like pos tagging, in different web applications and in local set ups).

Utilizo diversos recursos para trabajar con el texto, siempre dependiendo de la tarea que necesite resolver. Por ejemplo, para anotar morfosintácticamente textos del Corpus Técnico del IULA he utilizado diversas herramientas desarrolladas en el instituto y la herramienta BwanaNet para explotar los datos (http://bwananet.iula.upf.edu/), para obtener automáticamente las estructuras morfológicas y sintácticas en castellano y catalán he utilizado el Freeling (http://nlp.lsi.upc.edu/freeling/), y para anotar relaciones y estructuras discursivas en diversas lenguas he utilizado la RSTTool (http://www.wagsoft.com/RSTTool/).

Si los textos que utilizo están en PDF normalmente empleo dos estrategias: a) realizo la conversión directamente del PDF a texto con el Adobe Acrobat Profesional y limpio el "ruido" con un editor de texto como el Editplus, o b) utilizo el conversor de PDF a texto que incluye el sistema Terminus, un gestor de corpus y terminología (http://terminus.upf.edu/cgi-bin/terminus2.0/terminus.pl), y también limpio el texto con el Editplus.

Would a chaining of tools be of use for their project? (Reference to our demonstrator with problems for uploading files)

Me parecería interesante una herramienta de extracción de terminología o de detección de neologismos.

He intentado probar las herramientas del portal pero no he obtenido resultados. He probado con varios navegadores, subiendo el texto desde un archivo y pegando el texto en el recuadro, usando textos en UTF8 y con otras codificaciones, y enviándome los resultados a diferentes correos electrónicos.

Which other tools should be chained? (She mentions other available tools)
Teniendo en cuenta mis líneas de investigación se me ocurren varias tareas, como son: segmentación discursiva, detección de conectores discursivos, análisis de relaciones y estructuras discursivas. Pero soy consciente de que este tipo de tareas son muy complicadas y requieren muchos procesamientos lingüísticos.
Jorge Vivaldi (investigador del IULA) ha desarrollado junto con Horacio Rodríguez (investigador de la UPC) una herramienta de extracción de términos independiente de lengua y de dominio basada en la estructura de la Wikipedia. Los resultados son excelentes y pienso que sería una herramienta muy útil en el portal.