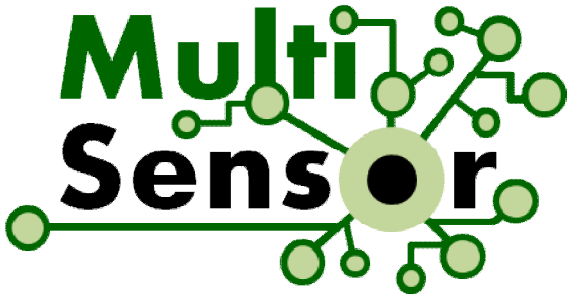


MULTISENSOR

Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation

FP7-610411

D2.1

Empirical study on media monitoring and internationalisation resources

Dissemination level:	Public
Contractual date of delivery:	Month 6, 30 April 2014
Actual date of delivery:	Month 6, 30 April 2014
Workpackage:	WP2 Multilingual and multimedia content extraction
Task:	T2.1 Empirical study
Type:	Report
Approval Status:	Final Draft
Version:	1.1
Number of pages:	172
Filename:	D2.1_EmpiricalStudy_2014-04-30_v1.1.pdf

Abstract

This empirical study identifies the resources and the type of information that needs to be extracted in the project and their encoding types. In addition it reports information retrieval and crawling techniques that could be employed for the extraction of this information.

Co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	20/03/2014	Draft	V. Aleksić (LT)
0.2	03/04/2014	Comments	S. Vrochidis (CERTH), I. Arapakis (BM-Y!)
0.3	15/04/2014	Update	V.Aleksić (LT)
0.4	16/04/2014	Document for internal review	V.Aleksić (LT)
0.5	24/04/2014	Review	A. Moumtzidou (CERTH)
0.6	29/04/2014	Minor revisions based on reviewer's feedback; additional contribution by partners	V.Aleksić (LT)
0.7	29/04/2014	Minor comments	A. Moumtzidou, S. Vrochidis (CERTH)
1.0	29/04/2014	Final document	V. Aleksić (LT)
1.1	29/04/2014	Final document correcting minor typos	V. Aleksić (LT)

Author list

Organisation	Name	Contact Information
Linguatec	Vera Aleksić	v.aleksic@linguatec.de
Linguatec	Gregor Thurmair	g.thurmair@linguatec.de
Linguatec	Reinhard Busch	r.busch@linguatec.de
DW	Tilman Wagner	tilman.wagner@dw.de
DW	Nicolaus Heise	nicolaus.heise@dw.de
pressrelations	Mirja Echkoff	mirja.eckhoff@pressrelations.de
pressrelations	Leszek Blacha	leszek.blacha@pressrelations.de
PIMEC	Teresa Forrellat	TForrellat@pimec.org
PIMEC	Sara Pedraz	export@pimec.org
BM-Y!	Ioannis Arapakis	arapakis@yahoo-inc.com

Executive Summary

This report presents the results and the findings of the empirical study on media monitoring and internationalisation resources.

We have conducted the study in two phases. First, we analysed existing workflows in multimedia monitoring, and user requirements in MULTISENSOR. The crucial questions were which data sources to use, and which kind of information to consider important for the analysis. The findings of our survey and questionnaire showed that the important information is (a) factual information such as named entities, concepts, relations, (b) internal metadata about the sources such as author, location, date, and (c) derived meta information about the sources such as topic, sentiment, key messages. In the second phase we empirically analysed a set of 79 different data sources. The lists of sources were provided by the use case partners and are representative for all three project use cases. The main task was to discover where the required pieces of information can be found. We have analysed newsletter articles, html pages, interactive data bases, single files (PDF), multimedia files (audio, video), and some social media. The analysis confirmed our general assumption that the important data can be retrieved from the metadata, from the content area of the sources (plain text, structured information, images, video, audio, posts) and from their URLs.

This study has been conducted as a descriptive, qualitative analysis. It contains observations on each single data source, not only on where and how the important information is encoded, but also e.g. whether the contents are freely accessible or not.

The subsequent tasks in the project can use the findings of the analysis to decide about the feasible and most appropriate techniques for information retrieval and extraction. The study gives a very brief overview of information retrieval and extraction techniques, without making decision, whether these or other approaches are feasible and appropriate.

Abbreviations and Acronyms

API	Application programming interface
ARPA	Advanced Research Projects Agency
ASR	Automatic speech recognition
CSV	Comma-separated values
GDP	Gross domestic product
GUI	Graphical User Interface
HTML	Hypertext Markup Language
IE	Information extraction
IR	Information retrieval
JPEG	Joint Photographic Experts Group (Image file format)
JSON	JavaScript Object Notation
JSONP	JSON (JavaScript Object Notation) with padding
MPEG	Moving Picture Experts Group (Audio/visual file format)
NE	Named entity
NER	Named entities recognition
OCR	Optical character recognition
PDF	Portable document format
PUC	Pilot use case
PUC1-J	Pilot use case 1 – Journalistic Media Monitoring scenario
PUC1-MM	Pilot use case 1 – Commercial Media Monitoring scenario
PUC2	Pilot use case 2 – SME internationalisation
REST	Representational State Transfer
RSS	Really Simple Syndication (or Rich Site Summary)
RTV	Radio Television
SME	Small and medium enterprises
URL	Uniform resource locator
WP	Work package
XML	Extensible Markup Language

Table of Contents

1 INTRODUCTION

The objective of WP2 in the MULTISENSOR project is to extract knowledge from multimedia input data, and present it in a way that later components can operate on them.

In order to ensure an efficient acquisition and selection of relevant input data out of the rapidly growing online available information, there is a need to understand the structure of the websites and resources to be monitored and crawled and to identify the type of information that can be extracted from there, as well as their encoding types. Hence, the current study has been carried out with two main objectives:

- to provide brief insights into the state-of-the-art techniques and approaches in media monitoring and information extraction
- and to empirically analyse the project relevant data sources, so that the subsequent tasks can decide about the most appropriate techniques for information retrieval and extraction

The main purpose of the study is to facilitate the targeted extraction of news and financial information to support the project use cases. MULTISENSOR has defined two pilot use cases: International media monitoring and SME internationalisation support. The media monitoring use case covers two scenarios: a journalistic media monitoring scenario and a commercial media monitoring use case. The common characteristic of the two scenarios is that both journalists and commercial analysts have to analyse a huge amount of data that come from very heterogeneous media sources, in order to detect and select the relevant ones. On the other hand, the internationalisation use case represents the scenarios where small or medium-sized companies (SMEs) that intend to reach foreign markets need not only reliable market analysis for their target country, but also a risk assessment, comparison with other markets and eventually also a decision support. The lack of knowledge about the relevant sources of information, as well as the language barriers are the main obstacles to them.

The outcome of the study will support the extraction tasks in WP2. Furthermore, Task 3.1 “Indicators for media monitoring and internationalisation”, Task 4.1 “Topic-based modelling”, as well as Task 7.2 “Crawlers and data channels infrastructure” will rely on the clues and outcomes of the study.

To perform this empirical study, a representative list of resources, distributed as evenly as possible along the use cases, has been provided by the project use-case partners and has been analysed. The study tries to determine the characteristic common features of the data sources of interest and how meaningful and relevant content, as well as important metadata, are conveyed.

This report is organised as follows: Section 2 lists the basic research questions of the study and describes our approach. Section 3 reviews our findings about the existing workflows and practices in media monitoring tasks, including also the description of their common workflows provided by the MULTISENSOR use case partners. In Section 4 we describe the data set on which the study is based. In Section 5 we report on our empirical analysis of the MULTISENSOR data set, describe in detail our findings related to the different resources, and present also the analysis of social media and networks. Section 6 reviews the applicable

information retrieval and information extraction techniques. Results and conclusions are given in Section 7. Appendices contain, among other things, detailed analyses for each single source.

2 QUESTIONS AND APPROACH OF THE STUDY

The amount of data in the World Wide Web is rapidly growing, which in turn led to an increasing demand for efficient identification and extraction of relevant data. More and more professionals and companies are recognising the importance and the urgency of not only a targeted search for the relevant information sources, but also of an efficient, automated aid in understanding, aggregation and semantic interpretation of them. One of the main objectives of the MULTISENSOR project is to support such tasks. The question arising is, what knowledge about the information sources is needed up front, in order to provide this support in an optimal way. The current study should provide some answers.

In order to analyse and draw safe conclusions regarding the data required for MULTISENSOR use cases and given the lack of theoretical knowledge regarding the characteristics of the resources containing interesting information for the media monitoring use case and the SMEs internationalisation use case, we realised the following procedure that involves two main phases. The first phase involves performing a survey on existing resources and workflows used for all use cases and includes the preparation and completion of a questionnaire by the expert users. The second phase on the other hand, involves the realisation of an empirical study on the resources provided also by the experts.

2.1 Survey on media monitoring and internationalisation resources and workflows

In order to better understand what is the state-of-the art in media and market monitoring approaches, and which questions might be important for all MULTISENSOR use cases, our first intention when designing this study was to get a picture of what are the common applications and workflows in media and market monitoring, analysis, evaluation, and decision support, from the first step (data selection) to the last one (presentation of the results to the user).

First, we collected information from the use case partners and a few selected experts from the MULTISENSOR user group by sending them a questionnaire. Since the purpose of the questionnaire was to get a general outline of the situation, we decided to design it as an open format questionnaire containing a small number of general questions (cf. below). The selection of the questions was based on the first draft of the user requirements defined for the project within WP8 (in particular regarding the type of information required for the modules forming the MULTISENSOR architecture), as well as on an initial literature review.

The analysis of the collected responses helped to gain an impression of state of the art on the one hand, and also resulted in an extended catalogue of several more specific questions on the other hand, which then served as a basis for the subsequent intensive literature research and investigation of common practices in media monitoring and internationalisation (more detailed description is given in Chapter 3.2).

Questions

Thus, the following basic questions have been considered important and have determined the survey phaseof the study: What are the state-of-the-art workflows and practices in the media and market monitoring tasks, and in SMEs internationalisation support? In particular:

- What are the typical application scenarios?
- Which information and data sources are important?
- What type of information is usually being extracted?
- How the information is aggregated, summarised and archived?
- How does the interaction with the user work and what is considered important?

Approach

In order to perform the first phase of the study, we have consulted the related literature, other related empirical studies such as (Kasper, et al. 2010), (Uhlmann, 2011), and (Siavash, 2011) and in general resorted to methods used in similar cases. Thus, towards this direction, we formed a questionnaire that covered the needs of MULTISENSOR use cases. This questionnaire contained the above-mentioned questions that weresent to the MULTISENSOR use case partners and experts from the user group, in order to learn more about their everyday practices, as well as about their requirements for the current project.

Procedure

Accordingly, the two main tasks in the survey phase were:

- Preparatory task that involves:
  - literature review
  - studying of the initial user requirements (WP8)
  - creation of the questionnaire
- Analysis of the questionnaire responses (cf. Appendix C)

One of the main outcomes of the survey phase was lists of relevant data sources for each MULTISENSOR use case. This was the starting point for the second phase, the empirical analysis of each item from the dataset.

2.2 Empirical analysis of the MULTISENSOR dataset

To efficiently extract relevant information according to a topic or a keyword requires profound knowledge of the data structure and adequate information extraction techniques on the one hand. On the other, it needs also optimised techniques for a targeted data retrieval (crawlers etc.), as well as for filtering and selection of the data from a huge data pool (optimisation through intelligent parsing and data separation techniques, such as HTML parsing, format conversions, data scraping etc.). Also in the case of using only a predefined and limited set of data sources, as is the case of the MULTISENSOR pilot projects, it is still a challenge to optimise the use of information resources and to manage the quantity of the relevant information to be stored.

An inevitable step in tailoring all those tasks for concrete use cases is to collect empirically gained insights into the structure of the relevant sources, and thus to support them in finding an optimal way to detect, extract and store the data in a controlled manner.

Accordingly, the second task of this study was to analyse a predefined set of relevant resources, and to empirically discover where the important information is placed, how it is encoded there, and to review some adequate techniques for their retrieval and extraction.

Questions

The following questions have been considered important and have determined the second phase of the study:

1. Given the initial set of data sources to be investigated, which information important for MULTISENSOR pilot use cases can be extracted from them? Based on the knowledge gained during the phase one, we considered it necessary to concentrate on:
   - For Use case 1: the content of the news articles themselves, the related multimedia content (images, videos, graphs), and the meta information such as language, location, author and source of the news articles
   - For Use case 2: financial and statistical information about the targeted markets, information about the products and main actors (producers, distributors and potential customers)
   - Social media information, relevant for both use cases: posts and responses themselves, as well as the meta information such as the number of posts, number of favourites, followers, location, visits, language, author etc., depending on what data the social media platform makes available.
2. Where and how this information is encoded? The hypothesis was that such information can be retrieved from the URL addresses, the web content itself (text, images, multimedia data), from the metadata of the websites, and in the case of the social media by directly using their APIs. We also looked at external web statistics.
3. By which extraction techniques could this information be retrieved and extracted? What are the state-of-the-art information retrieval and extraction techniques and tools? In particular for:
   - Web crawling
   - Parsing, data pre-processing and noise reduction
   - Information extraction

Approach

Being the empirical part of the study, the second phase included a closer investigation of each single data source. In particular, it comprised:

- First, an initial, rather shallow, analysis of each source: kind of source (news text, interactive data base, single files, search engine results, multimedia files etc.), URL, accessibility (free or restricted), relevance for our use cases. We originally also tried to find more about how relevant and reliable each single source is in its local and in global context, by looking at external web analytics providers such as e.g. Alexa(www.alexa.com).But, in the end, we have not included the concrete numbers into the analysis table, since they change from day to day.

  The results are summarised in Appendix A.

- In the second cycle, a deeper determination of the important content parts: Which kind of content do the single sources (text, audio, video, different file formats etc.) contain? Where is the meaningful information located and how can it be accessed? Which kind of meta information is important (time, location, language, author etc.), where is it located, and how can it be accessed?

The extensive results of the empirical study for each source can be found in Appendix B.

Generally, empirical research methods can be divided into two categories (Siavash, 2011):

- Quantitative methods which collect numerical data and analyse them using statistical methods (they tend to be more appropriate when theory is already well developed, or for purposes of theory testing and refinement)
- Qualitative methods which collect information drawn from observations, interviews and documentary evidence and analyse them using qualitative data analysis methods (they tend to be more appropriate for theory building and in exploratory research)

Based on this distinction, the research method that we conducted in this study was qualitative research, since it best fitted with our research questions and study objectives. In the context of this study, the main research questions were: (a) What are the different types of information to be found in the dataset? (b) How those different types of information are encoded? Both of them are exploratory questions.

Procedure

In order to achieve the study objectives, we adopted the analytic strategy and techniques described in (Creswell, 2013) to analyse the data and to answer the research questions. Creswell describes five steps as a generic analytic process for qualitative inquiries:

- Step 1 – Organise and prepare the data for analysis (with the help of experts we have selected an appropriate dataset, balanced in terms of data and languages relevant for each use case)
- Step 2 – Read through all the data (first, we performed a shallow analysis of the entire dataset – results can be found in Appendix A)
- Step 3 – Begin detailed analysis with a coding process (then we performed a deeper analysis of each single source to identify the most important information and their encoding types – results in form of single analyses can be found in the Appendix B)
- Step 4 – Use the coding to generate a description of the setting as well as categories (by applying this step we have identified several categories of data sources with similar features according to either formal or/and content criteria such as newspapers, multimedia, single files, interactive data bases etc.)
- Step 5 – Make an interpretation or meaning of the data (after having performed the analysis and the validation of all data sources we have drown the conclusions for each category identified in the dataset – cf. the conclusions in Chapters 5.1 – 5.6)

Accordingly, the main procedure steps during the empirical analysis of the data were to:

- Select the initial data set and divide it into an analysis and a test (hold-out sample) part
- Apply several analysis steps to each data source from the analysis part (cf. above)
- Draw conclusions
- Validate the findings by applying them on the hold-out test set
- Consider adequate extraction techniques

Validation

In order to validate our findings we used the *hold-out sample* approach ("The hold-out sample is the subset of the data available to a data mining routine used as the test set"[1]).

First we used part of the dataset to analyse it and extract the findings, then we validated them by verifying the knowledge gained during the first cycle of analysis on the rest of the dataset.

[1] http://www.statistics.com/glossary&term_id=774

3 MEDIA MONITORING AND INTERNATIONALISATION RESOURCES AND WORKFLOWS

Media monitoring is an important marketing instrument and basis for decision making, used not only by big players, but also by small and medium enterprises. Consequently, there is a rapidly growing number of tools and providers that support the companies in this task. Also journalists, analysts, and authors face the problem of enormous data volumes to be analysed, understood and aggregated. For them as well, there are huge savings to be made in terms of time, labour and costs by automating and simplifying their editorial and authoring processes.

3.1 Questionnaire

As already mentioned in Section 2, for a better understanding of those processes and the requirements with regards to the supporting tools, and therefore as a preparation step for the empirical analysis of the project data, we sent a questionnaire first to the MULTISENSOR use case partners, and then to a few selected users from the MULTISENSOR user group, in order to learn about the typical scenarios and workflows, data resources, techniques for their aggregation and summarisation, and the communication of the results to the user.

The focus was on describing and analysing the commonly used information resources by the use case partners and their existing workflows. The answers were intended to serve as a basis for the technology partners to understand the status quo and how it works today, and for the entire consortium to then define the resulting priorities. The second main intention was to ask the users to list resources that from their point of view might be relevant for MULTISENSOR.

The complete answers can be found in Appendices C (Questionnaire responses) and D (list of resources relevant for MULTISENSOR).

3.2 Follow-up survey and conclusions

Building upon the outcome of the questionnaire analysis, and in order to understand it in a broader context and to gain even deeper insights into most common work routines and practices in the field of media monitoring and internationalisation, we additionally reviewed a couple of market research studies such as (Kasper, et al. 2010) and (Uhlmann, 2011) addressing more precisely the following questions:

- What are the most frequent monitoring application fields?
- What are the typical application scenarios?
- Which data sources are usually being monitored?
- Which filtering techniques for the data sources are applied?
- How the data are stored?
- How the data sources are formally analysed and described (meta information)?
- How the stored data are (pre)selected and prioritised for the deeper analysis?
- Which kind of information is then identified and extracted from the data?
- Which conclusions are derived from the identified information?
- How the analysed data are then structured and formally presented?

- What are the main functions of the query component / user cockpit / dashboard?
- What are the main operating principles?

Table 1 summarises the most relevant findings and answers to the questions addressed by the questionnaire and our survey:

Most frequent monitoring application fields
• Reputation management • Event detection • Risk management • Crisis management • Issue management • Observance of the market and competition • Competitive management • Market analysis • Trend analysis • Influencer detection • Customer relationship management • Product management • Innovation management
Typical application scenarios
• Market and product monitoring and interaction with the costumers • Product launch control • Measuring the public opinion in relation to a topic • Measuring the tonality in relation to a topic • Control and/or evaluation of corporate communications • Identification and penetration of new target groups • Stakeholder management • Early recognition of risks and potentials • Support in finding solutions • Derivations of recommendations for action • Decision support for internationalisation, analysis and comparison of target markets • Profiling on the basis of personal data • Job descriptions / CV matching • Public relations and press activities • Support for (simplifying of) editorial and authoring processes
Data sources
• Print media • Online • TV / radio • Photo and video portals (e.g. YouTube, Flickr, Vimeo) • Social media: ○ Blogs

- Fora
- Twitter
- Social networks (e.g. Facebook, LinkedIn, Xing)
- Consumer portals
- News groups
- Chat rooms
- Message boards
- Social bookmarking services

- Data bases (press data bases, business data bases)
- Press services
- News aggregators
- Shopping and Paytip portals
- Other information portals (Eurostat, EurLex, national authorities' info portals)

Filtering techniques for the data sources

When collecting and crawling data, which criteria are applied to filter them before storing?

- Date and duration of the publication
- Geographical limitations (only Europe; only single countries/regions)
- Language
- Domain
- Author
- Frequency of (predefined) key words
- Alert functions, such as: when a threshold value is exceeded; when a time limit is reached

Frequent domains

- Automotive
- IT / Telecommunication
- Finance / Insurance
- Media / Advertisement
- Pharmacy / Medicine
- Biotechnology
- Food industry

Meta information considered to be important

News/Press:

- The printed and sold circulation of the source
- The coverage (when, where, how often, how many readers …)
- The reader structure (age, gender, income, educational level …)
- Clustering according to topic, media type, geographical reference

Social media and networks:

- Main influencers
- Followers

• The followers structure (age, gender, social status)
Important information to be identified and extracted from the data
• Key words • Tonality (positive/negative/neutral) • Named entities (geographical names, companies and organisations, person names, temporal expressions, numerical and currency expressions) • Co-occurrences • Concepts and relations between entities • Topics • Events
Conclusions derived from the identified information
• Company / sector profile (turnover, market share, competitors, …) • Topic profile (involved actors, attitudes, arguments…)
Storage and presentation of the analysed and structured data
• Generation of semantic networks • Population of ontologies • Construction of association graphs • Data bases • Summaries with highlighted keywords • PowerPoint presentations or Excel Sheets • Development of a customer-specific code book for further detailed data analysis and storage • Oral or written reports to the customers • Consulting rounds
Main functions of the query component / user cockpit / dashboard
• Administration of user profiles • Visualisation / presentation of the results to the user (bars, columns, lines, pie charts, tree-maps, maps …) • Interaction with the user (e.g. change and update of the meta information, definition of favourites) • User defined update of sources (possibility to exclude some sources from the analysis, or to include their own sources) • Search filters (media type, country …) • Access to the original data • User-predefined topics
Main operating principles

• Data collection:
  ○ fully automatic (crawlers, APIs)
  ○ manual (telephone and email surveys, interviews, literature search)
  ○ Accessing of external databases that offer e.g. full text articles via keyword

searches (for free or against payment) ○ Using data from third-party aggregators (e.g. Google blog search) ○ RTV clips can be gathered by purchasing these from specialised vendors, who use proprietary technologies for capturing, speech-to-text, search and audio/video-cutting • Information extraction: fully automatic / manual / automatic with manual corrections • Creation of reports and summaries: mostly manual, based on automatically acquired information

Table 1: Questionnaire and Survey Findings

Another important outcome of this part of the study was that country- and cultural differences should be taken into account from the very beginning of a media monitoring process. Already the appropriate choice of the data sources to be monitored in different societies is vital to ensure the relevance of the results. Thus, blogs held a much more important position in the shaping of public opinion in the USA than in Europe. In contrary, the fora are very important e.g. in Germany (potentially relevant information for PUC2). Furthermore, also the use of a national search engine, instead of the international one such as Google, can lead to more relevant results for a search query.

4 MULTISENSOR DATASET

4.1 Dataset

With the aim of performing the empirical analysis on a set of resources representative for all MULTISENSOR pilot use cases, we created data sources for each use case separately. The use-case partners provided lists of sources which are use-case relevant in terms of kind of information they contain and languages they cover. For PUC1, the lists have been provided by DW (the journalists scenario) and by pressrelations (the commercial scenario), and for PUC2 (SME internationalisation) by PIMEC. The complete lists can be found in Appendix D.

Here we give a brief overview of the dataset statistics:

Use-case relevance:

Total number of different sources:	**79**
PUC1 – journalist media monitoring scenario relevant sites:	39
PUC1 – commercial media monitoring scenario relevant sites:	33
PUC2 relevant sites:	23

Language coverage:

Bulgarian	3
English	50
French	17
German	35
Spanish	16

Encoding types (many sources contain different formats; the following table should only roughly illustrate the distribution of different encoding types):

Sources containing news articles (online newspapers in html format)	25
Websites of governmental authorities, international organisations, academia, and regulation bodies	30
Interactive portals / data bases / procurement archives	5
Single PDF files + other sources containing PDF	23
Sources containing multimedia (audio, video, images, e-books)	21
Blogs (including other sources containing blogs)	13
Other web content (Google search, Wikipedia)	2
Social media (e.g. Facebook, Twitter, Reddit)	Many links to them

More details can be found in Appendices A and B.

Appenidix A lists all sources, with:

- their URLs
- the languages covered
- short descriptors of the content type
- internal links to the project use cases
- accessibility (free, restricted)

Appendix B gives a more detailed description of each single source.

4.2 Information to be extracted

One of the central tasks of the project will be information extraction which is a task of finding factual information in, mostly unstructured, data sources and to structure it. Which information is commonly being identified by state-of-the-art IE tools, usually corresponds to a predefined set of concepts within a specific domain. Generally, an IE task is defined by a set of resources to be analysed following clearly specified information needs.

For the purposes of this empirical study, after having scanned the set of resources, and based on the evaluation of the questionnaire responses, we went over to the identification of the information needs. We compiled the requirements for the kind of information to be extracted into three main groups:

1. Factual information to be extracted from the content part of the sources:

Names:

- countries, cities, addresses
- companies, producers, competitors, marketers, distributors
- politicians, decision makers, opinion leaders
- brands, products
- dates, time periods
- currencies, prices

Other textual and numerical concepts:

- cardinal numbers (for size, density, growth, capacity …)
- fractions
- professional occupations, job titles, titles
- languages, religions
- ingredients, materials
- regulations, laws

Relationships between them, roles, functions, and attributes:

- general information about a country such as its capital, currency, leaders, religion, language …
- economic information about a country such as its population size and density, per capita income, GINI coefficient, GDP growth, inflation, market size …
- bilateral relation between countries
- foreign market sectors, import/export

- regulations in a country, law certifications

2. **Meta-information about the sources (explicit metadata)**
   - all media: country, language, media type (e.g. daily newspaper, popular press, advertising journal), date, headline (and text), author, Advertising Value Equivalence (AVE)
   - print: printed edition, distributed edition, print reach, page start, number of pages, number of illustrations
   - online media, blogs, forums: visits
   - RTV: viewers, duration (min:sec)
   - Twitter: followers
   - Facebook: likes
   - Google+: +1
   - YouTube: views
   - Backlinks within articles (used to display connections of articles or tweets etc.)
3. **Our (i.e. user's) interpretation of the content (derived meta information)**
   - main topics
   - tonality
   - judgment over exclusivity or weight
   - key messages adherence
   - sentiment

Thus, the next cycle of the empirical observation was to find more detailed answers to the questions: Where and how are the required pieces of information encoded? Which extraction techniques would be applicable? These issues will be treated in more detail in the next two chapters.

5 EMPIRICAL ANALYSIS OF THE MULTISENSOR DATASET

After a careful observation of a dataset, we identified that the keyplaces were all the interesting information mentioned earlier can primarily be found are metadata, URL, and the main content of the webpage. Thus, the empirical part of the study has been conducted in several analysis cycles in order to prove the hypothesis that such information can be retrieved from the metadata, the web content and the URL address of a website, and to closer describe how the information are encoded and distributed there.

Metadata

Metadata describe other data. They provide information about one or more aspects of the data, such as:

- what program was used to create them
- purpose of the data
- date of creation
- author
- location on a computer network where the data were created
- standards used
- description of the page
- keywords relevant to the page

For example, metadata of a text document may contain information about how long the document is, who the author is, when the document was written, what is the structure of the document, and a short summary.

An image may be described by metadata such as how large the picture is, the colour depth, the image resolution, when the image was created, and other data.

Web pages often include metadata in the form of meta tags. HTML web pages allow inclusion of a variety of types of metadata, from basic descriptive text, dates and keywords to further advanced metadata schemes such as the Dublin Core[2], e-GMS[3], MPEG Multimedia Metadata[4], and AGLS[5] standards.

Metadata may be included in the page's header or in a separate file. Most search engines use the content of these tags when adding pages to their search index.

URL

Definition: URL stands for Uniform Resource Locator. A URL is a formatted text string used by Web browsers, email clients and other software to identify a network resource on the

[2] http://dublincore.org/

[3] http://www.esd.org.uk/standards/egms/

[4] http://www.multimedia-metadata.info/

[5] http://www.agls.gov.au/

Internet. Network resources are files that can be plain Web pages, other text documents, graphics, or programs[6].

URL strings consist of three parts (substrings):

- network protocol
- host name or address
- file or resource location

These substrings are separated by special characters as follows:

protocol :// host / location

URL Protocol: The 'protocol' substring defines a network protocol to be used to access a resource. These strings are short names followed by the three characters '://' (a simple naming convention to denote a protocol definition). Typical URL protocols include http://, ftp://, and mailto://.

URL Host: The 'host' substring identifies a computer or other network device. Hosts come from standard Internet databases such as DNS and can be names or IP addresses.

URL Location: The 'location' substring contains a path to one specific network resource on the host. Resources are normally located in a host directory or folder.

URLs contain very often useful meta information such as date, language, title, and topic of a data source.

Content

As content we consider everything that we can see on a website, in a file, or in a database. Contents are in our dataset video and audio files, text articles, images and their descriptions, social media posts and all other visible information accompanying them (author, number of visits etc.).

We have analysed content, metadata and URL for each single source, and have tried to find common characteristics for different groups of sources represented in our analysis corpus (online newspapers, other html websites, PDF files, interactive portals and data bases, audios and videos, blogs, social media).

In the sequel, we analyse different type of sources, such as online newspapers, portals, and blogs and provide information regarding the type of information that different sources contain and where it can be found. We start each section with an example analysis, where we demonstrate our analysis approach step by step. In the second part of each section, we then describe our findings and summarise the common characteristics of the respective type of source.

5.1 Online newspapers

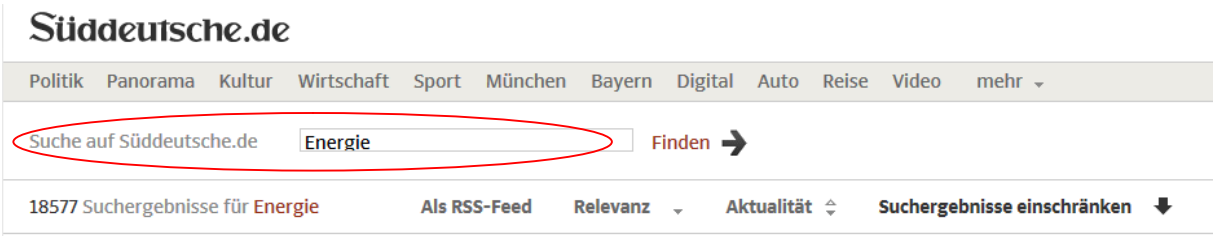
We first provide an analysis of a typical example of an online newspaper and then present the findings of a study realised on several online newspaper sites. The study involves the three main parts of the source that interest us which are their content, metadata and URL.

[6] http://compnetworking.about.com/od/internetaccessbestuses/g/bldef-url.htm

5.1.1 Sample analysis

As the typical example for a newspaper analysis, we take *Die Süddeutsche Zeitung* (www.sueddeutsche.de). The requirement is, for the PUC1 journalistic scenario, to retrieve and analyse all articles on the topic "Energie".

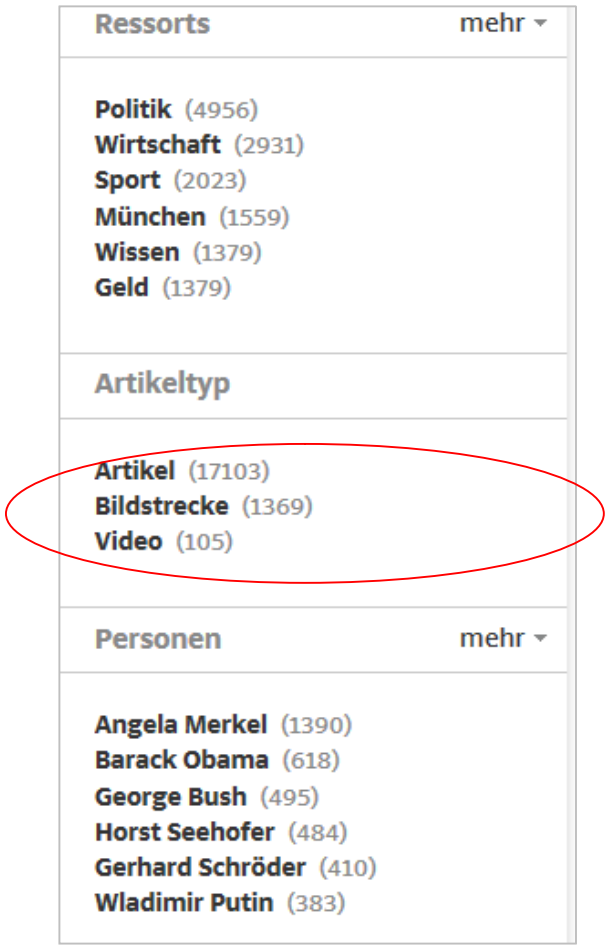
1. First, we check, whether the web page has a search function. If yes, we search for "Energie":



2. Next, we check whether our search query is represented in the URL. If yes, this more specific URL could be used as the seed URL for the web crawler later on:

```
http://suche.sueddeutsche.de/?query=Energie&Finden=Finden
```

3. Further, we look at the search results. They can be restricted according to:
   - Resort
   - Document type
   - Person names
   - Locations
   - Keywords
   - Companies



If some of the search restrictions are applied, this is represented in the URL as well:

http://suche.sueddeutsche.de/query/Energie/sort/-news/drilldown/%C2%A7ressort%3A%5EWirtschaft%24

Later on, for the web crawling, we even could consider using this kind of very specific URLs as seed terms, in order to minimise the recall of non-relevant search results (from other resorts etc.)

In our example, we can see that the 18.577 query results are distributed as follows:

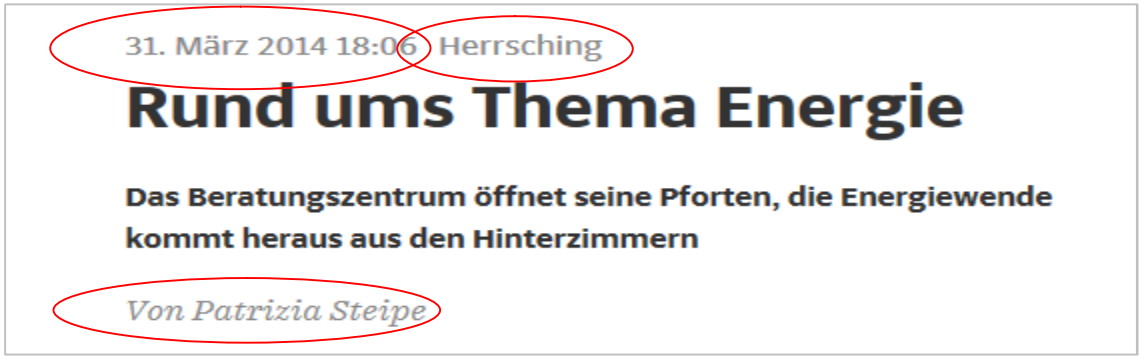
- Articles (text) (17.103)
- Slide shows (1.369)
- Videos (105)

4. In the next step, we look at sample files from each different encoding type (text, slide shows, and videos) in order to find out, how the relevant information is encoded. We look at the content part, the metadata and the URL.

Article

Content: the important content can be found in the title of the article and in the text (keywords etc.). Both will be subject of information analysis and extraction tasks.

Also important metadata (date, time and author) can be extracted from the article itself.


31. März 2014 18:06 Herrsching

Rund ums Thema Energie

Das Beratungszentrum öffnet seine Pforten, die Energiewende kommt heraus aus den Hinterzimmern

Von Patrizia Steipe

A deeper look at the retrieved articles reveals the following:

- Before storing it, the html file should be parsed and cleaned by removing advertising, imprint, links to other articles etc. (boilerplate removal)
- The results contain a huge number of thematically non-relevant articles (with "Energie" in all other meanings than electrical power). As a solution we can consider using more specific keywords, such as "Windenergie, Windkraft, Energiewende, Atomausstieg, Atomenergie, erneuerbare Energie, regenerative Energien, Stromerzeugung, Energieeinsparung, Energiekonzern …"

Metadata: From the metadata we could extract keywords, date and time, URL of the article (and of the images, if available on the page), content type (article), title, brief content description, and the language, as the most important information for our purposes:

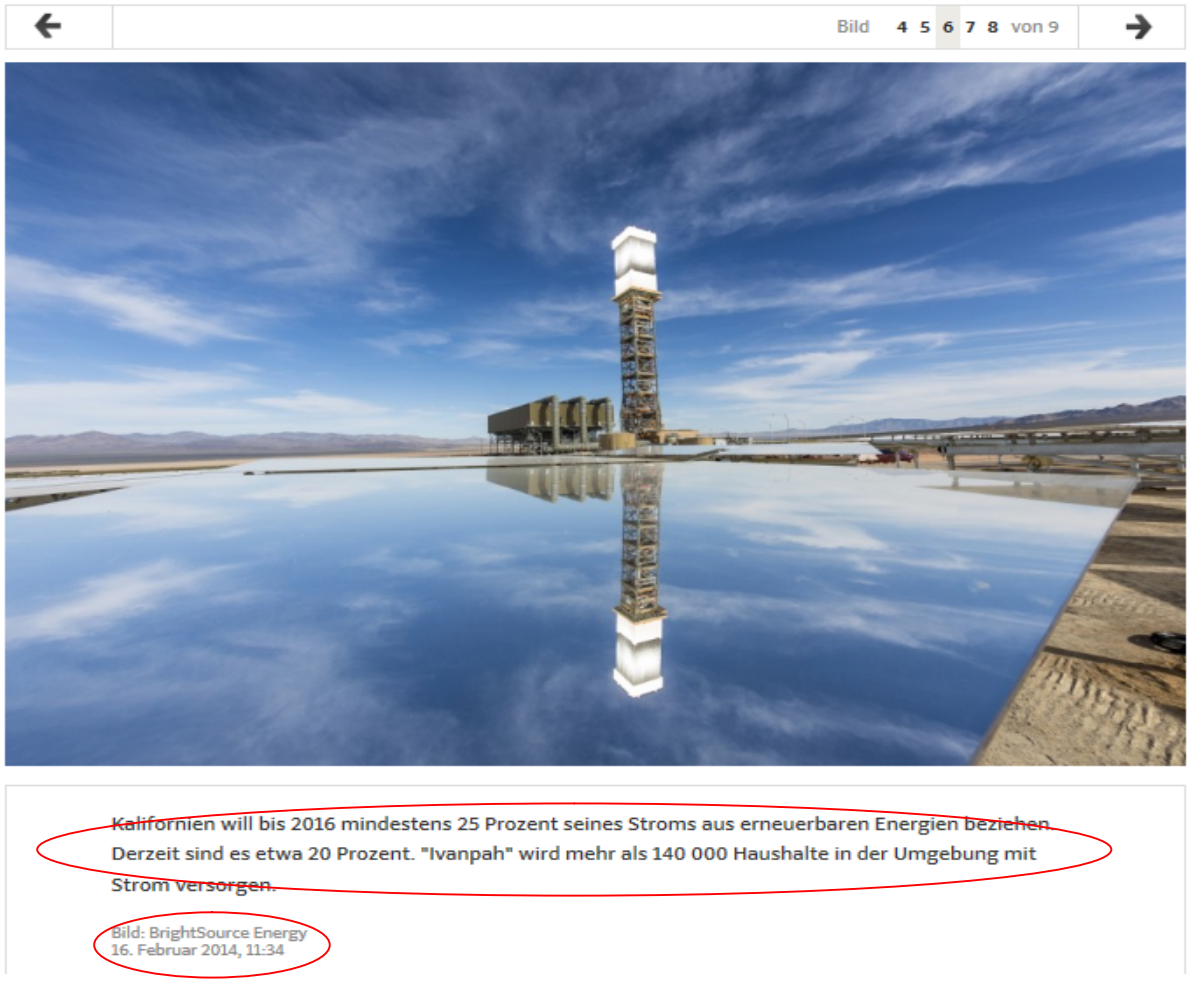
```
<meta name="keywords" content="Herrsching, Karl Roth, S&uuml;ddeutsche Zeitung, SZ">
<meta name="news_keywords" content="Herrsching, Energiewende, Karl Roth, Landratsamt, Weltfrieden, &Ouml;kostrom, Starnberg">
<meta name="robots" content="index,follow,noarchive,noodp">
<meta name="last-modified" content="Mo, 31 Mrz 2014 18:06:29 MESZ">
<meta property="og:url" content="http://www.sueddeutsche.de/muenchen/starnberg/herrsching-rund-ums-thema-energie-1.1926442">
<meta name="twitter:card" content="summary">
<meta name="twitter:site" content="@SZ">
<meta property="og:type" content="article">
<meta property="og:title" content="Herrsching &ndash; Rund ums Thema Energie">
<meta property="og:description" content="Das Beratungszentrum &ouml;ffnet seine Pforten, die Energiewende kommt heraus aus den Hinterzimmern">
<meta property="og:image" content="http://polpix.sueddeutsche.com/bild/1.1926446.1396281989/900x600/herrsching.jpg">
<meta property="og:site_name" content="Süddeutsche.de">
<meta property="og:locale" content="de_DE">
```

URL: from the URL we could extract the title of the article and the language:

http://www.sueddeutsche.de/muenchen/starnberg/herrsching-rund-ums-thema-energie-1.1926442

Slide show

Content: The important content is in images, as well as in their descriptions.



On the website we can identify: the image itself, a short description under the image, the source of the image and date and time in the bottom left corner.

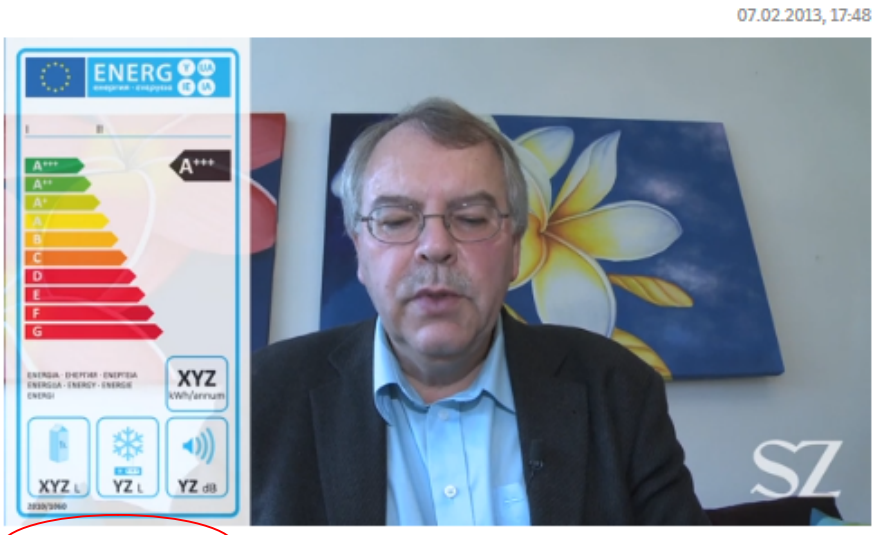
Metadata: from the metadatawe could extract the same information as in the article-metadata (cf. above): keywords, date and time, URLs of the article and of the images, title, brief content description, and the language. There is no hint for the distinction between an article and a slide show.

URL:the URL contains the title and the language (domain).

http://www.sueddeutsche.de/wirtschaft/sonnewaermekraftwerk-in-kalifornien-spiegel-in-der-wueste-1.1888297-6

Video

Content: The important content is in videos, as well as in their descriptions.



Energiekosten im Haushalt

Welche ~~Informationen~~ gibt Ihnen das Energielabel?

Daniela Dau und Patrick von Frankenberg

Die Angaben auf Energielabels sind für viele Verbraucher verwirrend. Was die bunten Balken, sowie die Buchstaben- und Zahlenkombinationen auf den Energielabels von Neugeräten bedeuten.

On the website we can identify: the video itself, its title, the names of the authors and a short description under the image.

Metadata: from the metadatawe could extract the same information as in the article-metadata (cf. above): keywords, date and time, URLs of the article and of the images, title, brief content description, and the language.

```
<meta name="medium" content="video">
<meta name="video_type" content="application/x-shockwave-flash">
```

Additionally, there is information available about the medium type (video) and the video application (x-shockwave-flash):

URL: The URL contains the title and the language (domain).

5.1.2 Conclusions

We have analysed more than 20 online newspapers and magazines. As described above in the sample analysis, we looked for: which type of content they offer, do they have a keyword search function and/or a topic search, how can the queries be filtered, and what are possible extraction ways for the information encoded there.

Content

In the content part of the newspapers and magazines mostly two main areas can be recognised, the header and the body. The header part very often contains relevant meta information such as title, author, date and time, and the location of the publication. Those parts could be reached and separated by HTML parsing.The most important content from newspapers and magazines is in their content containers in the body part, represented as

news texts, videos, images, blogs etc. to be analysed by information extraction and video, image and audio processing techniques.

Regarding the use case coverage, we have observed that the amount of content relevant for the journalistic use case (keyword “energy”) is very high in newspapers, ranging from a couple of thousands of articles to several hundreds of thousands. However, the search results for only the keyword “energy” contain a big number of less or non-relevant matches (“life energy”, sports etc.). In order to select the really relevant content before storing it into the repository we recommend using more specific keywords or combination of keywords. For example, instead of “énergie” as the keyword, much more relevant results are retrieved from the French sources by using the following, more specific, search terms: “UNFCCC, bouquet énergétique, changement climatique, consommation d'électricité, gaz carbonique, panneaux photovoltaïques, panneaux solaires, production d'électricité, rejets de CO, réchauffement planétaire, transition énergétique, économie d'énergie, énergie nucléaire, énergie solaire, énergie éolienne, énergies renouvelables, énergies vertes, réacteurs nucléaires, consommation d'énergie …”.

For the commercial use case (product monitoring with the examples of “Kitchenaid”, “Liebherr” and “AEG”) we realised that the number of search results is much lower in the newspapers than for the journalistic scenario, and really low in women’s and men’s magazines (popular press). Moreover, the keywords being ambiguous (Liebherr is a usual German surname, AEG also refers to an American music promoters), the results from the popular press contain more non-relevant matches than relevant ones. A combination with other more specific keywords would help to better retrieve relevant content. This in particular applies to the popular press. The trade press mostly contain more relevant articles (shopping, product comparison, product reviews). The main content is often in images, with very short accompanying text descriptions.

For the content distillation we will use the information extraction from text, as well as video, image and audio processing.

Metadata

In order to identify important formal information, we look at the metadata. There, we always will find some useful information pieces such as: title, language, date, author, keywords, description, URLs to images etc.

Many newspapers follow a metadata scheme such as the Open Graph protocol[7] or Dublin Core Scheme[8], like in the following examples:

```
<meta property="og:locale" content="en_us"/>

<meta property="og:type" content="video"

<meta property="og:image" content="http://decor8blog.com/wp-content/uploads/2008/07/160x160-decor8blog-grey-1.jpg"/>

<meta name="DC.date.issue" content="2014-04-07" />
```

[7] http://ogp.me/

[8] http://dublincore.org/

<meta name="DCSext.rChannel" content="Blogs" />

<meta name="DCSext.ContentHeadline" content="Financing more solar energy - MuniLand" />

<meta name="DCSext.rAuthor" content="Cate Long" />

<meta name="dcterms.creator" content="natalie.schuckardt" />

<meta name="dcterms.language" content="de" />

Metadata can be extracted by using standard metadata harvesting techniques.

URL

From the URLs of the newspaper articles very often the following information could be extracted: language, date, title, and topic. In view of the fact that a wide range of different URL formats can be observed, a URL parsing approach to be specially tailored for each website would be needed in order to extract the information.

5.2 Websites of authorities, organisations, academia, and regulation bodies

In this type of source we have grouped all websites from our dataset other than newspapers and magazines. Most of them are websites of governmental authorities, national or international organisations, academic institutions, and regulation bodies, but also web presentations of private persons or smaller companies or institutions. There are certainly far less commonalities among them than among different newspapers, but we did not consider it necessary to differentiate them further.

5.2.1 Sample analysis

As the example for a website analysis from this group, we will take the website of *German Federal Ministry for Economic Affairs and Energy* (http://www.bmwi.de/).

The requirement is, for the PUC1 journalistic scenario to retrieve and analyse all issues on "energy", press materials, and 'Mediathek' (multimedia archive), in all available languages.

1. First, we check in which languages the website is present, and how the language codes are encoded. We find out that the language codes can be found in the URL, for German, English and French:

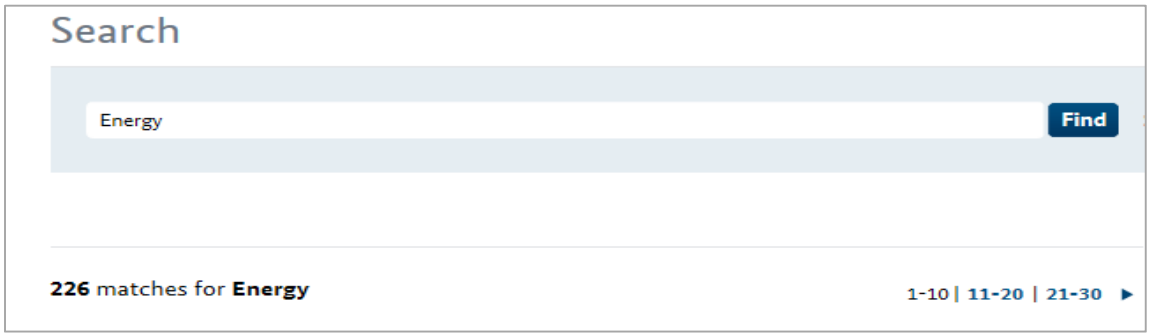
   http://www.bmwi.de/DE/root.html
   http://www.bmwi.de/EN/root.html
   http://www.bmwi.de/FR/root.html

2. Second, we check where and how on the page the issues on the topic "energy" can be found:
   - First, there is a pull-down menu under "Topics", containing the topic "Energy".

When selected, it is represented in the URL as well:

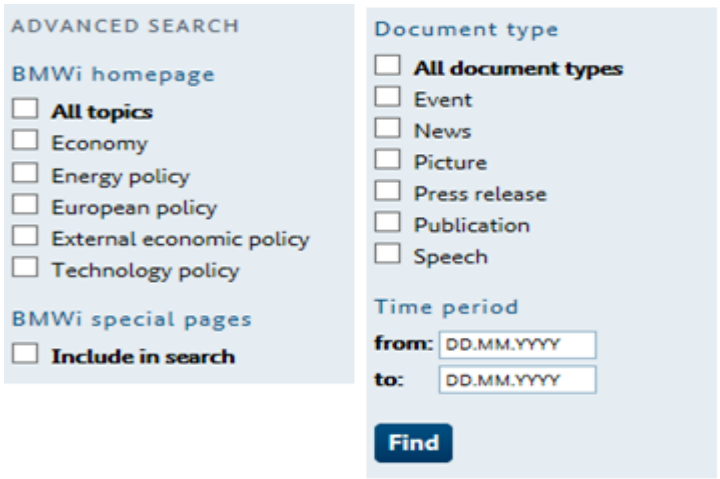
http://www.bmwi.de/EN/Topics/energy.html

- Second, there is also a search field, where we enter the query “energy”.



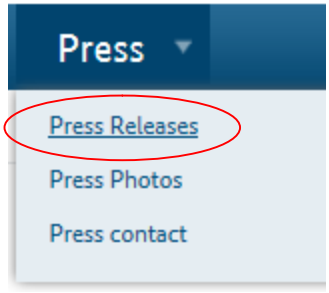
The query term is not represented in the URL: http://www.bmwi.de/EN/Service/search.html

The search results can be restricted according to the topic, document type (event, news, picture, press release, publication, speech), and the time period:



When used, those restrictions are not included into the URL.

3. Next, we search for the press releases on the website:
   - First, there is a pull-down menu under “Press”, containing “Press releases”. If selected, it is represented in the URL (http://www.bmwi.de/EN/Press/press-releases.html):

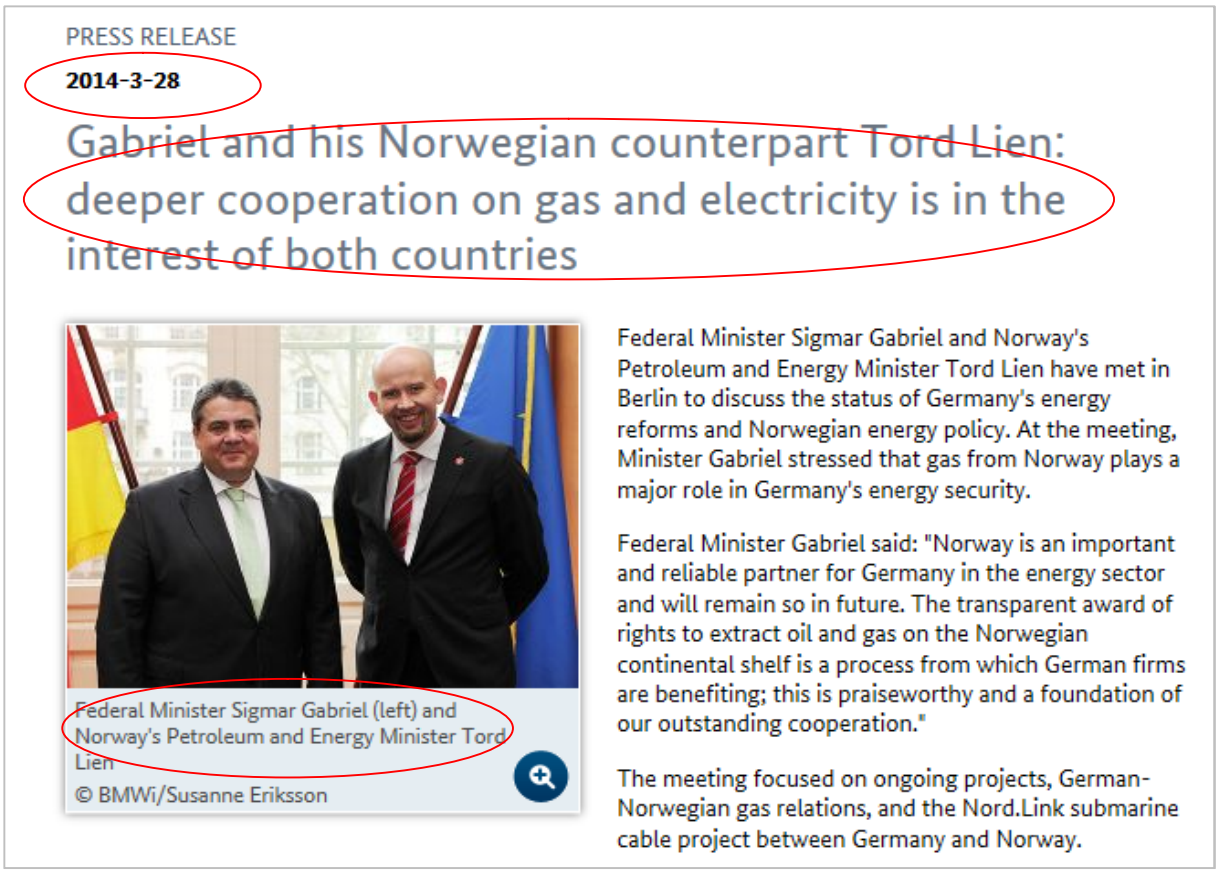


- Second, as already seen above, there is also a search field available, and the search results can be restricted according to the document type, press releases being among them.

Next, we look for the multimedia archive ('Mediathek'). It exists only in German. The contents are video and audio files, photos, slide shows, publications, social media etc.:

4. In the next step, we look at sample files from each different encoding type (articles, press releases, and multimedia) in order to find out, how the relevant information is encoded. We look at the content part, the metadata and the URL. For example:

Press releases are in HTML format, and they contain a date, title, text, and very often pictures with short descriptions:



PRESS RELEASE

2014-3-28

Gabriel and his Norwegian counterpart Tord Lien: deeper cooperation on gas and electricity is in the interest of both countries

Federal Minister Sigmar Gabriel (left) and Norway's Petroleum and Energy Minister Tord Lien

© BMWi/Susanne Eriksson

Federal Minister Sigmar Gabriel and Norway's Petroleum and Energy Minister Tord Lien have met in Berlin to discuss the status of Germany's energy reforms and Norwegian energy policy. At the meeting, Minister Gabriel stressed that gas from Norway plays a major role in Germany's energy security.

Federal Minister Gabriel said: "Norway is an important and reliable partner for Germany in the energy sector and will remain so in future. The transparent award of rights to extract oil and gas on the Norwegian continental shelf is a process from which German firms are benefiting; this is praiseworthy and a foundation of our outstanding cooperation."

The meeting focused on ongoing projects, German-Norwegian gas relations, and the Nord.Link submarine cable project between Germany and Norway.

The analysis procedure is similar as in the case of newspapers (see Chapter 5.1).

5.2.2 **Conclusions**

"Other HTML websites" are all non-newspapers and non-magazines from our lists of sources. They include web presentations of governmental and non-governmental national and international organisations and authorities, as well as European legislation bodies, but also smaller websites designed by private persons. As such, they provide a high diversity in HTML design and organisation. Consequently, it is difficult to find any real commonalities.

Content

Regarding the "content part" of those websites, we looked for: where is the important information (is it in only one part, page, topic, or thread of the website, or is the whole website topic-relevant for us), can it be reached by applying keyword queries or topic search, how is the content formally encoded, which techniques could we apply to reach it? In general, it can be said that such websites: Either can be crawled in full, beginning from the top-level URL and following all links, if the whole website is devoted to a specific topic; or only one specific page should be defined as the starting URL for the crawler and then no links, or only very specific ones should be followed. This can be a topic page, or a search query, if they are included into the URL.

Regarding the formal structure of the sites, there cannot be recognised stable patterns to be applied to all such sources already during the HTML parsing in order to recognise already structured facts (like title, author, date etc. as it was the case in newspapers). The parser can only recognise and remove boilerplates (advertisement, banner etc.) and let the content distillation up to following content analysis and extraction steps.

It also was difficult to apply a validation methodology (to keep a hold-out sample set in order to verify the analysis findings on it) to them. Each site is unique and it should be treated as such.

Metadata

The most frequently encountered meta information are keywords and description (a short summary of the content). Very often are missing: date, place, and author. However, in view of the fact that the big majority of such websites are more static than dynamic ones, with contents created once and not changing on a regular basis, those information can be considered less relevant. We have observed that the metadata annotation in websites other than news and press usually do not follow schemes such as OG or Dublin Core.

URL

The information to be found in the URL is highly individual as well. If websites exist in more than one language, usually the language code is included into the URL. If looking only for parts of the website containing a certain topic, the URL can be a good indicator for it, if the topic is included. For example:

http://www.minetur.gob.es/energia/es-ES/Paginas/index.aspx

http://www.minetur.gob.es/energia/en-us/Paginas/Index.aspx

5.3 Interactive portals / databases

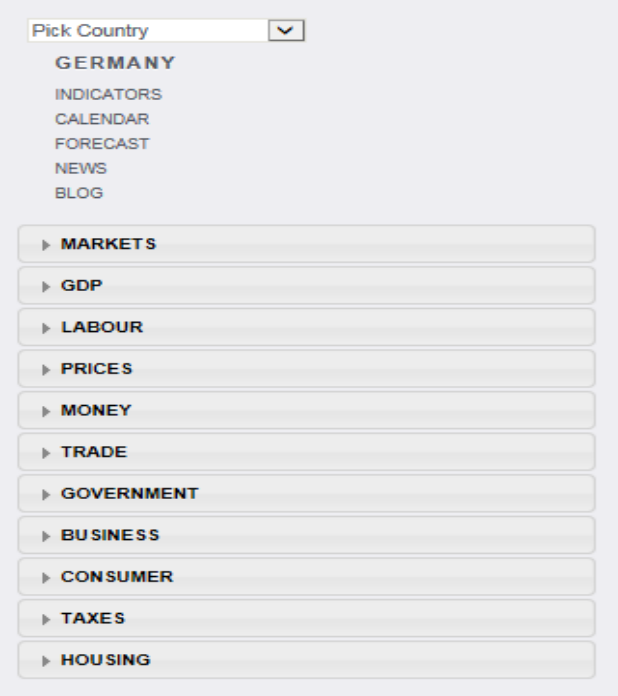
5.3.1 Sample analysis

As an example for portal analysis we take the website of Trading Economics. The requirement is, for PUC2, to retrieve and analyse all relevant economic indicators for Germany (Markets, GDP, Labour, Prices, Money, Trade, Government, Business, Consumer, Taxes & Housing).

1. First, we check in which languages the website exists, and how the language codes are encoded. We find out that the language codes can be found in the URL, for German, Spanish and French. Only for English there is no language code in the URL:

   http://www.tradingeconomics.com/germany
   http://de.tradingeconomics.com/germany
   http://fr.tradingeconomics.com/germany
   http://es.tradingeconomics.com/germany

   There are also other languages available, but not relevant for us.
2. Next, we check where on the site the required information can be found. In a navigation table on the left, we find all required issues:



Each of them contains further pull-down menu items, for example:

Markets:

- Stock Market
- Currency
- Government Bond 10Y

GDP:

- GDP per capita
- GDP per capita PPP
- GDP Constant Prices
- Gross Fixed Capital Formation
- Gross National Product
- GDP Growth Rate
- GDP Annual Growth Rate
- GDP

Labour:

- Population
- Employed Persons
- Job Vacancies
- Labour Costs
- Long Term Unemployment Rate

Productivity

Retirement Age Men

Retirement Age Women

etc.

3. Next, we select some of them and check, whether our search selection is represented in the URL. If yes, this more specific URL could be used for information retrieval later on:

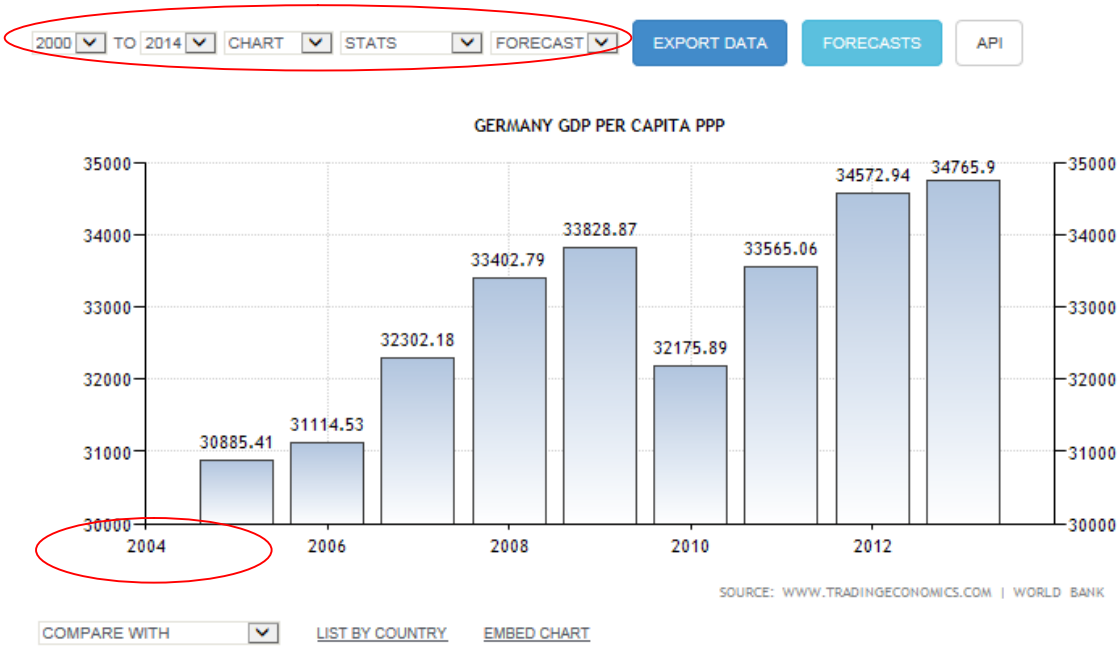
http://www.tradingeconomics.com/germany/currency
http://www.tradingeconomics.com/germany/population
http://www.tradingeconomics.com/germany/gdp-growth-annual

4. In the next step, we look at the functionality of the site, where is the data encoded and how are the results presented.

Content

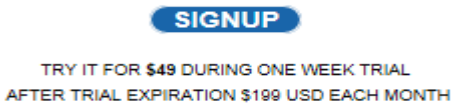
The site is an interactive portal, implemented as a JavaScript. The user can select:

- a time period (from – to)
- the presentation form (chart, line, column, area, candle, bar)
- the distribution of data (stats, mean, maximum, minimum, moving average, histogram
- forecasting model (arima, forecast, trend)
- comparison with another country



The results can be exported (as XML, CSV or JSON) or retrieved by an API.

This is possible only against a monthly payment (!):



Metadata

From the metadata we could extract only a description (which is not really valuable, since it is always the same, independent on our query), and keywords.

```
<head id="ctl00_Head1"><meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="viewport" content="width=device-width" />
<title>
</title><meta id="metaDesc" name="description" content="The Gross Domestic Product per capita in Germany was last recorded at 34765.90 US dollars in 2012, when adjusted by purchasing power parity (PPP). The GDP per Capita, in Germany, when adjusted by Purchasing Power Parity is equivalent to 158 percent of the worlds average. GDP per capita PPP in Germany is reported by the World Bank. From 1980 until 2012, Germany GDP per capita PPP averaged 27858.0 USD reaching an all time high of 34765.9 USD in December of 2012 and a record low of 20860.7 USD in December of 1980. The GDP per capita PPP is obtained by dividing the country's gross domestic product, adjusted by purchasing power parity, by the total population. This page provides - Germany GDP per capita PPP - actual values, historical data, forecast, chart, statistics, economic calendar and news." />
<meta id="metaKeyword" name="keywords" content="Germany GDP per capita PPP, Chart Graph, Data, Historical Data" />
<link href="../css/stylesheet-2014-03-11.css" rel="stylesheet" />
<link rel="icon" href="http://cdn.tradingeconomics.com/favicon.ico" />
```

URL

As described in the beginning of this section, from the URL we can extract: the language (in the example below: Spanish), the country name that is subject of our query (Italy) and the indicator we are interested in (wages):

```
http://es.tradingeconomics.com/italy/wages
```

5.3.2 Conclusions

Websites which require user interaction in order to retrieve needed information (fill out form fields, select parameters, define the presentation form for the results etc.) are difficult to be automatically crawled. For this, the best way to obtain the data is using a public API, if available, or an export function. In our test set, we had two examples of such pages: EUROSTAT (nr. 55) and Trading Economics (nr. 51). Initially, we considered analysing and using the PEPPOL portal as well (nr. 67), but since it is just now undergoing a process of transfer into a new "Open Peppol" it was difficult to work on at the moment.

Content

There are no commonalities in such interactive pages which can support an automatic information recognitions and extraction. Each of them has to be analysed in detail and separately, in order to decide which information is needed, and to then decide whether and how it can be retrieved.

Since the APIs and the export functionalities return already structured facts, there is no further text analysis needed. The facts just have to be stored in a predefined form, and used for populating the semantic repositories.

Metadata

Metadata here are not of a great relevance, because they are often only the standard ones, independent on our queries.

URL

The URLs can contain relevant information such as language, country the query is aiming at, and the indicators we ask for. But, this is highly dependent on the site.

5.4 Multimedia (audio, video)

5.4.1 Sample analysis

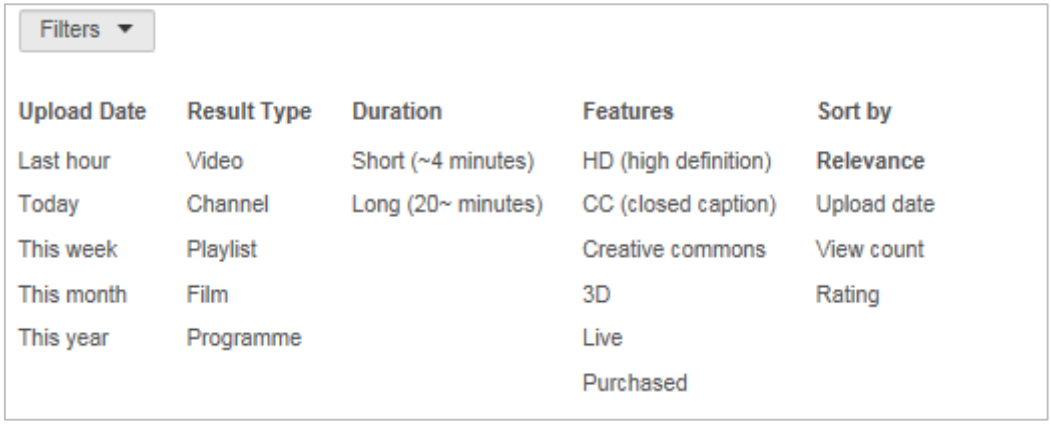
Some multimedia sources have been discussed above, as the results of search queries in newsletters or in other portals.

Here we take as example a video file from YouTube. The requirement is, for the PUC2 to retrieve and analyse all videos with the keyword “Kitchenaid”.

1. First, we go to YouTube, ask the query “Kitchenaid” and check whether it is included into the URL:

http://www.youtube.com/results?search_query=kitchenaid

2. Next, we check the possibilities to restrict the query. The filters available are:



If applied, the query restrictions are included into the URL:

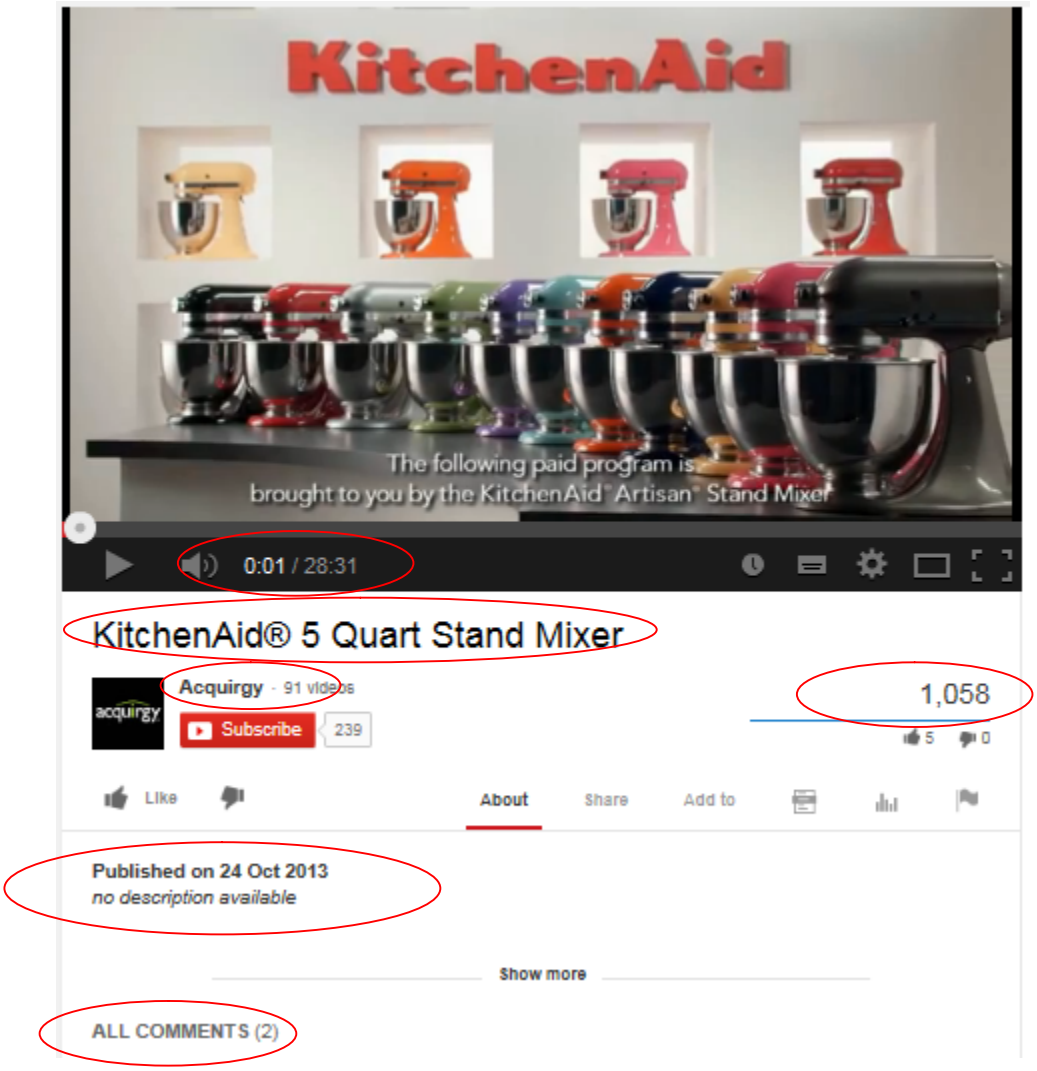
http://www.youtube.com/results?filters=week&search_query=kitchenaid&lclk=week
http://www.youtube.com/results?search_query=kitchenaid&filters=video%2C+today&lclk=today
http://www.youtube.com/results?search_query=kitchenaid&filters=long&lclk=long

3. In the last step, we go to the single matches and look for meaningful information in their content, metadata and URL.

Content

The important content is in videos themselves, as well as in their descriptions.

On the website we can identify: the video itself, its duration, title, the name of the author, number of visits, publishing date, a description (or “no description available”), number of the comments, and the comments themselves:



Metadata: In the metadata are only very few information that we are looking for available: title, URL (marked in green):

```
<title>kitchenaid - YouTube</title>
<ink rel="search" type="application/opensearchdescription+xml" href="http://www.youtube.com/opensearch?locale=en_GB"
title="YouTube Video Search">
<ink rel="shortcut icon" href="http://s.ytimg.com/yts/img/favicon-vfldLzJxy.ico" type="image/x-icon">
<link rel="icon" href="//s.ytimg.com/yts/img/favicon_32-vflWoMFGx.png" sizes="32x32">
<ink rel="alternate" media="handheld"
href="http://m.youtube.com/results?search_query=kitchenaid&amp;lclk=long&amp;filters=long">
<ink rel="alternate" media="only screen and (max-width: 640px)"
href="http://m.youtube.com/results?search_query=kitchenaid&amp;lclk=long&amp;filters=long">
<meta name="description" content="Share your videos with friends, family and the world">
<meta name="keywords" content="video, sharing, camera phone, video phone, free, upload">
```

Although language and keywords metadata exist as well (marked in red), they are not relevant for us, since they refer to the language of the internet browser you are using (not of

the video), and they only list general YouTube keywords (not keywords referring to the video content).

URL: As described in the beginning of this section, our search query and the filters are represented in the URL, but there is no other meaningful information to be extracted there:

http://www.youtube.com/results?filters=week&search_query=kitchenaid&lclk=week http://www.youtube.com/watch?v=4jFqwcLfgtE

There is a problem with downloading videos from YouTube, since it would be against the Terms and Conditions of YouTube. The API it provides does not allow one to download videos (http://apiblog.youtube.com/2010/01/youtubes-apis-and-refresher-on-our.html).

5.4.2 Conclusions

Multimedia contents (audio, video, images) occur in almost each online newspaper and magazine source, embedded into the articles or retrievable from a special multimedia area. Also many other HTML websites offer multimedia material as a part of their content or as links to external sources. Furthermore, our use cases require retrieval in public photo and video portals such as YouTube. An important question here is the free accessibility of them, and restrictions defined by Terms and Conditions of content providers in respect.

Content

We have observed that the important information can be identified not only in the media files themselves, but also in the accompanying descriptions and meta information in the "content" part of the websites. Usually they are composed of some meta-like tags such as author, date and time, title and description (cf. sample analyses in Chapters 5.1.1 and 5.4.1). They could be reached either already during the HTML parsing, or later on, within the textual information extraction task.

The analysis of the media contents themselves needs techniques of image processing, video processing and automatic speech recognition. The text recognised from speech undergoes then the same information extraction steps as news articles and other texts. However, the analysis has to cope with all limits of speech recognition. Despite the latest developments in ASR systems, there can still be misspellings and linguistic errors in the output. Furthermore, there exist no clear sentence boundaries, but rather utterance boundaries, which can be difficult to parse linguistically.

Metadata

Depending of the source of the multimedia files, they can contain different kind of metadata. If they are part of a newspaper, they mostly follow the same metadata syntax as the text articles there. Important information is e.g. the kind of medium, if available:

<meta property="og:type" content="media" />

<meta property="og:type" content="video"

<meta name = "medium" content = "video" />

For the further analysis, in particular for ASR, the language information would also be very valuable. Unfortunately, we have observed that the language meta tags, even if present,

often refer to the language of the browser or of the text article in which the media are embedded, instead of to the language spoken in the audio/video (and they can be different).

URL

The information included in the URL often follows the same patterns as the articles in which the multimedia files are embedded.

5.5 Blogs

5.5.1 Sample analysis

We analyse the blog "decor8" (http://decor8blog.com/). The requirement is, for PUC1 commercial scenario, to retrieve and analyse all articles with the keyword "Kitchenaid".

1. First, we check whether the web page has a search function. If yes, we search for "Kitchenaid":

2. Next, we check whether our search query is represented in the URL. If yes, we can consider using this more specific URL as the seed URL for the web crawler later on:

http://decor8blog.com/?s=kitchenaid&x=0&y=0

3. In the last step, we go to the single matches and look for meaningful information in their content, metadata and URL.

Content: The important content is in the text of the blogs, as well as in their descriptions.

On the website we can identify: the blog itself with text and images, its title, comments, the number of comments, and the date of the post.

Metadata: From the metadata we could extract the language, links to images, title, and the URL of the blog.

```
<meta name="p:domain_verify" content="4d0f4d41399d64269cfa3e6fcf2b1ed1"/>
<meta charset="UTF-8"/>
<meta name="copyright" content=""/>
<meta property="fb:admins" content="decor8"/>
<meta property="og:url" content="http://decor8blog.com/2013/04/16/crushing-on-smeg-refrigerators/"/>
<meta property="og:title" content="Crushing On Smeg Refrigerators"/>
<meta property="og:site_name" content="decor8"/>
<meta property="og:description" content="Okay so I&amp;#039;m thinking to buy a pure white SMEG refrigerator. I even started a smeg pinboard because it helps me to see all of my choices in one place (do you do"/>
<meta property="og:type" content="article"/>
<meta property="og:image" content="http://farm9.staticflickr.com/8257/8655845446_3c2270eefb_o.jpg"/>
<meta property="og:image" content="http://farm9.staticflickr.com/8120/8655831250_0b9fe9e8df_o.jpg"/>
<meta property="og:image" content="http://farm9.staticflickr.com/8259/8654726913_2e6e4dba3d_o.jpg"/>
<meta property="og:image" content="http://farm9.staticflickr.com/8249/8654727749_1949df4e92_o.jpg"/>
<meta property="og:image" content="http://farm9.staticflickr.com/8249/8655831436_bb728569bd_o.jpg"/>
<meta property="og:image" content="http://decor8blog.com/wp-content/uploads/2008/07/160x160-decor8blog-grey-1.jpg"/>
<meta property="og:locale" content="en_us"/>
<div class="meta clear">
```

URL: As described in the beginning of this section, our search query is represented in the URL.

http://decor8blog.com/?s=kitchenaid&x=0&y=0

From the single posts' URLs, we can extract the date of the post and its title:

http://decor8blog.com/2013/04/16/crushing-on-smeg-refrigerators/

5.5.2 Conclusions

When crawling blogs we are interested in blog posts, as well as different metadata such as language, author, visits etc.

Content

In the content part of the blogs, many pieces of very important meta information such as date and time of the post (frequency of updates), author name, language, visits etc. can be found. This kind of statistical information is interesting for the analysis of the blogger's influence and trends in social networks.

The texts and media files from the posts (i.e. their content) will undergo further multimedia and text analysis and information extraction.

Metadata

The meta data of blogs can range from no useful information at all to many relevant data, as in the following example:

```
<meta name="DCSext.rChannel" content="Blogs" />
<meta name="DCSext.ContentHeadline" content="Financing more solar energy - MuniLand" />
<meta name="DCSext.rAuthor" content="Cate Long" />
<meta name="description" content="There is another rapidly growing area in site-based residential or commercial solar installations." />
<meta name="keywords" content="solar" />
```

URL

Very often, if blogs are included into e.g. newspapers, already in the URL is indicated that the channel type is a "blog", and some other information such as language code or topic can also be included:

http://uk.reuters.com/search/blog?blob=energy

5.6 Social media

5.6.1 Sample analysis

As the example for a social media analysis we take Twitter. The requirement is, for PUC2, to retrieve and analyse all relevant tweets for the hashtag #yoghurt.

1. First, we search in Twitter Search for #yoghurt, and check, whether our search query is included into the URL:

https://twitter.com/search?q=%23yogurt

2. Next, we analyse the matches and check the content area of the site, the metadata and the URL.

Content: In theposts area we recognise: The Twitter name of the author, the post itself, images (if any), number of retweets, favourites, and the time of the posting:

The retweets can be reached by clicking onto “RETWEETS”. They appear in a pop-up window. There, we see the re-posts, the names of their authors and the number of followers of each of them:

Metadata: In the meta tags we recognise only a short description as the information potentially relevant for us:

```
<meta charset="utf-8">
<meta name="description" content="Die neuesten und besten Tweets auf #yoghurt. Lies, was Leute sagen und nimm am Gespräch teil.">
<meta name="msapplication-TileImage" content="//abs.twimg.com/favicons/win8-tile-144.png"/>
<meta name="msapplication-TileColor" content="#00aced"/>
<meta name="swift-page-name" id="swift-page-name" content="search">
```

URL: as described above, from the URL only the information on the hashtag name could be extracted:

```
https://twitter.com/search?q=%23yogurt
```

5.6.2 Conclusions

When collecting data from social media networks, we are interested in the posts, as well as in different metadata such as language, author, location, number of posts and answers, followers etc. Which data and metadata can be found is determined by the respective source.

Content

The content of social media follows other rules than in standard text articles. The language is interspersed by abbreviations, idioms, grammatically and orthographically incorrect words, and emoticons. The text analysis and information extraction have to cope with those problems.

Metadata

Most social media channels have public APIs. When they are crawled, the available meta information is already offered in a structured form. Depending on the channel, possibleinformation is language code, location, names, number of posts and re-posts, followers etc.

URL

Having retrieved most important data through the API, URL parsing might be not necessary to obtain further information.

6 APPLICABLE INFORMATION RETRIEVAL AND EXTRACTION TECHNIQUES

In this chapter we will very briefly present some of approaches that can be applied to obtain the required information. We mention both information retrieval and cleaning techniques (e.g. crawling, parsing, and format conversion) on the one hand, and information extraction techniques (analysis and extraction of implicit information, harvesting of explicit meta information) on the other. Finally, we also consider the enrichment of data using external metadata, which is closely connected to content extraction and involves retrieving additional content on a specific topic, apart from the ones found inside the initial source, which can provide a broader view of it. We describe them in an order in which they might be applied to retrieve, clean and analyse the data.

6.1 Web crawling

Crawlers are fully automated tools that scan the websites, follow all links from a site, retrieve the contents and store the results in a predefined repository.

Focused vs. generic crawling

We distinguish between a generic and a focused (targeted) crawling approach.

A generic crawler accepts as input a seed list (i.e. a number of web domains or URLs) and discovers up to a number of levels (1,2,3..) the links and fetches the content of those links. The advantage of a generic crawler is that it does not require frequent updates and modifications due to changes in the structure or services of a given source and thus it is easier to support. The disadvantage could be that a generic crawler collects huge amounts of data that are first stored, and the selection of relevant subsets of documents, e.g. based on keywords, has to be performed afterwards.

A targeted crawling is when the crawler tries to guess from the URL of a page if its content is relevant to some keywords **before** even fetching that content. The advantages of a focused crawler are that you spend less time, money & effort processing web pages that are unlikely to be of value. Disadvantage of a focused crawler could be that it can miss relevant pages if there does not exist a chain of hyperlinks that connects the starting pages to other relevant ones.

Crawling social media

When crawling social media sources, among other things, the following should be taken into account:

- The crawler should be able to recognise the structure of the sources and to differentiate the so called ‘entities’, i.e. the postings on the one hand, and their attributes, such as author, date of the posting, registration date of the user/author, and other meta information, on the other. For this, the methodology of scraping is the most recommendable one: the crawler tries to identify these structures during the crawling and to store them in predefined structured repositories. The other way around (crawl, store, and then structure) would require higher computing capabilities and lead to higher error rates.

For MULTISENSOR, since we will use the infrastructure that already exists and is being used at *pressrelations*, the approach will not be to crawl the contents of the posts and tweets, but to acquire them through public APIs. These APIs return already structured information.

- Not all sites can be continuously and completely crawled and indexed. The providers of the portals might set limits for the permissible loads
- Consequently, it might become necessary to implement some “observer” techniques, in order to identify and retrieve only new or changed data, and thus to treat available resources with care. Another possible approach is to apply a non-selective crawling and ignore content freshness. What will be kept in the end is determined at post-crawling phase, by a service that inspects the repository and filters out duplicates or old content.

Crawling multimedia data

To fetch and store multimedia files (video, audio) during the crawling can be very time-consuming and also difficult. An applicable way is to try to extract the links to images, audio and video files, and to fetch them at post-crawling phase. However, given that in some cases the multimedia content is accessed through the use of jQuery, Flash, or a media player, extracting these hidden links is not always possible. In the cases that the extraction of the links was successful, the content can be fetched afterwards and stored in a multimedia repository.

6.2 Data acquisition via API

Additionally or alternatively, the data can be acquired by consulting the public APIs which are offered by most social media channels. We list only a couple of examples:

FB Graph API

Facebook is an online social networking service. Users must register before using the site, after which they may create a personal profile, add other users as friends, exchange messages, and receive automatic notifications when they update their profile. Additionally, users may join common-interest user groups, organised by workplace, school or college, or other characteristics, and categorise their friends into lists such as "people from work" or "close friends".[9]

Facebook Graph API[10] is the new version of the old REST-API. It has been published in April 2010. The idea of this interface is that the user can query for each single object type (user, event, site) in the Facebook-Graph by using a uniform URL (https://graph.facebook.com/ID). By using further URLs it is possible to fetch different kinds of relations. The Graph API also provides a search function for all public contents.

Twitter Search API

Twitter is an online social networking and microblogging service that enables users to send and read short 140-character text messages, called "tweets". Registered users can read and

[9] http://en.wikipedia.org/wiki/Facebook

[10] https://developers.facebook.com/docs/graph-api

post tweets, but unregistered users can only read them. Users access Twitter through the website interface, SMS, or mobile device app.[11]

The Twitter Search API[12] is part of Twitter's v1.1 REST API. It allows queries against the indices of recent or popular Tweets.It has been published in 2006. It can be used to query a real-time index of tweets. This index however is limited to the last six to nine days. Older tweets cannot be queried.

Alternatively, **Twitter Streaming API**[13] can be used for real-time monitoring of tweet streams. A part of the Streaming API is the "Fire-Hose" method, which enables you to monitor all public status messages. It can be used only against payment.

Google+

Google+ is a social networking and identity service that is owned and operated by Google Inc. Google has described Google+ as a "social layer" that enhances many of its online properties, and that it is not simply a social networking website, but also an authorship tool that associates web-content directly with its owner/author. It is the second-largest social networking site in the world after Facebook.[14]

The Google+ API[15] has been released in September 2011. It still has very strong request limits. In October 2011, among many other features, a search function for public content has been integrated.

LinkedIn

LinkedIn is a business-oriented social networking service. It is mainly used for professional networking. Founded in 2002, in 2006 LinkedIn increased to 20 million viewers.As of June 2013, LinkedIn reports more than 259 million acquired users in more than 200 countries and territories.[16]

LinkedIn offers different Modules both for REST[17] and JavaScript[18] to extract specific data. You search for a profile, for a topic or a location.

IN.API Overview

IN.API.Profile()

IN.API.Connections()

IN.API.PeopleSearch()

[11] http://en.wikipedia.org/wiki/Twitter

[12] https://dev.twitter.com/docs/using-search

[13] https://dev.twitter.com/docs/api/streaming

[14] http://en.wikipedia.org/wiki/Google%2B

[15] https://developers.google.com/+/api/?hl=de

[16] http://en.wikipedia.org/wiki/Linkedin

[17] https://developer.linkedin.com/rest

[18] http://developer.linkedin.com/javascript

IN.API.MemberUpdates() and IN.API.NetworkUpdates()

IN.API.Raw()

The Raw method has options like GET, POST, PUT and DELETE[19].

To access the API you need a key, which you can use for both JavaScript and REST. You also need to be logged in[20].

Pinterest

Pinterest is a visual discovery tool that people use to collect ideas for their different projects and interests. People create and share collections (called "boards") of visual bookmarks (called "Pins") that they use to do things like plan trips and projects, organise events or save articles and recipes. Pinterest users can upload, save, sort and manage images, known as pins, and other media content (e.g. videos and gifs) through collections known as pinboards.[21]

You have the options to get the most repinned contents, the most clicked-through or results for a certain category. The API works with JSON. The responses are JSONP or jsonp.[22]

Domain API overview Request Access:

fields=pin.first_name,pin.last_name

add_fields=pin.first_name,pin.last_name,board.title

GET /v3/domains/<domain>/pins/top/repins/

Delicious

Delicious (formerly del.icio.us) is a social bookmarking web service for storing, sharing, and discovering web bookmarks. By the end of 2008, the service claimed more than 5.3 million users and 180 million unique bookmarked URLs.[23]

The Delicious API[24] works with RSS and JSON feeds. You can access to a list of bookmarks to a specific topic with the API (user, tag, combinations).

An example for a get-call is:

curl https://user:passwd@api.delicious.com/v1/posts/get?tag=webdev&meta=yes

StumbleUpon

StumbleUpon is a discovery engine (a form of web search engine) that finds and recommends web content to its users. Its features allow users to discover and rate Web

[19] https://developer.linkedin.com/documents/inapiraw

[20] https://developer.linkedin.com/documents/getting-started-javascript-api

[21] http://en.wikipedia.org/wiki/Pinterest

[22] https://developers.pinterest.com/api_docs/

[23] http://en.wikipedia.org/wiki/Delicious_(website)

[24] https://github.com/vjkaruna/delicious-api

pages, photos, and videos that are personalised to their tastes and interests using peer-sourcing and social-networking principles.[25]

The Su.pr API from StumbleUpon[26] works with HTTP-request and REST. You need authentication to use it. The current version is 1.0 but you can choose the version you want to use through a parameter (default 0.95).

The output format is either JSON (default) or XML. You need to have an account to use the API. Example GET-request:

http://su.pr/api/shorten?longUrl=http://www.stumbleupon.com

Reddit

Reddit is an entertainment, social networking service and news website where registered community members can submit content, such as text posts or direct links. Only registered users can then vote submissions "up" or "down" to organise the posts and determine their position on the site's pages. Content entries are organised by areas of interest called "subreddits".[27]

In the Reddit-API[28] there are several GET-methods to acquire for example a list of posts relevant to a certain topic.

GET /api/subreddits_by_topic.json

With the search function you can search in a link list for interesting topics like: relevance, new, hot, top, and comments. There's also a testing environment for the API: https://apigee.com/console/reddit

Digg

Digg is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Internet audience such as science, trending political issues, and viral Internet issues. It was launched in its current form on July 31, 2012, with support for sharing content to other social platforms such as Twitter and Facebook.[29]

There's no API of any sort available for Digg. There is a Google developing group for Digg-APIs, but it doesn't seem to be professional[30].

6.3 Metadata harvesting

Metadata harvesting (collection, extraction) aims at automatically collecting internal metadata from META tags, which are located in the “header” source code of a website, as

[25] http://en.wikipedia.org/wiki/StumbleUpon

[26] http://www.stumbleupon.com/help/business-tools/supr/supr-api/

[27] http://en.wikipedia.org/wiki/Reddit

[28] http://www.reddit.com/dev/api

[29] http://en.wikipedia.org/wiki/Digg

[30] https://groups.google.com/forum/#!forum/diggapidev

shown in our metadata analysis examples above. Metadata can be simple HTML meta tags or defined by a schema such as Dublin Core Schema[31].

Several methods can be used for automatic metadata harvesting, which can be divided into two main categories: machine learning approaches and rule-based approaches.

In general, machine learning (ML) methods are robust and adaptable to any document set, when one can find a gold set of training data. In fact, the basic idea of this approach is to learn segmentation models from training data and then to use them from classifying previously unseen data. ML techniques may include symbolic learning, support vector machines (Han, et al. 2003), hidden Markov models (Seymore, et al. 1999), as well as statistical methods. Hidden Markov Models (HMMs) are the most widely used learning approach for extracting information from sequential data. However, they are based on the assumption that the model features they represent are not independent from each other. For this reason, HMMs cannot easily exploit regularities of semi-structured real data. In order to deal with the problem of independent features, other models have been introduced such as maximum entropy-based Markov models (McCallum, et al. 2000) and conditional random fields (Lafferty, et al. 2001). In general, all learning techniques represent training data as a set of features which, in case of metadata harvesting, can be word and line-specific. The former may include intrinsic orthographic properties of words as well as information about words and n-grams extracted from reference corpora and gazetteers. Line-specific features, instead, include for example the number of words in the line, the line position in the document as well as information about the contained tokens. Learning techniques have been employed with promising results, although they proved to work properly with relatively homogeneous document sets. Effectiveness can significantly decline as the heterogeneity of the data collection increases, and an annotated training set is always mandatory.

On the contrary, rule-based approaches can be straightforwardly implemented without training, since they rely on a set of rules that define how to extract data based on human observation (see for example (Chowdhury, 1999) and (Ding, et al. 1999)). Template mining techniques based on pattern recognition and pattern / regular expression matching in meta elements are commonly used in rule-based applications for metadata harvesting (Moumtzidou,et al. 2010).

6.4 Parsing, pre-filtering, cleaning

In order to retain only the core content part of a crawled website, as well as to recognise the boundaries between the important blocks of it (title, text body), tools for boilerplate removal and text extraction from HTML pages are needed. They commonly include different HTML-parsing, filtering and extraction pipelines.

HTML parsers can be used to identify different parts of a document and their attributes, as well as to find relevant information there. An HTML document is mainly composed of three parts[32]:

[31] http://dublincore.org/documents/2003/06/02/dces/

[32] http://www.w3.org/TR/html401/struct/global.html

- a line containing HTML version information,
- a declarative header section (delimited by the HEAD element),
- a body, which contains the document's actual content. The body may be implemented by the BODY element or the FRAMESET element.

The HEAD element contains information about the current document, such as its title, keywords that may be useful to search engines, and other data that is not considered document content.

The BODY of a document contains the document's content. The content may be presented as text, images, graphics, audio, etc.

The crawled data should be cleaned from non-relevant data, such as advertisement, navigational text, related articles and other so-called “boilerplates” (Kohlschütter, et al. 2010). For this, open source software is available, e.g. boilerpipe[33]

Another important step at this stage might be the identification and removal of duplicate entries.

Furthermore, it can be worth looking at the possibility of identifying and eliminating less relevant sources already at this stage of the data processing, before the content analysis and information extraction start.

6.5 Language identification

For cases where the language of a (text) article cannot be extracted from the metadata or its URL, there is a need for an external language identification component which analyses the content of the article and proposes its language. Language is one of the obligatory input parameters for many information extraction tasks.

Language identification of the audio/video input is a very hard task, and there exist only very few programmes that attempt to do it such as Softpedia[34] or Nexidia[35].

6.6 Automatic Speech Recognition

Audio files, in order to analyse their content and to extract important information from them, should be converted into text by using speech recognition software.

Speech recognition is translation of spoken words into text. State-of-the-art speech recognition technology can include, besides the conversion of audio into text, also transcription of films and videos including automatic creation of subtitles, and media search techniques.

In MULTISENSOR, the Automatic Speech Recogntion (ASR) system developed by Linguatec will be used. It supports nearly all audio/video input formats. Based on RWTH-ASR technology(Rybach, et al. 2009), it is a speaker-independent, server-based, LVCSR (large vocabulary continuous speech recognition(Ney, et al. 1998) technology. This ASR framework

[33] http://code.google.com/p/boilerpipe/

[34] http://www.softpedia.com/get/Multimedia/Audio/Other-AUDIO-Tools/Language-Identification.shtml

[35] http://www.nexidia.com/government/products/nexidia_language_id_workbench

employs a series of state-of-the-art techniques: continuous density HMMs (Hidden Markov Models) for the acoustic modelling; MFCC (Mel Frequency Cepstral Coefficients) feature extraction; LDA (Linear Discriminant Analysis) and VTLN (Vocal Tract Length Normalisation) for the recognition; support of language models in ARPA-format and pronunciation variants; speaker adaptation by the means of CMLLR (Constrained Maximum-Likelihood Linear Regression); time-synchronous left-to-right beam search strategy for the decoding.

It must be noted that the processing of text converted from audio/video involves also dealing with potentially erroneous input, such as wrong recognition due to pure audio quality and background noise, absence of sentence boundaries in the output, and wrong capitalisation.

Since the ASR output contains timestamps for each recognised word, it can enable a later indexing and linking of recognised text to the audio signal. This function can support the user to "jump" into the audio (or video, if the input for ASR was a video file) directly to the place where a certain word was recognised.

6.7 Video and image processing

Image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Image processing usually refers to digital image processing, but optical and analogue image processing also are possible[36].

For video files, in addition to the image analysis, the analysis of the audio signal and translation of spoken words into text can be done by using ASR.

Both for image and video files, concepts can be detected from a given set of high level visual concepts such as water, man, worker, landscape. Additionally, for videos, event information can be extracted as well. Event detection takes into account the concept information from several sequential frames and also can take into account motion information as well thus introducing the time dimension to the whole processing procedure.

Many images, scanned text files and PDF files can contain text. For the text extraction from such non-machine-processable files, in order to obtain the text from them for further content analysis, there are two main techniques which can be applied.

- OCR (optical character recognition)
- PDF to TXT conversion

For this, several open source tools exist, such as FreeOCR[37], or PDf2txt[38]. There are also commercial tools available such as: Omnipage (Nuance)[39], FineReader (Abbyy)[40] for OCR, and Argus (Iceni)[41], Magellan (BCL Technologies)[42] for PDF conversion.

[36] http://en.wikipedia.org/wiki/Image_processing

[37] http://www.free-ocr.com/

[38] http://www.pdf2txt.de/

However, it must be noted that the processing (content analysis) of text converted from images and PDFs involves also dealing with potentially erroneous input, such as wrong recognition of characters, line breaks in the middle of sentences, wrong recognition of the text flow, etc.

6.8 Content analysis and information extraction

Information extraction is an area of NLP which deals with finding factual information in free text. In case of MULTISENSOR, it also includes analysis of multimedia sources, images, video, and audio content.

Factual information to be extracted are objects (entities) that include named entities, numerical and textual facts, occurrences, and states, with their arguments and attributes, and with semantic relations between them. The goal of IE is to extract them from a given pool of domain-specific documents, according to a set of predefined types of entities and relations, to then build a meaningful representation of their semantic content and to populate databases and ontologies. This way, they build a basis for creating even more complex structures such as summaries, trend forecasts, and decision suggestions.

In this chapter we aim at providing a brief overview of information extraction tasks. The classic information extraction tasks include named entities recognition (NER), concept extraction, co-reference resolution, entity linking and relation extraction, as well as event extraction.

The task of **NER** is to identify and classify named entities such as person names, organisations, place names, measurement and currency expressions, temporal expressions etc. It can additionally include extraction of supplementary information such as e.g. in the case of persons titles and gender, in the case of place names the allocation of the kind of entity such as country, city, river, mountain etc. NER involves a series of different linguistic processing tasks, starting from the tokenisation, normalisation (different spelling forms of the same name), and lemmatisation (in order to recognise inflected forms of the names), pattern matching and a shallow or deep syntactical analysis.

Concept extraction is usually based on a domain-specific pre-specified set of entities which are application relevant. There is a certain overlap between the NER task and concept extraction. The process of extracting concepts can be quite challenging, starting from the definition what is a concept to the clear distinction between NEs, keywords, events, and other concepts. Concept extraction involves, like NER, a series of different linguistic analysis steps, starting from the sentence splitting and tokenisation, to deep semantic analysis and parsing.

Co-reference resolution is, simply said, the identification of multiple mentions of the same entity in a document (or even in several documents). It is a very challenging task, having the

[39] http://www.nuance.com/for-individuals/by-product/omnipage/index.htm

[40] http://finereader.abbyy.com/

[41] http://www.iceni.com/argus.htm

[42] http://www.bcltechnologies.com/

job of recognising different spellings, inflected forms, full names vs. abbreviations, nominal or pronominal references to an already mentioned entity.

Entity linking and Relation extraction is the task of detecting and classifying relationships between entities in a document (or even in several documents). The number of the relations between the entities can theoretically be unlimited, but usually there is a pre-specified set of relations that should be recognised within a given domain and for a fixed domain-specific dataset. Relations are very often presented as triples (subject, predicate, object), e.g.:

CapitalOf(Sofia, Bulgaria), PresidentOf(Hollande, France).

Event extraction is a kind of relation recognition and of giving answers to the W-questions (who, when, where, what, why). It identifies the main actors of an event, its location, and other important facts such as number of actors, time, causes and effects of an event. It is considered to be the hardest information extraction task(Piskorski and Yangarber, 2013).

In the beginnings of IE research the emphasis was on processing single documents and monolingual (mostly only English). The trend now is to shift to cross-document and multiple-language approaches.

Moreover, in the early years the information extraction tasks were only rule-based, knowledge-driven, and developed by linguistic experts. Now, statistical methods are used increasingly and an increasing emergence of trainable systems, both unsupervised as well as supervised machine-learning approaches, can be observed.

6.9 Enrichment by external content and metadata

There exist open-source knowledge bases with structured content which could be linked to the results of an IE task, in order to enrich them by additional facts. We name only two:

DBpedia (from "DB" for "database") is a project aiming to extract structured content from the information created as part of the Wikipedia project. This structured information is then made available on the World Wide Web. DBpedia allows users to query relationships and properties associated with Wikipedia resources, including links to other related datasets.[43]

The **Google Knowledge Graph** is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. It provides structured and detailed information about a topic in addition to a list of links to other sites. The goal is that users would be able to use this information to resolve their query without having to navigate to other sites and assemble the information themselves.[44]

In addition to the meta information that will be extracted from the primary sources themselves, some required meta information and additional content can be acquired from external web statistic providers, e.g. from Alexa (www.alexa.com). There we can look for the information on:

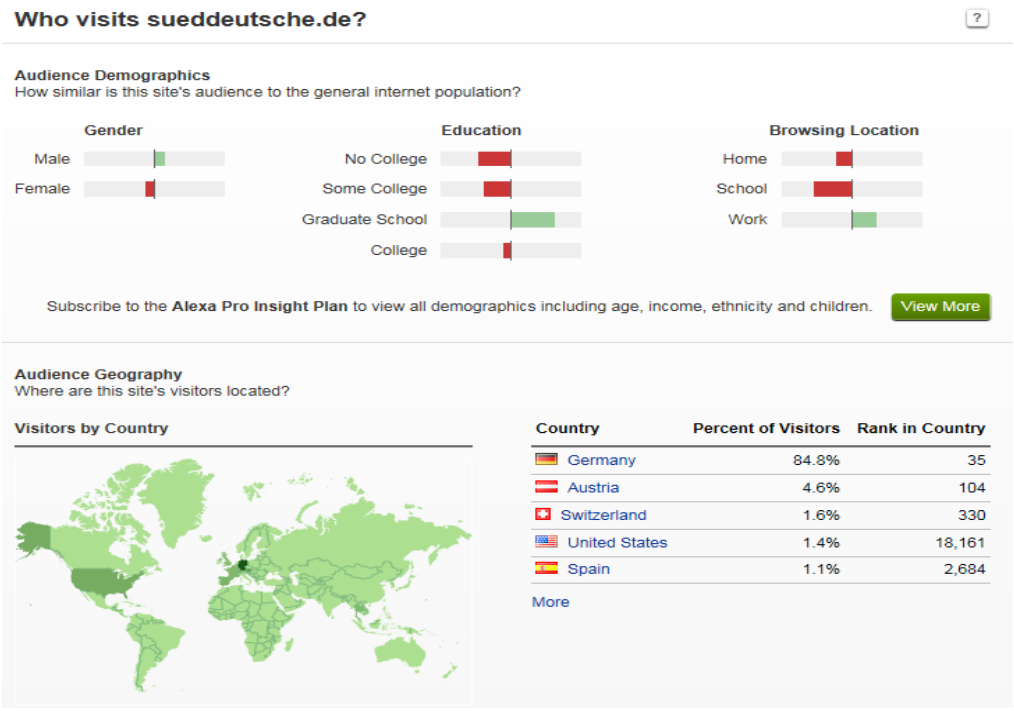
- How popular is a website?

[43] http://en.wikipedia.org/wiki/DBpedia

[44] http://en.wikipedia.org/wiki/Knowledge_Graph

- How engaged are the visitors to the website?
- Who visits it?
- Where do the visitors come from?
- Where do the visitors go next?
- What other sites link to the website?
- What sites are related to it?
- Where do the visitors go on the website?

We have collected those data for www.sueddeutsche.de as an example:



…etc.

For this service is no free API available. There is a traffic-API over the Amazon Web Services: http://aws.amazon.com/awis/ (Payment per request).

Alternative sites (payment required):

https://developer.compete.com/

https://developer.similarweb.com/traffic_api

7 FINDINGS AND CONCLUSIONS

In this deliverable we conducted an analysis of the data sources considered important for our use cases. The study was performed in two phases: (a) a survey on state of the art media monitoring approaches and data sources and (b) the empirical analysis of the MULTISENSOR data set.

Regarding the data sources and the kind of information to be extracted, the findings of the first analysis phase have shown that the most relevant information can roughly be divided into three groups:

- factual information to be extracted from the "content" part of the source, such as names, other textual and numerical concepts and relationships between them, as well as their roles, functions and attributes
- metainformation about the sources, such as language, media type, author, date, followers, likes, views etc.
- and derived meta information, i.e. our interpretation of the content, by means of identifying the main topic, the tonality and sentiment, or key messages.

In the second phase, we empirically analysed each of 76 different data sources and tried to identify, where the information listed above can be found. The main findings are:

- factual information is in the content part of the sources and can be reached, depending of the data medium, by text analysis, image and video processing and by speech recognition
- the meta information can be reached either directly from the meta tags, or by analysis of the content part, or by using public APIs in particular for the social media sources
- the implicit metadata derivation (sentiment, tonality etc.) was not subject of this deliverable.

Thus, as the main containers of the useful information we identified the metadata (date, time, location, author, keywords, descriptions etc.), the URL (language, date, topic), and finally the content part, as the main carrier of the information.

For the analysis, we divided the data sources according to their main format and function into: newspapers, other html sites, interactive portals, blogs, social media, and as a separate group - single files (mostly in the PDF format). For each of them (except of single PDF files) we used an analysis part of the corpus to identify the commonalities and draw conclusions, and a smaller hold-out corpus to verify them.

For each source we also listed some possible retrieval and extraction techniques. In general, the main techniquesare:

- web crawling, which includes several different approaches such as: adopting a generic crawling approach and selection of relevant content in the post-crawling phase; targeted crawling by using more specific URLs as the seed URLs, and crawling only a pre-defined depth of levels; crawling only the landing page without following any further links in certain cases

- HTML parsing, which can be used on the one hand to remove boilerplates from the websites, and on the other hand for recognition of relevant content parts (header, title, body etc.) in order to identify facts (date, author, title etc.) already during this processing step
- metadata extraction and URL parsing that can be used for recognition of important meta information
- format conversion step, which is needed to convert files such as PDF, Word, Excel, PowerPoint into plain text format, or to recognise text in images (OCR), or to separate the audio signal from the video
- content extraction techniques that depend on the medium and include information extraction from the text input, image and video processing and speech recognition

Regarding the richness of the data sources and use case coverage, we have observed that:

- For the use case 1 (journalistic scenario) the major news outlets contain huge amounts of data, but in order to select the most relevant ones, the search and selection criteria should be refined and thus for example it is better to use phrases such as “wind energy, renewable energies” instead of just using the keyword “energy. Other sources, such as governmental and non-governmental organisations, academiaare less rich, but they offer more domain-relevant data.
- For the use case 1 (commercial media monitoring scenario), the major news outlets contain significantly less data than the ones found for the journalistic scenario (the keywords tested were “Kitchenaid”, “Liebherr” and “AEG”). From the most popular press sources only very few results could be retrieved, and they are also very ambiguous. In the case of “Liebherr” the matches often refer to people with this surname, and not to the company in respect; in the case of “AEG” almost all matches in the popular press sources refer to an American music promoter and not to the company AEG and their products. However, the trade press delivers more promising results, many of them with very few textual content and much more images.
- For the use case 2 (SME internationalisation) all sources are much smaller, but they seem to be domain-relevant. Among them are interactive databases containing relevant statistical and economic facts. They need either a login, or can be queried by an API, but against payment.

Regarding the accessibility of the sources, most of them are freely available, but some are restricted by paywalls or require login to access data (in the Appendices A and B all of them are marked in red).

8 REFERENCES

Chowdhury, G., 1999. “Template mining for information extraction from digital documents”, Library Trends 48 (1), pp. 182–208.

Creswell, J. W., 2007. “Qualitative Inquiry and Research Design”, California, USA; London,UK; New Delhi, India: Sage Publication, Inc.

Ding, Y., Chowdhury G., and Foo, S., 1999. “Template mining for the extraction of citation from digital documents”, In Proceedings of the Second Asian Digital Library Conference, Taiwan, pp. 47–62.

Han H., Giles, C. L., Manavoglu E., Zha, H., Zhang Z. and Fox E. A., 2003. “Automatic document metadata extraction using support vector machines”, In Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital Libraries, pp. 37-48.

Kasper, H., Dausinger M., Kett, H., Renner, T., 2010. “Marktstudie Social Media Monitoring Tools”, Fraunhofer IAO, Stuttgart, Germany.

Kohlschütter, C., Fankhauser P., and Nejdl W., 2010. “Boilerplate Detection Using Shallow Text Features”, In WSDM ’10.

Lafferty, G., McCallum, A., and Pereira F., 2001. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, In Proc. 18th International Conf. on Machine Learning, pp. 282–289.

McCallum, A., Freitag, D., and Pereira F., 2000. “Maximum entropy Markov models for information extraction and segmentation”, In Proc. 17th International Conf. on Machine Learning, pp. 591–598.

Moumtzidou, A., Vrochidis, S., Tsatsou, D., Gkalelis, N., Dasiopoulou, S., Tonelli, S., Pianta, E., Tarvainen, V., Karppinen, A., Myllynen, M., Koskentalo T., 2010. “D2.1 Analysis of the codification of metadata and content in environmental service pages and design of the functional index repository”, PESCaDO, FP7-248594.

Ney, H., Welling, L., Ortmanns, S., Beulen, K., and Wessel, F., 1998. “The RWTH Large Vocabulary Continuous Speech Recognition System”, In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 853-856, Seattle, WA, USA.

Piskorski, J., Yangarber, R., 2013. “Information Extraction: Past, Present and Future”, In: Poibeau T. et al. (Eds.), Multi-source, Multilingual Information Extraction and Summarization, Springer series: Intelligent Systems Reference Library, Vol. 42, Springer, pp. 23-49.

Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., and Ney, H., 2009. “The RWTH Aachen University Open Source Speech Recognition System”, In Interspeech, pp. 2111-2114, Brighton, UK.

Seymore, K. McCallum, A., and Rosenfeld, R., 1999. “Learning hidden Markov model structure for information extraction”, In Proceedings of AAAI 99.Workshop on Machine Learning for Information Extraction, pp. 37–42.

Siavash, A.M.J., 2011. “The role of knowledge in internationalization of small- and medium-sized enterprise”, University of Umeå, Sweden.

Uhlmann, K., 2011. “Anforderungen an Monitoring Tools und die Online Reputationsanalyse in Schweizer Finanzdienstleistungsinstituten”, University of Freiburg, Switzerland.

A Appendix: Empirical study - Overview of dataset

In this section we provide an overview of all sample websites that were used for this empirical study.

Src. Nr.	Name / Abbreviation	URL	Language	Content	Use case relevance	Free/restricted availability
1	SZ	http://www.sueddeutsche.de/	German	News articles Images Video	PUC1-J PUC1-MM	free
2	FAZ	http://www.faz.net/	German	News articles Images Video/Audio	PUC1-J PUC1-MM	free
3	Die Zeit	http://www.zeit.de/index	German	News articles	PUC1-J PUC1-MM	free
4	Spiegel	http://www.spiegel.de/	German	News articles, videos, interactive maps and graphs	PUC1-J PUC1-MM	free
5	Le Figaro	http://www.lefigaro.fr/	French	News articles	PUC1-J PUC1-MM	Restricted (everything older than 2 days)
6	Le Monde	http://www.lemonde.fr/	French	News articles	PUC1-J PUC1-MM	Restricted (Entire articles only for subscribers)
7	Libération	http://www.liberation.fr/	French	News articles, videos	PUC1-J PUC1-MM	free
8	El Mundo	http://www.elmundo.es/	Spanish	News articles, videos, photos, blogs	PUC1-J PUC1-MM	free
9	El País	http://elpais.com/	Spanish	News articles	PUC1-J PUC1-MM	free
10	The Guardian	http://www.theguardian.com/uk	English	News	PUC1-J PUC1-MM	free
11	The Times	http://www.thetimes.co.uk/tto/news/	English	News articles, videos, images	PUC1-J PUC1-MM	Restricted (Entire articles only for subscribers)
12	The daily telegraph	http://www.telegraph.co.uk/	English	News articles	PUC1-J PUC1-MM	free
13	Reuters	http://www.reuters.com/ http://uk.reuters.com/ http://de.reuters.com/ http://fr.reuters.com/	English, German French Spanish	News, agency reports, images, videos, blogs, Stock exchange quotations	PUC1-J PUC1-MM	free

		http://es.reuters.com/				
14	AFP	http://www.afp.com/ http://www.afp.com/en/ http://www.afp.com/es/ http://www.afp.com/de/	French English Spanish German	News, agency reports, social media posts (Twitter, Facebook)	PUC1-J PUC1-MM	free
15	dpa	http://dpa.de/ http://www.dpa.de/English.82.0.html http://www.dpa.de/Espanol.83.0.html	German English Spanish	Only the web presentation of the agency	PUC1-J PUC1-MM	Restricted (the news articles can be reached only through dpa service)
16	agencia EFE	http://www.efe.com/efe/noticias/espana/1 http://www.efe.com/efe/noticias/english/4	Spanish English	News articles, photos, video, audio	PUC1-J PUC1-MM	Restricted (free are only articles from the "contenidos gratuitos"
17	German Federal Ministry for Economic Affairs and Energy	http://www.bmwi.de/DE/root.html http://www.bmwi.de/EN/root.html http://www.bmwi.de/FR/root.html	German English French	News press materials publications videos, photos (mediathek only in German)	PUC1-J	free
18	Agency for the Cooperation of Energy Regulators	http://www.acer.europa.eu/Pages/ACER.aspx	English	News, press releases, meeting minutes, official documents	PUC1-J	free
19	German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety	http://www.bmub.bund.de/ http://www.bmub.bund.de/en/	German English	News on climate and energy, speeches (in text format), videos	PUC1-J	free
20	French Ministry for Ecology, Development and Energy	http://www.developpement-durable.gouv.fr/	French	News articles, press releases, different publications	PUC1-J	free
21	European Commission: Energy	http://ec.europa.eu/energy/index_en.htm http://ec.europa.eu/energy/index_de.htm http://ec.europa.eu/energy/index_fr.htm	English German French	Articles, Multimedia (often only as links to external pages, most of them in English) Videos mostly only in English	PUC1-J	free
22	Spanish ministry of industry, energy and	http://www.minetur.gob.es/energia/es-ES/Paginas/index.aspx	Spanish English	Papers, reports, statistics, documentation, talks and speeches etc.	PUC1-J	free

	tourism	http://www.minetur.gob.es/energia/en-us/Paginas/Index.aspx				
23	UN Energy knowledge network	http://www.un-energy.org/	English	News, publications, newsletters. Tools	PUC1-J	free
24	International association for energy economics	http://www.iaee.org/en/	English	Publications and abstracts	PUC1-J	restricted (publications should be bought; abstracts are for free)
25	Gesellschaft für Energiewissenschaft und Energiepolitik e. V.	http://www.gee.de/ http://www.gee.de/?lang=en	German English	News, calendar, events …	PUC1-J	free
26	Verein für ökologisch-solidarische Energie- & Weltwirtschaft e.V.	http://power-shift.de/ http://power-shift.de/?lang=en	German (English not really available)	Articles, reports, studies, videos	PUC1-J	free
27	Climate Action Network Europe	http://www.caneurope.org/	English	News, press releases, publications, letters, public consultations; policies	PUC1-J	free
28	Öko-Institut e.V.	http://www.oeko.de/ http://www.oeko.de/en/	German English	Publications, projects, e-papers	PUC1-J	free
29	Bundesverband Erneuerbare Energien (BEE)	http://www.bee-ev.de/ http://www.bee-ev.de/BEE/English.php	German (English not really available; only a single abstract)	Articles, press releases, publications	PUC1-J	free
30	Regional Center for Energy Policy Research	http://www.rekk.eu/index.php?lang=en	English (original Hungarian)	Publications, reports, books	PUC1-J	free
31	Central European University - Center for Climate Change and Sustainable Energy Policy	http://3csep.ceu.hu/	English (mixed with Hungarian)	News, publications, project descriptions	PUC1-J	free
32	Central European University – Energy Policy Research Group	http://energy.ceu.hu/	English	News, events Publications (only abstracts, with links to book shops)	PUC1-J	free
33	European Energy	http://www.eera-set.eu/	English	Press releases, articles, interviews	PUC1-J	free

	Research Alliance					
34	Florence School of Regulation - Energy	http://fsr.eui.eu/FlorenceSchoolofRegulation/Energy/Index.aspx	English	Publications (video&audio, papers, presentations, books, reports …)	PUC1-J	free
35	Prof. Dr. Lorenz Jarass	http://www.jarass.com/home/index.php/DE/energie http://www.jarass.com/home/index.php/en/energie	German (English not really available)	Books, scientific papers, lectures	PUC1-J	free
36	Brigitte	http://www.brigitte.de/	German	Articles on fashion, nutrition, lifestyle, shopping, health, culture and travel. Forum	PUC1-MM	free
37	Bunte	http://www.bunte.de/	German	Articles about lifestyle, beauty, fashion, celebrities	PUC1-MM	free
38	Elle	http://www.elle.de/ http://www.elle.fr/ http://www.elleuk.com/ http://www.elle.es/	German French English Spanish	Articles about fashion, beauty, lifestyle, travel. Blogs	PUC1-MM	free
39	Elle Decoration	http://www.elle.de/elle-decoration-73332.html http://www.elledecoration.co.uk/	German English	Articles about home design	PUC1-MM	free (German) not clear (English)
40	Glamour	http://www.glamour.de/	German	Articles about fashion, beauty, celebrities	PUC1-MM	free
41	inStyle	http://www.instyle.de/	German	Articles about fashion and shopping, videos, blogs	PUC1-MM	free
42	Jolie	http://www.jolie.de/	German	fashion, beauty, lifestyle	PUC1-MM	free
43	GQ	http://www.gq-magazin.de/	German	Fashion magazine for men; automobile, technique, nutrition, fitness, travel	PUC1-MM	free
44	Architectural Digest	http://www.ad-magazin.de/ http://www.admagazine.fr/ http://www.revistaad.es/ http://www.architecturaldigest.com/	German French Spanish English	Articles about design, architecture, art, lifestyle, home. Images	PUC1-MM	free
45	Elektrojournal	http://www.elektrojournal.at/	German	Articles about electronic devices and products; articles, newsletters	PUC1-MM	free
46	Küchenmagazin	http://www.kueche-co.de/kuechenmagazin/	German	Descriptions and advice-giving articles on kitchen and kitchenware. Images	PUC1-MM	free
47	Elektronik JOURNAL	http://www.elektronik-journal.de/	German	Articles about power electronics, medical electronics equipment, electromechanics etc.	PUC1-MM	free
48	APPLIANCIST	http://www.appliancist.com/	English	Articles and images about home, kitchen and bathroom appliances	PUC1-MM	free

49	Spanish Office of Economy and Competitiveness	http://www.oficinascomerciales.es/icex/cda/controller/pageOfecomes/0,5310,5280449_5296122_5296234_0_DE,00.html http://www.oficinascomerciales.es/icex/cda/controller/pageOfecomes/0,5310,5280449_5296130_5296234_0_DE,00.html	Spanish	General information about Germany Practical information about Germany	PUC2	free free
50	Central Intelligence Agency – The world factbook	https://www.cia.gov/library/publications/the-world-factbook/geos/gm.html	English	Economic overview for Germany	PUC2	free
51	Trading Economics	http://www.tradingeconomics.com/germany	English	Economic indicators, exchange rates, stock market indexes …	PUC2	restricted (fee required)
52	German Missions in the United States	http://www.germany.info/Vertretung/usa/en/02__GIC/GIC/00/__Home.html	English	Information about Germany: holidays, traditions, facts, visa, passport, legal information	PUC2	free
53	German Center of Information for Spain and Latin America	http://www.alemaniaparati.diplo.de/Vertretung/mexiko-dz/es/03-PoliticaExterior/Lazos/0-LazosEspaniaYLatinoamerica.html	Spanish	General information about Germany and its foreign policy in Spain and Latin America	PUC2	free
54	Spanish Ministry of Foreign Affairs and Cooperation	http://www.exteriores.gob.es/Portal/es/SalaDePrensa/Paginas/FichasPais.aspx	Spanish	Information about Germany and bilateral relations between Spain and Germany	PUC2	free
55	EUROSTAT	http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database	English German French	Databases with statistics on economy, finance, industry, trade, population, social conditions etc. in European countries and regions	PUC2	free
56	German Federal Ministry of Economics and Technology	http://www.bmwi.de/English/Redaktion/Pdf/facts-about-german-foreign-trade-in-2012,property=pdf,bereich=bmwi2012,sprache=en,rwb=true.pdf	English	A single PDF file	PUC2	free
57	German Trade and Invest	http://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Industry-overviews/industry-overview-food-beverage-industry-en.pdf	English	A single PDF file	PUC2	free
58	Karen Juliane Schröder: Cannibalization on the	http://www.agric-econ.uni-kiel.de/arbeiten_PDFs/2012/MA2012S	English	A single PDF file	PUC2	free

	yoghurt market	chroederML.pdf				
59	Industry Analysis: Competitors	http://www.euromonitor.com/yoghurt-and-sour-milk-products-in-germany/report	English	Report about Yoghurt and Sour Milk Products in Germany	PUC2	Restricted (the full report has to be bought)
60	Industry Analysis: Competitors	http://en.wikipedia.org/wiki/M%C3%BCller_(company)	English	Wikipedia	PUC2	free
61	Industry Analysis: Competitors	https://www.google.es/search?q=ehrmann+joghurt&oq=ehrmann+jog&aqs=chrome.1.69i57j0l5.6484j0j4&sourceid=chrome&espv=210&es_sm=93&ie=UTF-8	German	Google search (search results)	PUC2	free
62	Europa – summaries of EU legislation (labeling, presentation and advertising of foodstuffs)	http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_en.htm (*_bg.htm, *_de.htm, *_es.htm, *_fr.htm)	English Bulgarian German Spanish French	summaries of EU legislation	PUC2	free
63	International Dairy Food Association (European health certification program)	http://www.idfa.org/files/resources/eu_health_certification_program_draft_091411.pdf	English	A single PDF file	PUC2	Page not found!
64	German Business Portal (Overview of market access of food and beverage)	http://www.ixpos.de/IXPOS/Navigation/EN/Your-business-in-germany/Business-sectors/Consumer-goods/food-and-beverage,did=263444.html	English	Industry overview of German food and beverage market	PUC2	free
65	IFS Food Packaging Guideline	http://www.ifs-certification.com/index.php/en/imprint-left-en/51-global-news/2005-news-2013-10-23-vplf-v2-en	English German	IFS Food Packaging Guideline	PUC2	Fee required The full guideline only in the IFS shop available
66	General requirements and standards for food and agricultural imports into Germany	http://www.spring.gov.sg/archives/ETAC/Documents/Germany.pdf	English	A single PDF file	PUC2	free
67	PEPPOL	http://www.peppol.eu/	English	Pan-European Public Procurement OnLine (PEPPOL) is an international project that aims at standardisation of cross-border electronically-supported	PUC2 (listed in theDoW)	Access restricted to registered users only

				public procurement procedures within the European Union.		
68	decor8blog	http://decor8blog.com/	English	Blog	PUC1-MM	free
69	Clean Technica	http://cleantechnica.com/	English	Blog	PUC1-J	free
70	Council on foreign relations – Energy, Security and Climate	http://blogs.cfr.org/levi/	English	Blog	PUC1-J	free
71	Volker Quaschning	http://volker-quaschning.de/index.php http://volker-quaschning.de/index_e.php	German English	Blog	PUC1-J	free
72	Smart Grid Watch Blog	http://smartgridwatch.wordpress.com/	English	Blog	PUC1-J	free
73	greenliving	http://greenlivingonline.com/blog	English	Blog	PUC1-MM	free
74	DIISIGN	http://www.diisign.com/	French	Blog	PUC1-MM	free
75	HOSPIMEDIA	http://blog.hospimedia.fr/	French	Blog	PUC1-MM	free
76	Banco Sabadell	http://blog.bancsabadell.com/	Spanish	Blog	PUC2	free
77	TED (tenders electronic daily)	http://ted.europa.eu/TED/main/HomePage.do	English German French Spanish Bulgarian	Supplement to the Official Journal of the European Union	PUC2 (listed in theDoW)	For registered users
78	SPOCS	http://www.eu-spocs.eu/index.php	English German French	SPOCS aims to build the next generation of online portals (Point of Single Contact or PSC), which every European country now has in place, through the availability of high impact cross- border electronic procedures.	PUC2 (listed in theDoW)	For registered users
79	MADB	http://madb.europa.eu/madb/indexPubli.htm	English (other languages listed but not really available)	The Market Access Database (MADB) gives information to companies exporting from the EU about import conditions in third country markets.	PUC2 (listed in theDoW)	Data provided by the German publisher Mendel, and subject of a license between the Mendel Verlag and EU (the user has to accept it before using the data base)

B Appendix: Empirical study – Detailed description of single data sources

1	SZ (Süddeutsche Zeitung)
URL	http://www.sueddeutsche.de/
Language	German
Content	News texts, Images (slide shows), Videos
Format	html
Keyword Search available Query included into URL Search results Paywall restricted	Yes Yes Example: http://suche.sueddeutsche.de/?query=Energie&Finden=Finden Single articles, slide shows, Videos No
Query filters included into URL Query filters: - Resort - Document type - Person names - Locations - Keywords - Companies - Author	Yes Examples: http://suche.sueddeutsche.de/query/Energie/sort/-news/drilldown/%C2%A7ressort%3A%5EPolitik%24 http://suche.sueddeutsche.de/query/Energie/sort/-news/drilldown/%C2%A7documenttype%3AArtikel http://suche.sueddeutsche.de/query/energie/sort/-news/drilldown/%C2%A7personnames%3A%5E%22Angela%20Merkel%22%24 http://suche.sueddeutsche.de/query/energie/sort/-news/drilldown/%C2%A7locations%3A%5EEuropa%24 http://suche.sueddeutsche.de/query/energie/sort/-news/drilldown/%C2%A7concepts%3A%5EEuro%24 http://suche.sueddeutsche.de/query/energie/sort/-news/drilldown/%C2%A7companies%3A%5ESiemens%24 http://suche.sueddeutsche.de/query/energie/sort/-news/drilldown/%C2%A7author%3A%5E%22Markus%20Balser%22%24
Useful content	Date and time, author (von XX), title, description (for videos and images), the content itself (text, video, blog …)
Useful metadata	<meta name="description" content="Anadarko hat von South Dakota bis Chicago verseuchte Gebiete hinterlassen. Nun muss der Konzern mehr als fünf Milliarden Dollar zahlen. "> <meta name="news_keywords" content="Anadarko, Justiz, Umwelt, Umweltverschmutzung, Uran, Öl, Wirtschaft, USA, Dollar, Chicago"> <meta name="last-modified" content="Fr, 4 Apr 2014 12:36:02 MESZ"> <meta property="og:locale" content="de_DE">
Useful info in the URL	Language, search term, Query filters
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (language, date, keywords, description) Information Extraction for text Image processing for slide shows Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “Energie”): Very high number of texts (18.000) and images (more than 1.000); more than 100 videos Puc1-MM (e.g. “Kitchenaid”: only 4 items; “Liebherr”: no matches; “AEG”: ca.150 matches)

2	FAZ
URL	http://www.faz.net/
Language	German
Content	News texts, Images (slide shows), Video, Audio, Blog
Format	html
Keyword Search available Query included into URL Search results Paywall restricted	Yes Yes. Example: http://www.faz.net/suche/?query=Energie&resultsPerPage=20&suchbegriffImage.x=19&suchbegriffImage.y=9 Single articles from the archive Yes (images, slide shows)
Topic search available Topic included into the URL Search results Paywall restricted	Yes (http://www.faz.net/themen/) Yes. Examples: http://www.faz.net/aktuell/wirtschaft/thema/energiegipfel http://www.faz.net/aktuell/wirtschaft/thema/aeg http://www.faz.net/aktuell/technik-motor/thema/liebherr Single articles on the topic No
Query filters included into URL Query filters: - Time period - Resort (politics, economy, etc.) - Document type (article, blog, video, audio, slideshow) - Editorial content / Readers' opinion - Author / Reader	Yes Example: http://www.faz.net/suche/?offset=&cid=&index=&query=energie& offset=&allboosted=&boostedresultsize=%24boostedresultsize& from=01.02.2014&to=07.04.2014&chkBox_2=on& BTyp=redaktionelleInhalte&chkBoxType_6=on &author=&username=&sort=date&resultsPerPage=20
Useful content	Date and time, author (von XX), title, description (for videos and images), the content itself (text, video, blog …)
Useful metadata	<meta name="last-modified" content="2014-01-29T16:35:06+0100"/> <meta property="og:description" content="Nach dem Atomunfall in Fukushima hat Deutschland in Windeseile die Energiewende ausgerufen. Aber worauf basiert die Zäsur? Die in..."/> <meta name="keywords" content="UN, Strahlenbilanz, Tsunami, Erdbeben, Energiewende, Fukushima, Japan"/> <meta name="language" content="Deutsch"/> <meta http-equiv="Content-Language" content="de"/>
Useful info in the URL	Keywords, Topic
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (language, date, keywords, description) Information Extraction for text Image processing for slide shows Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. "Energie"): Very high number of texts (14.000) and images (more than 1.000); more than 70 videos, only 1 audio, no slideshows … Puc1-MM (e.g. "Kitchenaid": only 3 items; "Liebherr": ca. 40 matches, no multimedia; "AEG": ca.280 matches)

3	Die Zeit
URL	http://www.zeit.de/index
Language	German
Content	News texts, videos
Format	html
Keyword search available Query included into URL Search results Paywall restricted	Yes Yes. Example: http://www.zeit.de/suche/index?q=energie News articles; book reviews No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (http://www.zeit.de/schlagworte/themen/A/index) Yes Examples: http://www.zeit.de/schlagworte/themen/energie/index http://www.zeit.de/schlagworte/themen/energiepolitik/index http://www.zeit.de/schlagworte/themen/haushaltsgeraete/index Single articles on the topic No
Query filters included into URL Query filters: - Time period (today, 24 hours, 7 days, 30 days, exact) - All items / only reviews	Yes Examples: http://www.zeit.de/suche/index?q=energie&tmode=today&sort=aktuell&rezension=0 http://www.zeit.de/suche/index?q=energie&tmode=7d&sort=aktuell&rezension=1
Useful content	Date and time, title, description (for videos and images), the content itself (text, video, blog …), number of reader comments, comments themselves
Useful metadata	<meta name="date" content="2006-04-20T14:00:00+0200"> <meta name="keywords" content="Leben, Globalisierung, Industrie, Wirtschaft und Konjunktur, Wettbewerb, Elektroindustrie, Portraits, Haushaltsgeräte, China, Siemens AG, Globalisierung, China, Huawei, Indien, Bangalore, Kolumbien, Norwegen, Pakistan, Kolkata, Peking, Silicon Valley"> <meta name="description" content="Siemens lässt Software in Bangalore und Peking entwickeln. Die Mitarbeiter dort gehören zu den Gewinnern der Globalisierung – und machen sich gegenseitig Konkurrenz."> <meta property="og:description" content="Soma und Venugopal Sharma leiten für Siemens jeweils ein Team von jungen Software-Ingenieuren in Bangalore">
Useful info in the URL	Keywords, Topic
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, keywords, description) Information extraction for text Image and video processing + ASR for videos
Use case coverage / relevance	Puc1-J (e.g. “Energie”): Very high number of texts (34.000) and reviews (ca. 160) Puc1-MM (e.g. “Kitchenaid”: only 1 item; “Liebherr”: ca.1.000 matches, but most of them for “lieber Herr” (only exact search not possible); “AEG”: ca.2.000 matches)

4	Der Spiegel
URL	http://www.spiegel.de
Language	German
Content	News articles, videos, interactive maps and graphs
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes. Example: http://www.spiegel.de/suche/index.html?suchbegriff=energie Single news articles, as well as links to topics (cf. next) No
Topic search available Topic included into the URL Search results Paywall restricted	Yes http://www.spiegel.de/thema/index-a.html Yes. Examples: http://www.spiegel.de/thema/erneuerbare_energien/ http://www.spiegel.de/thema/haushaltsgeraete/ http://www.spiegel.de/thema/aeg/ special topic pages (articles, videos, interactive maps, graphs to the topic in respect) No
Query filters included into URL Query filters: - Source (Spiegel Online (SPOX); Der Spiegel (SP); manager-magazin.de (MMOX)) - Full text, or only title and the headlines - Author - Time period	Yes Examples: http://www.spiegel.de/suche/index.html?suchbegriff=energie&quellenGroup=SPOX&suchbereich=kopftext&suchzeitraum=week&fromDate=01.04.2014 http://www.spiegel.de/suche/index.html?suchbegriff=energie&quellenGroup=SP&quellenGroup=MMOX&suchzeitraum=ab2005&fromDate=01.01.2005
Useful content	Date, title, description (for videos and images), the content itself (text, video, blog …), author (only an abbreviation)
Useful metadata	<meta name="date" content="2014-04-06T12:27:00+0200" /> <meta property="og:description" content="Der Kompromiss zwischen Bund und Ländern bei der Reform der Energiewende wird Stromkunden teuer zu stehen kommen. Nach Informationen des SPIEGEL werden Verbraucher in den kommenden sechs Jahren mit rund zehn Milliarden Euro belastet." /> <meta name="keywords" content="Mehrkosten, Verbraucher, EEG, Stromkunden, Deal, Ökostrom, Wirtschaft, Verbraucher & Service, Hannover Messe, Energiewende, Erneuerbare-Energien-Gesetz, Strompreis, Erneuerbare Energien, Energiewirtschaft, Stromnetze, Strom, Energieeffizienz" />
Useful info in the URL	Language, keywords, topic, search filters … Number of search query matches (http://www.spiegel.de/suche/index.html?suchbegriff=energie&offsets=24562&pageNumber=20)
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (data, keywords, description) URL parsing for: keywords, topics, language, number of matches for a search query Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “Energie”): Very high number of texts (24.000) + many relevant topics such as Energieeffizienz, Energietechnologie, Energiewende, Energiewirtschaft … Puc1-MM (e.g. “Kitchenaid”: only 2 matches; “Liebherr”: ca.90 matches; “AEG”: ca.1.800 matches) + topics such as AEG, Haushalstsgeräte, Haustechnik etc.

5	Le Figaro
URL	http://www.lefigaro.fr/
Language	French
Content	News texts, embedded videos
Format	html
Keyword search available Query included into URL Search results Paywall restricted	Yes Yes. Example: http://recherche.lefigaro.fr/recherche/recherche.php?ecrivez=%C3%A9nergie&go=Rechercher - News articles - News flash („Flash actu“) Paywall restriction for all articles from the archive (older than two days. No restriciton for „flash actu“
Query filters included into URL Query filters: - Ressort - Time period - Keywords - Source - Document type - Person names - Location - Company	Yes Examples: http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=énergie&tri%5B0%5D=Rubrique&critere%5B0%5D=Sciences&rubrique=Sciences http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=énergie&tri%5B0%5D=Date&critere%5B0%5D=Ces+7+dernier+jours&date=cettesemaine http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=%C3%A9nergie&tri%5B0%5D=Mot-cle&critere%5B0%5D=environnement&mot-cle=environnement http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=%C3%83%C2%A9nergie&tri%5B0%5D=Source&critere%5B0%5D=Figaro+Magazine&source%5B0%5D=FIGARO+MAGAZINE&source%5B1%5D=FIGARO+MAGAZINE+SANTE http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=%C3%83%C2%A9nergie&tri%5B0%5D=Type&critere%5B0%5D=Articles&type=ART http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=%C3%83%C2%A9nergie&tri%5B0%5D=Lieu&critere%5B0%5D=Allemagne&geo=Allemagne http://recherche.lefigaro.fr/recherche/recherche.php?page=articles&ecrivez=%C3%83%C2%A9nergie&tri%5B0%5D=Soci%C3%A9t%C3%A9&critere%5B0%5D=EDF&company=EDF
Useful content	Author, Date and time (Publié le 14/04/2014 à 06:00), title, content itself (video, text)
Useful metadata	<meta name="description" content="Équipée de capteurs photovoltaïques, une nouvelle génération d'appareils électroniques, écologiques et économiques, commence à émerger."/> <meta name="news_keywords" content="High-tech, Logitech, Kudo, Eton, TAG Heuer, WeWi, Énergie solaire, High-Tech"/> <meta name="DC.date.issued" content="2014-04-07T12:01:13+02:00"><!--heure de publi-->
Useful info in the URL	Search keyword, search filters, language
How to extract info?	Web crawling (generic or customised) (RSS feed) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, keywords, description) Information extraction for text analysis Video processing + ASR for videos
Use case coverage / relevance	Puc1-J (e.g. “énergie”): ca. 1.200 news articles (but only the newest ones available for free); ca. 450 ‚flash actu‘ articles Puc1-MM (e.g. “Kitchenaid”: only 4 items for free (flash actu), 1 article; “Liebherr”:3 flach actu, 15 articles; “AEG”: ca.35 flash actu, ca. 50 articles

6	Le Monde
URL	http://www.lemonde.fr/
Language	French
Content	News texts, embedded videos
Format	Html
Keyword search available Query included into URL Search results Paywall restricted	Yes Yes Example: http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&qt=recherche_globale News articles No
Query filters included into URL Query filters: - Exclude keyword - Only title / entire document - Author - Date	Yes Examples: http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=renouvelables&qt=recherche_texte_titre&author=&period=since_1944&start_day=01&start_month=01&start_year=1944&end_day=11&end_month=03&end_year=2014&sort=desc http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=&qt=recherche_texte_titre&author=&period=since_1944&start_day=01&start_month=01&start_year=1944&end_day=11&end_month=03&end_year=2014&sort=desc http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=&qt=recherche_titre&author=&period=since_1944&start_day=01&start_month=01&start_year=1944&end_day=11&end_month=03&end_year=2014&sort=desc http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=&qt=recherche_titre&author=Darge&period=since_1944&start_day=01&start_month=01&start_year=1944&end_day=11&end_month=03&end_year=2014&sort=desc http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=&qt=recherche_titre&author=&period=for_1_day&start_day=01&start_month=01&start_year=1944&end_day=11&end_month=03&end_year=2014&sort=desc http://www.lemonde.fr/recherche/?keywords=%C3%A9nergie&page_num=1&operator=and&exclude_keywords=&qt=recherche_titre&author=&period=for_1_week&start_day=10&start_month=03&start_year=2014&end_day=11&end_month=03&end_year=2014&sort=desc
Useful content	Author (Par XX), , Date and time (14.04.2014 à 11h32), title, content itself (video, text)
Useful metadata	<meta property="og:description" content="Point de vue L'Ukraine est dépendante de la Russie pour son approvisionnement en gaz et en pétrole. Les tensions avec Moscou l'exposent à de nouveaux conflits."> <meta property="og:locale" content="fr_FR"> <time datetime="2014-03-31T16:10:28+02:00" itemprop="dateModified">31.03.2014 à 16h10</time>
Useful info in the URL	Search keyword, search filters, language, date
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (description, language, date) Information extraction for text analysis Video preprocessing + ASR for videos
Use case coverage / relevance	Puc1-J (e.g. “énergie”): ca. 105.000 news articles Puc1-MM (e.g. “Kitchenaid”: only 8 articles, “Liebherr”:19 articles; “AEG”: ca.330 articles

7	Libération
URL	http://www.liberation.fr
Language	French
Content	News articles Videos
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes Example: http://www.liberation.fr/recherche/?q=%C3%A9nergie Single articles as well as links to dossiers (articles, videos and interviews from the archive to the search topic) No
Query filters included into URL Query filters: - Time period - Source (all, website, journal) - Resort	Yes Examples: http://www.liberation.fr/recherche/?q=%C3%A9nergie& period=last_365&period_start_day=0&period_start_month=0&period_start_year=0&period_end_day=0&period_end_month=0&period_end_year=0& editorial_source=&paper_channel=&sort=-publication_date_time Resort (rubrique du quotidien) is not included into the URL
Useful content	Author , Date and time (Gabriel SIMÉON 13 avril 2014 à 18:36), title, description (video), content itself (video, text)
Useful metadata	<meta property="og:type" content="article" > <meta property="og:type" content="media" /> <meta name="news_keywords" content="Iran, négociation, armement nucléaire, Agence internationale de l'énergie atomique (AIEA), diplomatie" >
Useful info in the URL	Search keywords, restrictions Language and Date: http://www.liberation.fr/terre/2012/10/02/la-biodiversite-ce-n-est-pas-seulement-les-baleines-et-les-ours-blancs_850358
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (keywords, media type) URL parsing (for language and date) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “énergie”): 1.000 results (no statistics on how many articles and how many videos) Puc1-MM (e.g. “Kitchenaid”: 5 matches, “Liebherr”: 11 matches; “AEG”: 70 matches (most of them referring to the concert promoter AEG))

8	El Mundo
URL	http://www.elmundo.es/
Language	Spanish
Content	News articles, videos, photos, blogs
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes Example: http://ariadna.elmundo.es/buscador/archivo.html?q=Liebherr&t=1&n=10&s=1&w=70 News articles, Blogs, Fotos, Videos No
Query filters included into URL Query filters: - Time (today, yesterday, last week, last month, 2014, 2013 …2000) - Resort - Source (supplements Motor, Yodona etc.)	Yes Examples: http://ariadna.elmundo.es/buscador/archivo.html?q=energ%eda&t=1&n=10&fd=0&td=0&w=70&s=1&no_acd=1&seccion=internacional&parametric_year=2013 http://ariadna.elmundo.es/buscador/archivo.html?q=energ%eda&t=1&n=10&fd=0&td=0&w=70&s=1&no_acd=1&suplementos=motor
Useful content	Author , Date and time (Actualizado: 24/03/2014 19:51 horas), title, description (for videos), content itself (video, text)
Useful metadata	<meta property="article:published_time" content="2014-04-07T18:20:43+02:00"/> <meta property="og:description" content="La Comisión Nacional de los Mercados y de la Competencia (CNMC) ha cifrado en 1.700 millones de euros el recorte sobre la retribución de las energías renovables en 2014. Así consta"/>
Useful info in the URL	Search Keywords, restrictions Date: http://www.elmundo.es/economia/2014/04/07/5342d059e2704ee4648b4580.html
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, description) URL parsing (for language and date) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “energía”): 23.500 articles (+ 1.300 from American edition), 1.500 blog results, 130 photos, more than 100 videos Puc1-MM (e.g. “Kitchenaid”: no matches, “Liebherr”: 9 matches; “AEG”: 4 matches (all of them referring to the concert promoter AEG))

9	El País
URL	http://elpais.com/
Language	Spanish
Content	News articles
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes No News articles No
Topic search available Topic included into the URL Search results Paywall restricted	Yes Yes Examples: http://deportes.elpais.com/tag/energia/a/ http://deportes.elpais.com/tag/electrodomesticos/a/ Collections of news articles related to the search topic No
Useful content	Author , Date and time (13 ABR 2014 - 20:39 CET), title, content itself (article text)
Useful metadata	<meta name="lang" content="es" /> <meta name="description" content="La eólica acapara el grueso del ajuste con una rebaja de 608 millones, el 34% de sus primas. En la termosolar es del 13%" /> <meta name="keywords" content="recortar, 1.671 millones, renovable, eólica, eólico, acaparar, grueso, ajuste, rebaja, 608 millones, 34 %, prima, reducción, termosolar, fotovoltaica, fotovoltaico, ser, 13 %" /> <meta name="DC.date.issue" content="2014-04-07" />
Useful info in the URL	Resort, date http://economia.elpais.com/economia/2014/04/07/actualidad/1396888065_742601.html
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (language, date, keywords, description) URL parsing (for resort (topic) and date) Information Extraction for text analysis
Use case coverage / relevance	Puc1-J (e.g. “energía”): more than 18.000 articles + several topic pages (Energía eléctrica, Energía eólica, Energías renovables, Energía solar, Reforma energética ...) Puc1-MM (e.g. “Kitchenaid”: ca. 150 matches, “Liebherr”: 4 matches; “AEG”: ca. 170 matches (all of them referring to the concert promoter AEG))

10	The Guardian
URL	http://www.theguardian.com/uk
Language	English
Content	News articles, videos, blogs
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes No A pop-up window with links to single articles No paywall restrictions, but only the first 10 pages of the search results are accessible
Topic search available Topic included into the URL Search results Paywall restricted	Yes Yes Examples: http://www.theguardian.com/environment http://www.theguardian.com/environment/renewableenergy Videos, articles, blogs No
Useful content	Author, Place (Damian Carrington, Berlin), day, date and time (Sunday 13 April 2014 11.19 BST), title, content itself (article text), number of comments, comments
Useful metadata	<meta name="description" content="Financial incentives for homeowners off the gas grid to switch to technologies such as biomass boilers" /> <meta name="DC.date.issued" content="2014-04-09"> <meta property="article:tag" content="Environment" /> <meta property="article:tag" content="Energy efficiency" /> <meta property="article:tag" content="Energy bills" /> <meta property="article:tag" content="Money" /> <meta property="article:tag" content="Carbon emissions" /> <meta property="article:tag" content="Climate change" /> <meta name="author" content="Adam Vaughan" /> <meta name="news_keywords" content="Renewable energy,Environment,Energy efficiency,Energy bills,Money,Carbon emissions,Climate change,Environment" />
Useful info in the URL	http://www.theguardian.com/environment/southern-crossroads/2014/apr/09/tony-abbott-carbon-price-wester-australia-senate-election
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, description) URL parsing (for language and date) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “energy”): more than 230.000 articles (but only ca. 100 of them reachable from the browser) + topic pages (Environment, Energy , Energy efficiency, Energy monitoring, Wind power...) Puc1-MM (e.g. “Kitchenaid”: ca. 120 matches, “Liebherr”: ca. 130 matches (most of them as last name not connected to the brand); “AEG”: ca. 500 matches (all of them referring to the concert promoter AEG) + few topic pages, e.g. Kitchen Gadgets: the test

11	The Times
URL	http://www.thetimes.co.uk/tto/news/
Language	English
Content	News articles, videos, images
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes Example: http://www.thetimes.co.uk/tto/public/sitesearch.do?querystring=energy&p=tto&pf=all&bl=on Articles (containing videos, image galleries and interactive graphics) – Yes (full articles, as well as videos, images and graphics only accessible for subscribers)
Query filters included into URL Query filters: - Time period (last seven days, 30 days, exact time period) - Section - Publication (The Times, The Sunday Times) - Journalist - Contains video, image gallery, or interactive graphics	Yes Examples: http://www.thetimes.co.uk/tto/public/sitesearch.do?querystring=energy&p=tto&pf=all&sectionId=690&bl=on#/tto/public/sitesearch.do?querystring=energy&filters=date_published_7days:[NOW/DAY-7DAY TO NOW]&offset=0&hits=0&bl=on&service=searchframe http://www.thetimes.co.uk/tto/public/sitesearch.do?querystring=energy&p=tto&pf=all&sectionId=690&bl=on#/tto/public/sitesearch.do?querystring=energy&navigators=sectionname0:Business&offset=0&hits=0&bl=on&service=searchframe http://www.thetimes.co.uk/tto/public/sitesearch.do?querystring=energy&p=tto&pf=all&sectionId=690&bl=on#/tto/public/sitesearch.do?querystring=energy&navigators=article_authors:Angela+Jameson&offset=0&hits=0&bl=on&service=searchframe
Useful content	Author (Lindsay McIntosh Scottish Political Correspondent), date and time (Last updated at 12:01AM, April 8 2014), title, content itself (only a few lines; the rest only for subscribers)
Useful metadata	<meta name="dashboard_published_date" content="Mon Mar 29 2010 10:13 UTC+1013" /> <meta name="description" content="Ambitious programme aims to create 2,500 jobs and cut household energy bills by at least £1bn over the next ten years" /> <meta property="fb:locale" content="en_GB"/>
Useful info in the URL	Query string, section http://www.thetimes.co.uk/tto/business/industries/utilities/article1690959.ece
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, description, language) URL parsing (section, topic) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “energy”): ca. 75.000 articles (more than 900 of them containing videos, more than 600 with image galleris, and ca. 80 with interactive graphics) Puc1-MM (e.g. “Kitchenaid”: ca. 40 (3 of them with multimedia), “Liebherr”: ca. 80 matches (most of them as last name not connected to the brand); “AEG”: ca. 300 matches (most of them referring to the concert promoter AEG)

12	The daily telegraph
URL	http://www.telegraph.co.uk/
Language	English
Content	News articles
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes Example: http://www.telegraph.co.uk/search/?queryText=energy&Search= Single articles. Links to related topic areas with many articles and videos (if available). No
Query filters included into URL Query filters: - Section (telegraph.co.uk, News, Sport, Travel, Culture, Finance, Lifestyle, Blogs, …)	Yes Examples: http://www.telegraph.co.uk/search/?queryText=energy&site=telegraph_finance http://www.telegraph.co.uk/search/?queryText=energy&site=telegraph_blogs
Useful content	Author (By Anna White, Enterprise and property correspondent), date and time (7:00AM BST 13 Apr 2014), title, content itself, number of comments, comments
Useful metadata	<meta name="description" content="Solar farms must not become "the new onshore wind", Greg Barker, the <meta name="keywords" content="Solar Power,Energy,Earth" /> <meta name="last-modified" content="2014-04-04" />
Useful info in the URL	Language, topic: http://www.telegraph.co.uk/earth/energy/solarpower/10744891/Energy-minister-vows-to-curb-the-spread-of-solar-farms.html
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (date, description, keywords) URL parsing (language, topic) Information Extraction for text analysis
Use case coverage / relevance	Puc1-J (e.g. “energy”): ca. 100.000 from different sections Puc1-MM (e.g. “Kitchenaid”: ca. 250 matches, “Liebherr”: ca. 200 matches (most of them as last name not connected to the brand); “AEG”: ca. 700 matches (most of them referring to the concert promoter AEG)

13a	Reuters – UK edition *
URL	http://uk.reuters.com/
Languages	English
Content	News articles, images, videos, blogs
Format	html
Keyword search available Query included into URL Search results Paywall restricted	Yes Yes Example: http://uk.reuters.com/search?blob=energy News: http://uk.reuters.com/search?blob=energy Blogs. http://uk.reuters.com/search/blog?blob=energy Videos: http://uk.reuters.com/search/video?blob=energy Pictures: http://uk.reuters.com/search/pictures?blob=energy No
Useful content	Day, date and time (Mon Apr 14, 2014 11:00am BST), title, content itself (text, video) , author (by XX), number of comments, comments
Useful metadata	<META name="description" content="U.S. Environmental Protection Agency Recognises Consumers Energy as ENERGY STAR® Partner of the Year <META name="REVISION_DATE" content="Tue Apr 08 14:31:04 UTC 2014"> Blogs: <meta name="DCSext.rChannel" content="Blogs" /> <meta name="DCSext.ContentHeadline" content="Financing more solar energy - MuniLand" /> <meta name="DCSext.rAuthor" content="Cate Long" /> <meta name="description" content="There is another rapidly growing area in site-based residential or commercial solar installations." /> <meta name="keywords" content="solar" /> Video: <meta id="MetaDescription" name="description" content="March 12 - Germany's biggest utility plans to halve its dividend for 2013 and shut more than a quarter of its power plants. As Hayley Platt reports its in response to a surprise rise in renewables across Europe."></meta> <meta id="MetaKeywords" name="keywords" content="emerging markets, Houston, nuclear power, Ukraine, Suez, ITT, management consulting, utility companies, Russia, business energy oil, renewable energy, Reuters, Video, News, Business, Finance, Technology ">
Useful info in the URL	Date, kind of medium (article, blog, picture) http://uk.reuters.com/article/2014/04/08/britain-utilities-idUKL5N0MS46Z20140408 http://blogs.reuters.com/muniland/2013/10/15/financing-more-solar-energy/ http://uk.reuters.com/news/pictures/searchpopup?picId=630902634
How to extract info?	Web crawling (generic or customised) Html parsing (boilerplate removal) to obtain the text Metadata extraction (data, keywords, description) URL parsing for: keywords, topics, date, kind of medium Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. “Energy”): ca. 108.000 news articles; 157.000 blog posts; ca. 450 videos, ca. 420 pictures Puc1-MM (e.g. “Kitchenaid”: ca. 100 news articles; 3 blog posts; no videos, no pictures Puc1-MM (e.g. “Liebherr”: ca. 100 news articles; 3 blog posts; no videos, no pictures Puc1-MM (e.g. “AEG”: ca. 50 news articles; 1 blog post; no videos, no pictures

* the same applies for the US edition (http://www.reuters.com/)

13a	Reuters (German edition)*
URL	http://de.reuters.com
Languages	German
Content	News Stock exchange quotations (not relevant for PUC1;not listed in the sources for PUC2)
Format	html
Keyword search available Query included into URL Paywall restrictions	News: **No** Stock exchange: **Yes** (but, only search for companies' abbreviations, to get the stock exchange quotes and news related to the company) Stock exchange: Yes Example: http://de.reuters.com/investing/stocks/quote?symbol=SIEM.NS&fs=1 No, for articles retrievable from the website.
Useful content	News: Day, date and time (Montag, 14. April 2014, 12:07 Uhr), title, content itself (text) Stock exchange: table with sector, industry, different trade information:  **Airbus Group NV (AIR.PA1) (Paris Stock Exchange)** **sector:** Industrials · **industry:** Aerospace / Defense · As of 7 Apr 2014 53.51 EUR — Price Change -0.49 — Percent Change -0.91% Research a stock: AIR.PA1 GO symbol lookup Last Trade €53.51 — Day's High €53.71 Trade Time 7 Apr 2014 — Day's Low €53.23 Change -0.91% — 52-wk High -- Prev Close €54.00 — 52-wk Low -- Open €53.50 — Beta 0.78 Volume 523,494 — Avg. Vol --
Useful metadata	<META name="description" content="Brüssel (Reuters) - Die EU-Kommission kommt mit ihren Leitlinien für Beihilfen im Energiesektor der europäischen Industrie noch weiter entgegen... <META name="REVISION_DATE" content="Thu Apr 10 06:52:53 UTC 2014"> <META name="DCSext.rCountry" content="DE">
Useful info in the URL	Language: http://de.reuters.com/article/topNews/idDEBEEA3900K20140410
How to extract info?	Crawler (generic, to collect all available articles and to select the required ones in post-crawling phase) Html parsing (boilerplate removal) to obtain the text Metadata extraction (data, keywords, description) URL parsing for: keywords, topics, date, kind of medium Information Extraction for text analysis (Stock exchange tables – not relevant for PUC1!)
Use case coverage / relevance	Info not retrievable, since no Search field available

* the same applies to the Spanish (http://es.reuters.com/) , French (http://fr.reuters.com/) , and all other non-English editions

14	AFP
URL	http://www.afp.com/fr/ http://www.afp.com/de/ http://www.afp.com/es/ http://www.afp.com/en/
Language	French (German, Spanish, English)
Content	News articles, social media posts (Twitter, Facebook) (news articles often have embedded videos)
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes Examples: http://www.afp.com/fr/search/site/%C3%A9nergie/ http://www.afp.com/de/search/site/Energie/ http://www.afp.com/es/search/site/energ%C3%ADa/ http://www.afp.com/en/search/site/energy/ Articles from different APF sections (L'Agence, L'info, Innovation …), Facebook and Twitter posts No (but Facebook posts cannot be accessed by following the hyperlinks; they always land on a blank page)
Useful content	News : title, place, date, content itself (text) Blogs : publication date, Blog name, Blog URL, RSS link, Comment count, Post URL, Post Id
Useful metadata	-
Useful info in the URL	Language: http://www.afp.com/fr/node/1217449 http://www.afp.com/es/node/2266551
How to extract info?	Web crawler (generic or customised) HTML parsing (boilerplate removal) to obtain the text Metadata extraction by html parsing (?) URL parsing (for language) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. "Energy"): French ca. 880 items, German ca. 150 items, English ca. 750 items, Spanish ca. 350 items Puc1-MM (e.g. "Kitchenaid": 0 items, German 0 items, English 0 items, Spanish 0 items Puc1-MM (e.g. "Liebherr": 0 items, German 0 items, English 0 items, Spanish ca. 350 items Puc1-MM (e.g. "AEG": ca. 15 items, German 0 items, English 0 items, Spanish ca. 350 items

15	dpa
URL	http://www.dpa.de/ http://www.dpa.de/English.82.0.html http://www.dpa.de/Espanol.83.0.html
Language	German, English, Spanish
Content	The dpa websites do not contain crawlable news articles or other contents. Dpa service is available against monthly fees.
dpa service	The DPA news agency produces over 800 daily reports from the entire world, in the subjects of politics, business, culture, sport and other news stories. Along with the international offices the twelve regional offices produce reports dealing with German politics, business, culture and sport. The DPA Photo Service provides customers about 350 photos daily.
conditions	DPA customers are provided the service for a monthly fee (fee is dependent on the size of the organisation), additional fees are required for organisations that do not provide content to DPA.

16	agencia EFE
URL	http://www.efe.com/efe/noticias/espana/1 http://www.efe.com/efe/noticias/english/4
Language	Spanish, English
Content	News articles, photos, video, audio
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	English: No (only if registered user) Spanish: Only in the free content area ("contenidos gratuitos"): http://www.efelibredescarga.com/LibreDescarga) No News articles including multimedia content Yes (except of "contenidos gratuitos)
Useful content	Place, date, title, content itself (text); Videos are links to YouTube (there: number of views, number of likes and dislikes, publishing date, number of comments, comments)
Useful metadata	<meta Name="description" content="La deuda del conjunto de las Administraciones Públicas españolas se situó en febrero en 987.945 millones de euros, lo que supone el 96,56 % del PIB, según datos facilitados por el Banco de España." /> <meta Name="keywords" content="espana,deuda,deuda,publica,espanola,alcanza,febrero,Portada" /> <meta property="og:type" content="article" />
Useful info in the URL	Language: http://www.efe.com/efe/noticias/espana/mundo/-/1/4/0 http://www.efe.com/efe/noticias/english/portada/die-huge-fire-chile/4/63/2293209
How to extract info?	Web crawler (generic or customised) HTML parsing (boilerplate removal) to obtain the text Metadata extraction URL parsing (for language) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (e.g. "Energy"): English no search function, Spanish 22 pages with ca. 20 issues on each of them Puc1-MM (e.g. "Kitchenaid": English no search function, Spanish 0 items Puc1-MM (e.g. "Liebherr": English no search function, Spanish ca. 0 items Puc1-MM (e.g. "AEG": English no search function, Spanish ca. 2 items

17	German Federal Ministry for Economic Affairs and Energy
URL	http://www.bmwi.de/ http://www.bmwi.de/EN/root.html http://www.bmwi.de/FR/root.html
Languages	German, English, French
Content	Ministry statements, press releases, videos and audios, photos
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes No News, pictures, press releases, publications, speeches, videos No
Topic search available Keyword included into the URL Search results Paywall restricted	Yes (menu “Topics”) Yes: http://www.bmwi.de/DE/Themen/energie.html http://www.bmwi.de/EN/Topics/energy.html http://www.bmwi.de/FR/Sujets/energie.html News No
Format	Ministry statements as PDF files. Example: http://www.bmwi.de/BMWi/Redaktion/PDF/Stellungnahmen/EEG/agfw,property=pdf,bereich=bmwi2012,sprache=de,rwb=true.pdf Press releases as HTML documents. Example: http://www.bmwi.de/DE/Presse/pressemitteilungen,did=630900.html Mediathek (only in German) embedded into the HTML file. Example: http://www.bmwi.de/DE/Mediathek/videos,did=600094.html
Useful content	Date, title, description (for videos, photos), content itself (text, video)
Useful metadata	<meta name="keywords" content="federal minister, federal ministry, energy, European Union, business, work, environment, economy, energy, technology, policy, investigation, development, company, concern, middle class, promotion of the economy, international trade, high-tech, firm, venture, bank, foreign trade, trade, import, export, fond, Energy policy, Energy, Topics"> <meta property="og:description" content="Economic efficiency, security of supply and environmental compatibility: these are the central aims of German energy policy. In Germany, the Federal Ministry of Economy and Technology has the..." />
Useful info in the URL	Language, Topic: http://www.bmwi.de/FR/Sujets/energie.html
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Metadata extraction (for keywords and description) URL parsing (for language and topic) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	Puc1-J (“Energy”): French ca. 110 items, German ca. 1.600 items, English ca. 230 items

18	Agency for the Cooperation of Energy Regulators
URL	http://www.acer.europa.eu/Pages/ACER.aspx
Language	English
Content	News, press releases, meeting minutes, official documents
Format	Html, PDF, Word, PowerPoint
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.acer.europa.eu/Search/Pages/results.aspx?k=energy single documents No
Query filters included into URL Query filters: - Result type (PDF, Word, PowerPoint) - Site (Acer or any) - Modified date (24 hours, past month, past 6 months, past year, earlier)	Yes Examples: http://www.acer.europa.eu/Search/Pages/results.aspx?k=energy&r=fileextension%3D%22pdf%22 http://www.acer.europa.eu/Search/Pages/results.aspx?k=energy&r=site%3D%22http%3A%2F%2Fwww%2Eacer%2Eeuropa%2Eeu%22
Useful content	News (html): title, date, text itself Other files (pdf, doc …) – files themselves (to be analysed)
Useful metadata	-
Useful info in the URL	Topic: http://www.acer.europa.eu/Gas/Framework%20guidelines_and_network%20codes/Pages/default.aspx http://www.acer.europa.eu/Electricity/Market%20monitoring/Pages/default.aspx
How to extract info?	Web Crawler Format converters (PDF, DOC, PPT to TXT) Information extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) ca. 1.200 matches

19	German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety
URL	http://www.bmub.bund.de/ http://www.bmub.bund.de/en/
Language	German, English
Content	News on climate and energy, speeches (in text format), videos
Format	Html, PDF, flash (video)
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.bmub.bund.de/en/search/?id=1892&no_cache=1&L=1&tx_solr%5Bq%5D=energy&x=0&y=0 websites, press releases, downloads (PDF), files (PDF), videos, events (event information) No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (menu "Topics") Yes: http://www.bmub.bund.de/en/topics/climate-energy/ http://www.bmub.bund.de/themen/klima-energie/ http://www.bmub.bund.de/themen/atomenergie-strahlenschutz/ topic pages with links to articles, publications (can be either ordered or downloaded as PDF) No (except of publications to be bought)
Query filters included into URL Query filters: - Type (pages, press releases, downloads, speeches, files, events, videos) - Time (>1 week; 1week-to-1month, 1-6 months, 6months-1year, more than 1 year) - Legislative period (18th, 17th, 16th, 15th, 14th)	Yes Examples: http://www.bmub.bund.de/en/search/?no_cache=1&tx_solr%5Bq%5D=energy&tx_solr%5Bfilter%5D%5B0%5D=type%253Avideos http://www.bmub.bund.de/en/search/?no_cache=1&tx_solr%5Bq%5D=energy&tx_solr%5Bfilter%5D%5B0%5D=age%253AhalfYear http://www.bmub.bund.de/en/search/?no_cache=1&tx_solr%5Bq%5D=energy&tx_solr%5Bfilter%5D%5B0%5D=archive%253Aperiod16
Useful content	Place, date (Berlin, 10.04.2014), title, text itself Multimedia are often only links to external pages (TV etc.) – thus, the information available depends on the channel
Useful metadata	<meta name="description" content="The water on our planet is in a constant cycle of precipitation and transpiration. It is not a finite resource like oil or gas. It is not possible to use up water. It is merely used and reintroduced into the water cycle." /> <meta name="keywords" content="Federal, Ministry, for, the , Environment, Nature, Conservation, and, Nuclear, Safety, water, drinking, resource, Germany, waste, water cycle, water balance, waste water, save energy, water supply, waste water treatment, avoid, unnecessary, pollution, Drinking Water Ordinance, Trinkwasserverordnung, Federal, Ministry, of, Health" /> <meta name="language" content="en" /> Video: <meta name="description" content="Video - Erneuerbar Mobil -" />(??)
Useful info in the URL	Topics, Language: http://www.bmub.bund.de/en/topics/water-waste-soil/water-management/drinking-water/
How to extract info?	Web crawler (generic or customised) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Metadata extraction (for keywords, description, language) URL parsing (for language and topic) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	PUC1-J ("energy") English ca. 600 matches; German ca. 3.200 matches

20	French Ministry for Ecology, Development and Energy
URL	http://www.developpement-durable.gouv.fr/
Language	French
Content	News articles, press releases, different publications
Format	Html, PDF
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.developpement-durable.gouv.fr/spip.php?page=search&recherche=%C3%A9nergie Links to single matches (articles, reports …) No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (menu with different topics) Yes: http://www.developpement-durable.gouv.fr/-Energie-Air-et-Climat-.html http://www.developpement-durable.gouv.fr/-Developpement-durable-.html Links to articles and publications No
Query filters included into URL Query filters: - Type of document (news; calls for proposals, press releases, speeches, press archives, forms, publications) - Time period (year) - Topic	Yes (but only the time period) Examples: http://www.developpement-durable.gouv.fr/spip.php?page=search&recherche=énergie&doc=&annee=2014&theme=
Useful content	Html: date, title, text itself PDF: each of them to be looked at separately before the information extraction
Useful metadata	<meta name="description" content="La Convention internationale des énergies marines renouvelables Thetis EMR se déroule les 9 et 10 avril à Cherbourg. Elle rassemble les acteurs des énergies marines renouvelables pour deux jours de (...)" />
Useful info in the URL	Language, Topic
How to extract info?	Web crawler (generic; all items should be retrieved) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Metadata extraction (for description) URL parsing (for language and topic) Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“énergie”) ca. 3.200 matches

21	European Commission: Energy
URL	http://ec.europa.eu/energy/index_en.htm http://ec.europa.eu/energy/index_de.htm http://ec.europa.eu/energy/index_fr.htm
Language	German, French, English
Content	Articles, Multimedia (often only as links to external pages, most of them in English only) Videos mostly only in English
Format	Html, PDF Many files available only as ZIP archives
Keyword search available Search results Paywall restricted	Yes Google search results on the website itself (e.g. for French: énergie site:ec.europa.eu/energy) No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (from the topic menu) Yes http://ec.europa.eu/energy/renewables/index_en.htm http://ec.europa.eu/energy/nuclear/index_en.htm news, press releases, newsletters, events … No
Useful content	all
Useful metadata	<meta http-equiv="Content-Language" content="en"> <meta name="Keywords" content="European Union, European Commission,EU,energy,nuclear,Euratom,safety,security,ENSREG,enef,waste management, radiation protection, decommissioning, nuclear energy forum"> <meta name="Description" content="European Commission – Nuclear energy in Europe"> <meta name="Date" content="02/12/2009"> <meta name = "medium" content = "video" />
Useful info in the URL	Language, Topic
How to extract info?	Web crawler (generic; all items should be retrieved) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Metadata extraction (for keywords, description, language, date, kind of medium) URL parsing (for language and topic) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	PUC1-J (“energy”) French ca. 2.800 matches, English ca. 12.700 matches, German ca. 2.800 matches

22	Spanish ministry of industry, energy and tourism
URL	http://www.minetur.gob.es/energia/es-ES/Paginas/index.aspx http://www.minetur.gob.es/en-US/Paginas/index.aspx
Language	Spanish, English
Content	Papers, reports, statistics, documentation, talks and speeches etc. (many links on the English site actually go to the Spanish documents)
Format	Html, PDF, mp3
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://buscador.060.es/search?client=mityc&proxystylesheet=mityc&ie=utf-8&oe=utf-8&filter=1&lg_i=en&output=xml_no_dtd&numgm=5&site=MIT_ENE&q=energy links to articles and PDF files No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (from the menu) Yes: http://www.minetur.gob.es/energia/es-ES/Paginas/index.aspx http://www.minetur.gob.es/energia/en-us/Paginas/Index.aspx “Highlighted Services” (reports, registers, statistics…), Press Office News No
Useful content	Contents themselves (text, audio) to be analysed
Useful metadata	<meta name="keywords" content="El, Ministerio, Press, Office, Press, Releases, Notas, Prensa, 2014, El, Ministerio, Industria, desmiente, que, renuncie, aplicar, precio, electricidad, horas"/>
Useful info in the URL	Language, Topic, Date http://www.minetur.gob.es/en-US/GabinetePrensa/NotasPrensa/2014/Paginas/20140314-luz-tarifa-precio-energia.aspx
How to extract info?	Web crawler (generic; all items should be retrieved) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Metadata extraction (for keywords) URL parsing (for language, topic, date) Information Extraction for text analysis ASR for audio Image and video processing + ASR for video
Use case coverage / relevance	PUC1-J (“energy”) English ca. 1.800 matches, Spanish ca. 15.600 matches

23	UN Energy knowledge network
URL	http://www.un-energy.org/
Language	English
Content	News, publications, newsletters Tools (internal and external links), e.g: - Measuring energy access: http://www.un-energy.org/measuring-energy-access - Multidimensional Energy Poverty Index: http://www.un-energy.org/sites/default/files/share/une/stats/content.swf
Format	Html, PDF, shockwave flash,
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes http://www.un-energy.org/search?cx=018133420830908500077%3Agnzvtpafeeo&cof=FORID%3A11&query=energy&as_sitesearch=www.un-energy.org&form_build_id=form-488a2526547f4fb2a961fde23748e73c&form_id=google_cse_searchbox_form&sitesearch=&sa.x=15&sa.y=10 news, PDF files, newsletters, tools No
Query filters included into URL Query filters: - Source (UN-Energy member sites, or items from "this knowledge network")	Yes http://www.un-energy.org/search?cx=018133420830908500077%3Agnzvtpafeeo&cof=FORID%3A11&query=energy&as_sitesearch=www.un-energy.org&form_build_id=form-ef443f44890f520d8368688137483bd0&form_id=google_cse_results_searchbox_form&siteurl=http%3A%2F%2Fwww.un-energy.org%2F&sa.x=20&sa.y=6
Useful content	Contents themselves (text, video) Tools: tables, graphs, amps - cannot be exported
Useful metadata	-
Useful info in the URL	-
How to extract info?	Web crawler (generic; all items should be retrieved) HTML parsing (boilerplate removal) to obtain the text Information Extraction for text analysis (not feasible to extract info by querying and using the interactive "tools")
Use case coverage / relevance	PUC1-J ("energy") English ca. 1.260 matches from the knowledge network

24	International association for energy economics
URL	http://www.iaee.org/en/
Language	English (for other offered languages, German, French and Spanish, only a part of the GUI is localised, but the contents are all in English)
Content	Publications (to be bought! Many of them cost $0, but nevertheless they are accessible only through the shopping cart) Abstracts and keywords are freely available Other freely available contents are more administrative ones (About, Membership, Students, Calendar of events…), but also some papers in PDF format. Some areas (search the membership directory etc.) require log-in. Newsletters are also accessible only through an input form, requiring the selection of the time period, author etc…)
Format	Html, PDF, ebook (Publications can only be reached through an *.aspx (Active Server Page Extended File) pop-up window)
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes No Links to the single PDF and *.aspx files Yes (Full publication papers) No (abstracts, different students' papers, events etc.)
Useful content	All
Useful metadata	<meta name="description" content="Energy association dealing with policy and economics of oil, natural gas, electricity restructuring, transportation, exploration, energy conferences, environmental, alternative fuels, and OPEC studies"> <meta NAME="Keywords" Content="energy economics, energy conferences, energy association, energy journal, alternative energy, Energy, Oil, Gas, Electricity, Renewable energy, Renewables, Fossil fuels, Nuclear energy, Coal, Gasoline, Diesel, Wind power, Wind energy, Solar energy, Bio fuels, Energy policy, climate, trade, environment, security, supply, poverty, GHG emission, Wall Street, Finance"> (but, they are the same for all pages, independent of the current text)
Useful info in the URL	Language http://www.iaee.org/en/publications/fullnewsletter.aspx?id=29
How to extract info?	Web crawler (for all freely available issues) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Information Extraction for text analysis
Use case coverage / relevance	PUC1-J ("energy") English ca. 13.200 items (but from the Browser only the first 10 pages with search results accessible)

25	Gesellschaft für Energiewissenschaft und Energiepolitik e. V.
URL	http://www.gee.de/ http://www.gee.de/?lang=en
Language	German and English (but the pages are very mixed; even if German selected, the site contains English texts, and vice versa)
Content	Very few news, event calendar, a small table with references to publications; calls for papers (not really rich content)
Format	html
Keyword search available Paywall restricted	No No
Useful content	There is no really useful content (?)
Useful metadata	<meta http-equiv="Content-Language" content="de-DE" />
Useful info in the URL	Language
How to extract info?	Web crawler HTML parsing
Use case coverage / relevance	PUC1-J ("energy") very small amount of data; seems not to be really relevant for the use case

26	Verein für ökologisch-solidarische Energie- & Weltwirtschaft e.V.
URL	http://power-shift.de/
Language	German (although there exist an English version of the site (http://power-shift.de/?lang=en), both the GUI and most of the content are in German only)
Content	Articles, reports, studies, videos
Format	Html, PDF
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://power-shift.de/?s=energie Links to articles, reports, news … (with embedded videos) No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (menu “Energiepolitik”) No (only as cat=12): http://power-shift.de/?cat=12 The same issues as in the keyword search No
Useful content	Html issues: title, date, content itself (text, video) PDF files are mostly only copies of the html, to be alternatively downloaded
Useful metadata	<meta name="description" content="PowerShift ** Corporate Europe Observatory ** Attac Frankreich ** Friends of the Earth Europe, ** Sierra Club ** Blue Planet Project ** Transnational Insti..." /> <meta name="keywords" content="Biomasse, Energiepolitik, Handelspolitik, Internationale Investitionspolitik, Rohstoffpolitik, Rohstoffstrategie" /> <meta http-equiv="Content-Language" content="de-DE" />
Useful info in the URL	Language
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) German: 42 matches

27	Climate Action Network Europe
URL	http://www.caneurope.org/
Language	English
Content	News, press releases, publications, letters, public consultations; policies
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.caneurope.org/component/finder/search?q=energy&Itemid=368 Links to single issues No
Topic search available Topic included into the URL Search results Paywall restricted	Yes (menu “Policies”) Yes: http://www.caneurope.org/policywork/issues/eu-ets (emission trading scheme) http://www.caneurope.org/policywork/issues/energy-saving http://www.caneurope.org/policywork/issues/eu-energy-and-climate-policy http://www.caneurope.org/policywork/issues/renewables http://www.caneurope.org/policywork/euintlissues/un-climate-negotiations http://www.caneurope.org/policywork/euintlissues/climate-finance http://www.caneurope.org/policywork/euintlissues/development The same issues as in the keyword search No
Query filters included into URL Query filters: - Time period - Category (all categories from the website) - Country (all, or only Belgium) - Type (articles, web links …)	Yes (only time is “human-readable”): http://www.caneurope.org/component/finder/search?q=energy&w1=before&d1=2014-04-01&w2=after&d2=2014-03-26&t%5B%5D=49&t%5B%5D=29&t%5B%5D=9
Useful content	Text, Images, graphs Links to PDF files contain some meta-information: date, file size, and some other details in an additional pop-up-window (abbreviation for author etc.)
Useful metadata	<meta name="keywords" content="climate change, EU, Brussels, Climate and energy policy, ETS, Emissions Trading Scheme, carbon markets, Energy Savings, Energy Efficiency Directive, Coal, renewable energy, European climate and energy package, UNFCCC, climate finance, adaptation fund, climate news, climate, climate network, climate network europe, climate action, energy for the planet, wendel trio, IPCC, climate and development" />
Useful info in the URL	Topic
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) English: 157 matches

28	Öko-Institut e.V.
URL	http://www.oeko.de/ http://www.oeko.de/en/
Language	German, English
Content	Publications, projects, e-papers
Format	Html PDF publications (can be downloaded) PDF e-papers (only to be displayed in the browser with a servlet; not downloadable)
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes No Divided into: - publications (html, PDF) - projects (either only project names, or links to executive summaries) - e-papers (links to short descriptions, title, editorial etc. and to online searchable, but not downloadable PDF files) - links to internal or external websites There is a “Subscribe” area (http://www.oeko.de/en/publications/e-paper/subscribe/) for e-papers
Topic search available Topic included into the URL Search results Paywall restricted	Yes (menu “research/issues” on the bottom of the page): Yes: http://www.oeko.de/en/research-consultancy/issues/energy-and-climate/ http://www.oeko.de/en/research-consultancy/issues/nuclear-engineering-and-facility-safety/ etc. articles on the topic No
Useful content	all
Useful metadata	<meta name="language" content="en"> <meta name="date" content="2014-03-31" />
Useful info in the URL	Language, topic http://www.oeko.de/en/research-consultancy/issues/sustainable-consumption/
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text PDF to TXT (for pdf files which can be downloaded) Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) ca. 45 e-papers, ca. 200 links to websites, many links to projects and publications (no number of matches available)

29	Bundesverband Erneuerbare Energien (BEE)
URL	http://www.bee-ev.de/
Language	German (there is also an English version of the site available (http://www.bee-ev.de/BEE/English.php), but it contains only the abstract of the Home-page)
Content	Articles, press releases, publications
Format	Html, PDF
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes http://bee-ev.de/Suche/index.php?query=Energie&search=1 Links to the single articles (all of them BEE internal) No
Topic search available Topic included into the URL Search results Paywall restricted	There are menus “Energiepolitik” and “Energieversorgung” Yes: http://bee-ev.de/Energiepolitik/Energiepolitik.php http://bee-ev.de/Energieversorgung/Energievorsorgung.php Web pages on the topic with further links to German, European and international issues. No
Useful content	Date, place (*Berlin, 13. April 2014:*), author (Jens Tartler Pressesprecher), all contents (text)
Useful metadata	-
Useful info in the URL	Language, Topic
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text PDF to TXT conversion Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) German ca. 390 matches (but many of them lead to longer lists of PDF files)

30	Regional Center for Energy Policy Research
URL	http://www.rekk.eu/index.php?lang=en
Language	English Original site is in Hungarian; thus, many pages and files (pdf etc.) are not translated into English. They often have a remark: “There are no translations available”, but not always.
Content	General information about the centre, its research and teaching profile, events etc. Publications (PDF) – many of them in Hungarian only Reports (PDF) Books (PDF)
Format	Html, PDF
Useful content	In PDF files
Useful metadata	<meta name="title" content="Forecasting Hungarian gas consumption between 2012-2015" /> <meta name="author" content="Decsák Zsuzsa" />
Useful info in the URL	Language: http://www.rekk.eu/index.php?lang=en
How to extract info?	Web crawler (generic, since all available items are required) HTML parsing (boilerplate removal) to obtain the text PDF to TXT conversion Metadata and URL parsing for author and language Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) ca. 80 matches

31	Central European University - Center for Climate Change and Sustainable Energy Policy
URL	http://3csep.ceu.hu/
Language	mostly English but, there are also some Hungarian pieces
Content	General information about the university, its research and teaching profile, events etc. News Publications (PDF) Links to projects' descriptions
Format	Html, PDF
Useful content	In PDF files
Useful metadata	-
Useful info in the URL	-
How to extract info?	Web crawler (generic, since all available items are required) HTML parsing (boilerplate removal) to obtain the text PDF to TXT conversion Information Extraction for text analysis
Use case coverage / relevance	PUC1-J ("energy") when using the search function: 10 pages of results (each of them ca. 10 items)

32	Central European University – Energy Policy Research Group
URL	http://energy.ceu.hu/
Language	English
Content	News, events Publications (only abstracts, with links to book shops)
Format	html
Useful content	News: date (27/02/2014), title, author, text itself Very few really relevant info about publications: title, year, author, abstract
Useful metadata	-
Useful info in the URL	Date: http://energy.ceu.hu/news/2014-03-31/aleh-cherp-presented-research-at-the-annual-conference-of-the-asia-and-the-pacific-p
How to extract info?	Web crawler (generic, since all available items are required) HTML parsing (boilerplate removal) to obtain the text URL parsing for date Information Extraction for text analysis
Use case coverage / relevance	PUC1-J ("energy") when using the search function: 9 pages of results (each of them ca. 10 items)

33	European Energy Research Alliance
URL	http://www.eera-set.eu/
Language	English
Content	General information about the organisation (members, events, programme) Press releases, Published articles, interviews (PDF) Newsletters (PDF)
Format	Html Word documents PDF
Keyword search available	No
Topic search available	No
Useful content	In PDF files
Useful metadata	-
Useful info in the URL	-
How to extract info?	Web crawler (generic; selection according to the keyword “energy” could be done in post-crawling stage) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Information Extraction for text analysis
Use case coverage / relevance	PUC1-J (“energy”) information not retrievable, since no search function

34	Florence School of Regulation - Energy
URL	http://fsr.eui.eu/FlorenceSchoolofRegulation/Energy/Index.aspx
Language	English
Content	General information about the school (donors, governance, activities) – seems to contain less interesting content Publications (video&audio, papers, presentations, books, reports …)
Format	Html PDF (publications) Flash (video, webinar recordings) Books (only abstracts; + links to full texts for download via EUI repository); mostly PDF
Keyword search available	No
Topic search available Search restrictions: - Area (Energy, Transport, Climate) - Author / editor - Type (Video&Audio, Workshop paper, webinar output, working paper, research report, presentation, policy brief, journal article, journal, booklet, book chapter, book) - Time (year)	No But, there is a search mask for publications. For all on "Energy": http://fsr.eui.eu/Publications.aspx?FSRPublicationsListing1_Areas=0%2f304%2f1847%2f1848%2f1857
Useful content:	In PDF and multimedia files Videos are mostly on YouTube;
Useful metadata	-
Useful info in the URL	-
How to extract info?	Web crawler (generic; selection according to the keyword "energy" could be done in post-crawling stage) HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Flash files (?) ASR for video Image and video analysis + ASR for video Information Extraction for text analysis
Use case coverage / relevance	PUC1-J ("energy") 2014 publications

35	private website Prof. Jarras
URL	http://www.jarass.com/home/index.php/DE/energie
Languages	German There is an English version of the website available (http://www.jarass.com/home/index.php/en/energie), but it does not contain useful English information; only a few parts of the GUI have been translated
Content	Books (only summaries and tables of content + links to book stores) scientific papers lectures
Format	html PDF (publications)
Useful content	In PDF files
Useful metadata	-
Useful info in the URL	Language, topic: http://www.jarass.com/home/index.php/DE/energie
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text PDF to TXT converter Information Extraction for text analysis
Use case coverage / relevance	PUC1-J ("energy") Ca. 10 books, 10 pages with 10-12 papers each, ca. 20 lectures

36	Brigitte
URL	http://www.brigitte.de/
Language	German
Content	Articles on fashion, nutrition, lifestyle, shopping, health, culture and travel Forum (original posts, answers, user names (aliases), date and time of the post)
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.brigitte.de/suche?query=kitchenaid Articles (including image galleries) and forum threads No
Useful content	Articles: Titel, author (Text: Sonja Niemann), the content itself (text), comments, number of comments Forum: date, time, author (name, date of his/her registration,and number of posts), topic, post, answers
Useful metadata	<meta name="keywords" content="Küchenmaschine, normale, küchenmaschinen, nachdenken, schaffen, profi, bereich, vollkornteig, [editiert], teigmaschine, ausschließlich, brotteig, maschine, kneten, brauchst, kleinen, anschaffung, würde, küchenmaschine" />
Useful info in the URL	Language, topic: http://bfriends.brigitte.de/foren/haushalt-tipps-und-tricks/372956-kuechenmaschine.html
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text (HTML parsing for the metadata from the forum?) Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) news articles and shop: 4 matches; forum: 76 posts PUC1-MM (“Liebherr”) news articles and shop: 0 matches; forum: 46 posts PUC1-MM (“AEG”) news articles and shop: 5 matches; forum: 236 posts

37	Bunte
URL	http://www.bunte.de/
Language	German
Content	Articles about lifestyle, beauty, fashion, celebrities
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.bunte.de/search/results?search_api_views_fulltext=kitchenaid&=Anwenden&op=Suche list of articles No
Topic search available	Yes (but, no topics relevant for our use case)
Useful content	Text, images – to be analysed
Useful metadata	<meta name="dcterms.title" content="Ferrari: Autokonzern baut Luxushotel in Themenpark" /> <meta name="description" content="In zwei Jahren eröffnet ein Ferrari-Freizeitpark in Spanien. Der Sportwagenhersteller will dort auch das weltweit erste Ferrari-Hotel eröffnen, das die Herzen alle Rennsportfans höher schlagen lassen soll." /> <meta name="dcterms.creator" content="natalie.schuckardt" /> <meta name="date" content="2014-04-10T16:21:25+02:00" /> <meta name="dcterms.language" content="de" />
Useful info in the URL	Language
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction for description, author, date and language Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 0 matches PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 0 matches

38	Elle
URL	http://www.elle.de http://www.elle.fr/ http://www.elleuk.com/ http://www.elle.es/
Language	German, French, English, Spanish
Content	Articles about fashion, beauty, lifestyle, travel Blogs
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.elle.de/suche-72638.html?search_search=kitchenaid&=Suchen http://www.elle.fr/recherche/recherche-globale?searchText=kitchenaid&getsection=all http://www.elleuk.com/content/search?SearchText=kitchenaid&SearchButton= http://www.elle.es/content/search?SearchText=kitchenaid articles, shop items No
Useful content	
Useful metadata	
Useful info in the URL	
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) German: 0 matches , French: 10 matches , English: 0 matches, Spanish: 0 matches PUC1-MM (“Liebherr”) German: 0 matches , French: 2 matches , English: 0 matches, Spanish: 0 matches PUC1-MM (“AEG”) German: 0 matches , French: 12 matches (all of them refer to the music promoter AEG) , English: 0 matches, Spanish: 0 matches

39	Elle Decoration
URL	http://www.elle.de/elle-decoration-73332.html http://www.elledecoration.co.uk/ http://www.elle.fr/Deco http://www.elle.es/elledeco
Language	German, English, French, Spanish
Content	Articles on home design
Format	html
Keyword search available Search results Paywall restricted	Yes (German, French, Spanish) No (English) The same as for Elle (Nr. 38) No (German, French, Spanish) Not clear (English): there are no entries freely available; the title page gives info on subscription
Useful content	
Useful metadata	
Useful info in the URL	
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text URL parsing for language Information Extraction for text analysis
Use case coverage / relevance	Same as in 38 (except of English, which does not offer any search functions)

40	Glamour
URL	http://www.glamour.de/
Language	German
Content	Articles about fashion, beauty, celebrities
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.glamour.de/content/search/?SearchText=kitchenaid articles, shop issues, prize competitions No
Useful content	
Useful metadata	
Useful info in the URL	
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text URL parsing for language Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) only 1 prize competition PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 2 matches (only prize competitions)

41	inStyle
URL	http://www.instyle.de/
Language	German
Content	Articles about fashion and shopping, videos, blogs
Format	
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.instyle.de/search/node/kitchenaid articles, shopping recommendations No
Useful content	
Useful metadata	
Useful info in the URL	
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text URL parsing for language Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 1 shopping recommendation PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 0 matches

42	Jolie
URL	http://www.jolie.de/
Languages	German
Content	Fashion magazine: fashion, beauty, lifestyle
Format	Html, flash
Keyword search available Query included into URL Search Results Paywall restriction	Yes Example: http://www.jolie.de/suche/index.html?q=kitchenaid Articles, Image galleries, Videos Forum (posts, answers, time and date, user name) No
Useful content	
Useful metadata	<meta property="og:type" content="video" <meta name="language" content="deutsch">
Ueful info in the URL	Language, content type: http://www.jolie.de/artikel/schlimme-weihnachtsgeschenke-1678396.html http://www.jolie.de/forum/hochzeit/6431-hochzeitsgeschenk-6.html http://www.jolie.de/bildergalerien/william-kate-hochzeitsgeschenke-1370428.html
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction or URL parsing (for content type, language) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 16 matches PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 10 matches

43	GQ
URL	http://www.gq-magazin.de/
Language	German
Content	Fashion magazine for men; automobile, technique, nutrition, fitness, travel
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.gq-magazin.de/content/search?SearchText=kitchenaid articles, images, videos, star portraits No
Topic search available Topic included into the URL Search results Paywall restricted	Yes Yes: http://www.gq-magazin.de/tags/b/bmw articles, images, videos, star portraits No
Useful content	
Useful metadata	<meta name="news_keywords" content="Louis Vuitton,BMW,Taschen" /> <meta name="description" content="Wird passend gemacht. Louis Vuitton präsentiert Gepäck für den Elektrosportler BMW i8 <meta property="og:type" content="article" /> <meta property="og:type" content="video" />
Useful info in the URL	Language, topic, content type http://www.gq-magazin.de/auto-technik/videos/bmw-r1200-gs
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction or URL parsing (for content type, language) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 0 matches PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 0 matches

44	Architectural Digest
URL	http://www.ad-magazin.de/ http://www.admagazine.fr/ http://www.revistaad.es/ http://www.architecturaldigest.com/
Language	German, French, Spanish, English
Content	Articles about design, architecture, art, lifestyle, home The main content type is image
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.ad-magazin.de/?s=kitchenaid&submit=Suche http://www.admagazine.fr/search/index?q=kitchenaid http://www.architecturaldigest.com/search?query=kitchenaid&sort=score+desc (no in the Spanish edition) Articles with embedded images and videos No
Useful content	
Useful metadata	-
Useful info in the URL	Language, topic: http://www.admagazine.fr/architecture
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text URL parsing (for topic, language) Information Extraction for text analysis Image and video processing + ASR for video
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) German: 0 matches , French: 0 matches , English: 9 matches PUC1-MM (“Liebherr”) German: 2 matches , French: 0 matches , English: 27 matches (according to the search list; but only one of them contained “Liebherr”) PUC1-MM (“AEG”) German: 0 matches , French: 0 matches, English: 2 matches

45	Elektrojournal
URL	http://www.elektrojournal.at/
Language	German
Content	Magazine on electronic devices and products; articles, newsletters
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.elektrojournal.at/_suchergebnis-119583.html?w=aeg Articles from the journal, readers' comments No
Useful content	
Useful metadata	<meta name="keywords" content=", elektro, Elektro, Elektrotechnik, elektrojournal, technik, Information, Nachrichten, Magazin, Bericht, Reportage, Buch, Katalog, Verzeichnis, Links, Produkt" /> <meta name="description" content="Liebherr setzt zum Energiespar-Überholmanöver an: Neue Geräte mit A+++ - elektrojournal.at - österreichs nachrichtendienst für die elektrobranche" /> <meta name="date" content="2011-09-07" /> <meta name="Content-Language" content="de" />
Useful info in the URL	-
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction (for keywords, description, date, language) Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM ("Kitchenaid") 0 matches PUC1-MM ("Liebherr") 126 matches PUC1-MM ("AEG") 707 matches

46	Küchenmagazin
URL	http://www.kueche-co.de/kuechenmagazin/
Language	German
Content	Descriptions and advice-giving articles on kitchen and kitchenware Many images
Format	html
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.kueche-co.de/suchergebnis/?searchquery=aeg&searchmax=100#crawlersearch divided into: Kitchen, Studio, Kitchenware No
Useful content	
Useful metadata	<meta name="description" content=" Mit den neuen Einbauherden und Backöfen von AEG wird das Backen und Garen jetzt noch einfacher. Preiswerte und günstige AEG Elektrogeräte bei Küche&Co. ">
Useful info in the URL	Topic: http://www.kueche-co.de/kuechenausstattung/aeg/
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction (description) Information Extraction for text analysis Image processing for images
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 0 matches PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 136 matches in

47	Elektronik JOURNAL
URL	http://www.elektronik-journal.de/
Language	German
Content	Magazine on power electronics, medical electronics equipment, electromechanics etc.
Format	Html website with downloadable single issues of the journal in **PDF** format
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.elektronik-journal.de/?s=aeg List of PDF issues, in which the search keyword occurs No
Useful content	
Useful metadata	-
Useful info in the URL	Language, Date of the issue, main Topic: http://www.elektronik-journal.de/2013/11/06/elektromechanik-3/
How to extract info?	Web crawler - Save only PDF files PDF to txt conversion URL parsing (language, date, topic) Information Extraction for text analysis
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) 0 matches PUC1-MM (“Liebherr”) 0 matches PUC1-MM (“AEG”) 0 matches

48	APPLIANCIST
URL	http://www.appliancist.com/
Language	English (but links to the products can lead to articles in other languages, e.g. German)
Content	online magazine about home, kitchen and bathroom appliances
Format	Html Many photos included
Keyword search available Keyword included into the URL Search results Paywall restricted	Yes Yes: http://www.google.com/cse?cx=partner-pub-5274853571753910%3Aj8zra-e3yxy&ie=ISO-8859-1&q=kitchenaid&sa=Search&siteurl=www.appliancist.com%2F&ref=www.appliancist.com%2Fsteam_ovens%2Fsteam-oven-kitchenaid-kosp6610.html&ss=2401j1258753j10#gsc.tab=0&gsc.q=kitchenaid&gsc.page=1 links to internal (appliancist) and external websites with product descriptions, prizes, advertisement etc.
Topic search available Topic included into the URL Search results Paywall restricted	Yes (list of topics in a side menu) Yes: http://www.appliancist.com/refrigerators/ articles, with images and product desriptions No
Useful content	
Useful metadata	<meta name="description" content="AEG Electrolux UltraCaptic Compact & Go vacuum cleaner" />
Useful info in the URL	Topic
How to extract info?	Web crawler HTML parsing (boilerplate removal) to obtain the text Metadata extraction (description) URL parsing for topic Information Extraction for text analysis Image processing for images
Use case coverage / relevance	PUC1-MM (“Kitchenaid”) ca. 330 matches PUC1-MM (“Liebherr”) ca. 200 matches PUC1-MM (“AEG”) ca. 630 matches

49a	Spanish Office of Economy and Competitiveness
URL	http://www.oficinascomerciales.es/icex/cda/controller/pageOfecomes/0,5310,5280449_5296122_5296234_0_DE,00.html
Languages	Spanisch
Content	Information about Germany General data (Location, size, climate, demography, society, history)
Format	Html PDF (if following the links under “related documents”)
How to extract info?	Web Crawler: - store only the landing page - and the corresponding PDF documents by following the links under “Documentación relacionada” HTML parsing PDF to TXT conversion Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) 1 webpage + 1 related PDF document

49b	Spanish Office of Economy and Competitiveness
URL	http://www.oficinascomerciales.es/icex/cda/controller/pageOfecomes/0,5310,5280449_5296130_5296234_0_DE,00.html
Languages	Spanisch
Content	Information about Germany Practical Information about how to access the market, visas, city websites, bank holidays, banks,...
Format	Html PDF (if following the links under “related documents”)
How to extract info?	Web Crawler: - store only the landing page - and the corresponding PDF documents by following the links under “Documentación relacionada” HTML parsing PDF to TXT conversion Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) 1 webpage + 4 related PDF documents

50	Central Intelligence Agency – The world factbook
URL	https://www.cia.gov/library/publications/the-world-factbook/geos/gm.html
Language	English
Content	Economic overview for Germany: info about sectors, GDP, unemployment rate, labor force, imports, exports, debt, exchange rates,…
Format	- Html (a tree-like overview, which can be expended, or collapsed one by one, or all at once) - (or alternatively, the complete information can be exported as a single PDF file: https://www.cia.gov/library/publications/the-world-factbook/geos/print/country/countrypdf_gm.pdf)
Useful metadata	-
Useful info in the URL	-
How to extract info?	Web crawling (only the landing page; or alternatively the PDF file) Html parsing PDF to TXT conversion (alternatively) Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) : this one page

51	Trading Economics
URL	http://www.tradingeconomics.com/germany/indicators
Languages	English
Content	Economic indicators, exchange rates, stock market indexes, governmental bond yields and commodity prices for more than 200 countries
Format	html
Keyword search available Query included into URL Search results Paywall restricted	Yes Yes Example: http://www.tradingeconomics.com/search.aspx?q=germany&sa=Search&cx=partner-pub-3400948010513654:2035820411&cof=FORID:10&ie=UTF-8 Tables Charts in different presentation forms (chart, line, column, area, candle, bar) **Yes (API available)**
Useful content	Economic indicators (for Germany, as required for PUC2): Markets, GDP, Labour, Prices, Money, Trade, Government, Business, Consumer, Taxes & Housing
Useful metadata	Keywords: <meta id="metaKeyword" name="keywords" content="Germany GDP per capita PPP, Chart Graph, Data, Historical Data" />
Useful info in the URL	the language, the country name that is subject of our query, and the indicator we are interesting in: http://es.tradingeconomics.com/italy/wages
How to extract info?	API (html, csv, json) (against payment on a monthly rate of 399 $) **TRADING ECONOMICS API** - DIRECT ACCESS to 300.000 INDICATORS The Trading Economics API provides direct access to our data from your favourite application or custom software. Providing several request methods to query our databases, it is the best way to export data in HTML, CSV or JSON formats. You can try the sample queries displayed on this page without registering. These examples use a guest account which is limited in scope to a few indicators and a few rows of output. To remove the examples limitation, please use your own username and password instead of the guest account. The full API feature is only available for registered users with a professional plan which paid in advance for at least a quarter. Please register or upgrade your account if you would like to gain access to our API. And please email us at contact@tradingeconomics.com if you have any questions regarding the API usage or pricing.
Use case coverage / relevance	PUC2 (“yoghurt”) - all indicators for a certain country (e.g. Germany)

52	German Missions in the United States
URL	http://www.germany.info/Vertretung/usa/en/02__GIC/GIC/00/__Home.html
Language	English
Content	Information about Germany: holidays, traditions, facts, visa, passport, legal information
Format	html
Useful content	
Useful metadata	<meta name="keywords" content="Foreign policy, Federal Foreign Office, Germany" />
Useful info in the URL	Language, topic http://www.germany.info/Vertretung/usa/en/05__Legal/02__Directory__Services/06__Customs/__Customs.html
How to extract info?	Web crawling (generic crawler; all pages contain potentially relevant information) Html parsing Metadata extraction for keywords URL parsing for language and Topic Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) all pages

53	German Center of Information for Spain and Latin America
URL	http://www.alemaniaparati.diplo.de/Vertretung/mexiko-dz/es/03-PoliticaExterior/Lazos/0-LazosEspaniaYLatinoamerica.html
Language	Spanish
Content	General information about Germany and its foreign policy in Spain and Latin America
Format	html
Useful content	
Useful metadata	<meta name="keywords" content="Política exterior, Ministerio Federal de Relaciones Exteriores, Alemania" /> <meta name="description" content="El Viceministro de Asuntos Exteriores alemán, Stephan Steinlein, y el Embajador de Chile en Alemania, Jorge O'Ryan Schütz, firmaron el jueves 20 de febrero de 2014 un convenio sobre un Programa de Vacaciones y Trabajo ("Working holiday Program")." />
Useful info in the URL	Language, topic http://www.alemaniaparati.diplo.de/Vertretung/mexiko-dz/es/03-PoliticaExterior/Lazos/WHPChile.html
How to extract info?	Web crawling (generic crawler; all pages contain potentially relevant information) Html parsing Metadata extraction for keywords and description URL parsing for language and Topic Information extraction for text analysis
Use case coverage / relevance	PUC2 ("yoghurt") all pages

54	Spanish Ministry of Foreign Affairs and Cooperation
URL	http://www.exteriores.gob.es/Portal/es/SalaDePrensa/Paginas/FichasPais.aspx
Language	Spanish
Content	Information about Germany and bilateral relations between Spain and Germany
Useful info	
Format	Html PDF
Useful metadata	
Useful info in the URL	
How to extract info?	Web crawler - everything under "Alemania"; recommendable seed URLs: http://www.exteriores.gob.es/Documents/FichasPais/ALEMANIA_FICHA%20PAIS.pdf http://www.exteriores.gob.es/Embajadas/Berlin/es/Paginas/inicio.aspx http://www.exteriores.gob.es/Portal/es/ServiciosAlCiudadano/SiViajasAlExtranjero/Paginas/DetalleRecomendacion.aspx?IdP=4 http://www.exteriores.gob.es/Consulados/Dusseldorf/es/Paginas/inicio.aspx http://www.exteriores.gob.es/Consulados/Francfort/es/Paginas/inicio.aspx http://www.exteriores.gob.es/Consulados/Hamburgo/es/Paginas/inicio.aspx http://www.exteriores.gob.es/Consulados/Munich/es/Paginas/inicio.aspx http://www.exteriores.gob.es/Consulados/Stuttgart/es/Paginas/inicio.aspx html parser PDF to TXT conversion Information extraction for text analysis
Use case coverage / relevance	PUC2 ("yoghurt") - everything about Germany (8 seed URL sites)

55	EUROSTAT
URL	http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database
Languages	English, German, French
Content	Databases with statistics on economy, finance, industry, trade, population, social conditions etc. in European countries and regions
Format	Searchable database
Where is the important information	In a data navigation tree; to be reached by browsing the tree, or by a search function
How to extract the info?	One can download individual datasets or the complete database by using the bulk download facility. Individual data sets can be downloaded from each terminal node in the navigation tree in one of the following formats :Excel, csv, HTML, PC-Axis, SPSS, TSV, PDF On the bulk download you will find: • all information **updated twice a day,** at 11:00 and 23:00, • the **datasets in tsv (tab separated values), dft and sdmx format**, which can be easily used to import the data in a tool of your choice, • **guidelines** on how to automate the download of datasets, • a **manual**containing all detailed information on the bulkdownload facility, • the **table of contents** that includes the list of the datasets available, • the "**dictionaries**" of all the coding systems used in the datasets.
Use case coverage / relevance	PUC2 (“yoghurt”) all tables (?)

Name	Size	Type	Date	
[comext]		DIR	19/03/2014 11:57:23	
[comp]		DIR	20/03/2014 10:38:05	
[data]		DIR	20/03/2014 11:01:19	
[dic]		DIR	20/03/2014 11:01:22	
[metadata]		DIR	14/03/2014 15:30:21	
Bulkdownload_Guidelines.pdf	51.20 KB	pdf	28/06/2013 11:46:13	Download
COMP_20140320110043.zip	33.57 KB	zip	20/03/2014 11:02:03	Download
ESTAT_20140320110043.zip	8.91 MB	zip	20/03/2014 11:01:55	Download
read_me.pdf	195.28 KB	pdf	23/07/2010 10:55:16	Download
table_of_contents.xml	15.41 MB	xml	20/03/2014 11:00:40	Download
table_of_contents_de.pdf	417.62 KB	pdf	20/03/2014 11:00:32	Download
table_of_contents_de.txt	1.04 MB	txt	20/03/2014 11:00:43	Download
table_of_contents_en.pdf	392.19 KB	pdf	20/03/2014 11:00:31	Download
table_of_contents_en.txt	994.24 KB	txt	20/03/2014 11:00:43	Download
table_of_contents_fr.pdf	414.62 KB	pdf	20/03/2014 11:00:32	Download
table_of_contents_fr.txt	1.07 MB	txt	20/03/2014 11:00:43	Download

56	German Federal Ministry of Economics and Technology
URL	http://www.bmwi.de/English/Redaktion/Pdf/facts-about-german-foreign-trade-in-2012,property=pdf,bereich=bmwi2012,sprache=en,rwb=true.pdf
Language	English
Content	Facts about German foreign trade, export/import, ranking in the 10 largest trade nations, Germany's foreign trade partners (only a single file)
Format	PDF with many tables and graphs
How to extract info?	Download the PDF PDF to txt converter (problem tables!?) Information extraction
Use case coverage / relevance	PUC2 ("yoghurt") this one PDF file

57	German Trade and Invest
URL	http://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Industry-overviews/industry-overview-food-beverage-industry-en.pdf
Language	English
Content	Report about food and beverage industry in Germany Many tables, graphs and images (only a single file)
Format	PDF
How to extract info?	Download the PDF PDF to TXT conversion (or pdf-to-html for images and graphs) Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) - this one file

58	Karen Juliane Schröder: Cannibalization on the yoghurt market
URL	http://www.agric-econ.uni-kiel.de/arbeiten_PDFs/2012/MA2012SchroederML.pdf
Language	English
Content	Master thesis on “Cannibalisation on the yoghurt market” (only a single file)
Format	PDF
How to extract info?	Download the PDF PDF to TXT conversion Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) - this one pdf file

59	Industry Analysis: Competitors
URL	http://www.euromonitor.com/yoghurt-and-sour-milk-products-in-germany/report
Language	English
Content	Country report: Yoghurt and Sour Milk Products in Germany
Format Paywall restricted	Html PDF (a sample report: Sample Yoghurt and Sour Milk Drinks Market Research Report) Excel sheet (a sample data sheet: Sample Yoghurt and Sour Milk Drinks Data The complete report can be bought for 675 €
How to extract info?	Web crawling (only the landing page + the sample reports (?)) Format conversion (PDF to TXT) Structured data extraction (from Excel)
Use case coverage / relevance	PUC2 (“yoghurt”) - report summary and samples - the actual report only against payment

60	Industry Analysis: Competitors
URL	http://en.wikipedia.org/wiki/M%C3%BCller_(company)
Language	English
Content	Wikipedia entry about a German diary producer
Format	html
How to extract info?	Web crawler (only the landing page) Or Wikipedia API HTML parser (to parse structured Wikipedia information such as:  Information extraction for the text analysis (company structure, list of products)
Use case coverage / relevance	PUC2 (“yoghurt”) - this page

61	Industry Analysis: Competitors
URL	https://www.google.es/search?q=ehrmann+joghurt&oq=ehrmann+jog&aqs=chrome.1.69i57j0l5.6484j0j4&sourceid=chrome&espv=210&es_sm=93&ie=UTF-8
Language	German
Content	**Google search** results for "Ehrmann joghurt" Links to relevant pages Images
Format	html
How to extract info?	Web crawler Html parser
Use case coverage / relevance	PUC2 ("yoghurt") - more than 60.000 matches in the Web

62	Europa – summaries of EU legislation (product labeling and packaging)
URL	http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_en.htm (*_bg.htm, *_de.htm, *_es.htm, *_fr.htm)
Language	All European languages
Content	Summary of EU legislation on Labelling, presentation and advertising of foodstuffs
Format	html
Useful info in the URL	Language: http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_en.htm http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_de.htm
How to extract info?	Web crawler (only the landing page) Html parser Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) - this one page

63	International Dairy Food Association (European health certification program)
URL	http://www.idfa.org/files/resources/eu_health_certification_program_draft_091411.pdf page does not exist anymore!

64	German Business Portal (Overview of market access of food and beverage)
URL	http://www.ixpos.de/IXPOS/Navigation/EN/Your-business-in-germany/Business-sectors/Consumer-goods/food-and-beverage,did=263444.html
Language	English
Content	Industry overview of German food and beverage market
Format	html
How to extract info?	Web crawler (only the landing page) Html parser Information extraction for text analysis
Use case coverage / relevance	PUC2 (“yoghurt”) - this one page

65	IFS Food Packaging Guideline
URL	http://www.ifs-certification.com/index.php/en/imprint-left-en/51-global-news/2005-news-2013-10-23-vplf-v2-en
Language	English, German
Content	IFS Food Packaging Guideline - only the introduction and a brief descrizion - the full guideline can be bought in the IFS shop (10 €)
Format	- Html - full guideline PDF
How to extract info?	Web crawling Html parsing PDF conversion (for the full guideline version)
Use case coverage / relevance	PUC2 (“yoghurt”) - brief summary + one document (against payment)

66	General requirements and standards for food and agricultural imports into Germany
URL	http://www.spring.gov.sg/archives/ETAC/Documents/Germany.pdf
Language	English
Content	General requirements and standards for food and agricultural imports into Germany
Format	PDF Only one table
How to extract info?	The PDF file seems to be not convertible into a processable format (e.g. txt)
Use case coverage / relevance	PUC2 (“yoghurt”) -(if successfully converted) – one table with structured information

67	PEPPOL (Pan European Public Procurement Online)
URL	http://www.peppol.eu/
Language	English
Content	A European Commision's portal which supports companies, in particular Small and Medium Sized Enterprises (SMEs), to bid for public sector contracts anywhere in the EU.
Format	html
INFO:	PEPPOL is right now undergoing a process of "transfer" into what will be the new Open PEPPOL. It's not a portal in itself but a system that makes possible the interoperability of Platforms, both public and private. The sustainability of PEPPOL is still under discussion and it's being complex to work with them for the time being.
Use case coverage / relevance	PUC2 ("yoghurt") (?)

77	TED (tenders electronic daily)
URL	http://ted.europa.eu/TED/main/HomePage.do
Language	English, German, French, Spanish, Bulgarian
Content	Supplement to the Official Journal of the European Union. It provides free access to business opportunities. It is updated five times a week with approximately 1500 public procurement notices from the European Union, the European Economic Area and beyond. The user can browse, search and sort procurement notices by country, region, business sector and more.
Format	html
Useful content	Registered users are allowed: To access the entire content of TED, including the archive. To personalise search profiles, according to their needs. To get e-mail alerts based on their search profiles. To personalise RSS feeds for their web sites and RSS readers.
Paywall restrictions	No (the content is for free) But: registration needed
Use case coverage / relevance	PUC2 (“yoghurt”) (?)

78	SPOCS
URL	http://www.eu-spocs.eu/index.php
Language	English, German, French
Content	SPOCS is a large-scale pilot project launched by the European Commission in May 2009. It aims to build the next generation of online portals (Point of Single Contact or PSC), which every European country now has in place, through the availability of high impact cross- border electronic procedures.
Format	html
Useful content	From the project description: Businesses seeking to expand into other countries often struggle to comply with all the regulations they need to follow. Applying for licenses, permits and completing other administrative procedures in another country can be very complicated. SPOCS provides seamless electronic procedures by building cross- border solutions based on user's country's existing systems. For example: An Italian real estate agent wants to expand his business to Bremen, Germany. The information provided by the Bremen PSC does not currently specify the Italian documents required. SPOCS should allow the user to do all the administrative procedures online and cross-border via the Point of Single Contact
Paywall restrictions	No (the content is for free) But: registration needed
Use case coverage / relevance	PUC2 ("yoghurt") (?)

79	The Market Access Database (MADB)
URL	http://madb.europa.eu/madb/indexPubli.htm
Language	English (other languages listed but not really available)
Content	Information for Export from and into EU: Tariffs, Procedures and Formalities, Statistics, Trade Barriers, Sanitary and Phytosanitary Issues, Rules of origin
Format	html
Useful content	The Market Access Database (MADB) gives information to companies exporting from the EU about import conditions in third country markets. In particular: **Tariffs**: Duties & taxes on imports of products into specific countries **Procedures and Formalities:** Procedures & documents required for customs clearance in the partner country **Statistics**: Trade flows in goods between EU and non-EU countries **Trade barriers:** Main barriers affecting your exports **SPS (Sanitary and Phytosanitary Issues):** Food safety/animal health/plant health measures **Rules of Origin:** Preferential agreements & rules of origin In order to reach this information the user has to select in an interactive input mask the name of the country of interest and the product code (both can be selected from lists)
Paywall restrictions	No (the content is for free) Only acceptance of the Terms and Conditions of Use is required.
How to extract the info?	We have not found a way to automatically export the relevant data. For this, we can consider contacting the European Commision and asking for support.
Use case coverage / relevance	PUC2 (“yoghurt”) (?)

B.1 Blogs

URL	name	Language	keywords	Use case
http://decor8blog.com/	Decor8	English	Kitchenaid, Liebherr, AEG, …	PUC1-MM
http://cleantechnica.com/	Clean Technica	English	energy	PUC1-J
http://blogs.cfr.org/levi/	Council on foreign relations – Energy, Security and Climate	English	energy	PUC1-J
http://volker-quaschning.de/index.php http://volker-quaschning.de/index_e.php	Volker Quaschning	German English	Energie energy	PUC1-J
http://smartgridwatch.wordpress.com/	Smart Grid Watch Blog	English	energy	PUC1-J
http://greenlivingonline.com/blog	greenliving	English	Kitchenaid, Liebherr, AEG, …	PUC1-MM
http://www.diisign.com/	DIISIGN	French	Kitchenaid, Liebherr, AEG, …	PUC1-MM
http://blog.hospimedia.fr/	HOSPIMEDIA	French	Kitchenaid, Liebherr, AEG, …	PUC1-MM
http://blog.bancsabadell.com/	Banco Sabadell	Spanish	yoghurt	PUC2

C Appendix: Questionnaire on media monitoring practices

C.1 Questionnaire

"The purpose of this questionnaire is to collect the input and data for Task 2.1 (Empirical study). The empirical study will identify the type of information that needs to be extracted in order to facilitate the targeted extraction of news and financial information to support the media monitoring and SME internationalisation use cases.

The focus is on describing and analysing the existing work flows and the available information resources from the use case partners. This study will serve as a basis for the technology partners to understand the status quo and for the entire consortium to then define the resulting priorities.

Feel free to provide as much information as possible to describe your existing workflows and resources, as this will be important for the technical partners to understand how it works today.

Questions:

- What are the typical application scenarios?
- Which different steps and workflows are needed to produce the required results?
- Which information and data resources are being used?
- What type of information is then being extracted?
- How the information is aggregated, summarised and archived?
- How the outcomes are reported to the customers/recipients?

C.2 PUC1, journalistic media monitoring - Response from DW

1) Describe your typical application scenario(s)

Please each use case partner describe the typical application scenarios in media monitoring (PM), SME internationalisation (PIMEC) and journalistic work flow (DW).

DW journalists monitor what is happening all over the world, to write about news, give background information on ongoing stories and keep people informed. It is also part of DW's work to educate and to entertain people. As an international broadcaster from Germany, DW covers events in Germany, but also worldwide with a German perspective. It is also DW's role to inform about Germany, help people who are trying to learn the language and give insights into German culture.

In order to do this it is essential to watch different channels, ranging from other local, national and international media outlets (like regional and national newspapers, e.g. Berliner Zeitung or Süddeutsche Zeitung or the New York Times, Times of India, The Guardian) to blogs and social media, wire services (dpa, reuters, etc.), official government news sources (press conferences, press releases) or direct interviews and statements.

MULTISENSOR would be used in the early stages of research, when a journalist sits down to get background information on a topic or a story he/she is working on. The application would help to go through a large number of resources and to summarise the information

already available on a certain topic or person and help detect a new angle on a story or an interesting way of shedding light onto a topic.

So in general, the typical application scenario would be the research process at the beginning of a story.

2) For each application scenario please describe <u>in detail</u> the different steps and workflows to produce the required results

Please describe in detail all steps in your workflows from e.g. the data collection and cleaning, to the analysis and the presentation of the results to the user.

Steps of the journalistic workflow

1) Creating an idea and a theory to write about (identification of topic)
2) Research first information, finding facts and data through research
    a. News outlets
    b. Other media outlets
    c. Social media
    d. Online news alerts
    e. Press releases
3) First quick validation of facts and data
    a. Secondary sources
    b. Telephone contacts
    c. Background checks
    d. Fact checking
4) Double checking story idea/theory, adding to the story
5) Second round of research for more background info
    a. Background information on people, facts, events, places used in story
    b. Researching for interview partners, experts, institutions (expertise, background, popularity)
    c. Telephone and direct contacts for more details/background information
6) Validate story: Check content for correct a) facts, b) authenticity, c) relevance d) balance
7) Verification through colleagues/editor in charge (4-eyes-principle)
8) Publication of the story
    a. Through own channels
    b. Republishing through other channels (second run)

3) Which information and data resources do you use?

Please list the exact data sources, data type, encoding types, size etc. of the resources which serve as the data basis for your applications (cf. also Table 1.2 from the DoW, p.19 of 74).

- Social Media channels (Twitter, Facebook, Reddit, etc.)
- News wires (like Reuters, dpa, AP, AFP, ...)
- Online DataBases (like http://www.munzinger.de/, etc.)
- Library data bases (http://staatsbibliothek-berlin.de/, https://www.deutsche-digitale-bibliothek.de/, http://www.loc.gov/, ...)
- Official media channels (like, sueddeutsche.de, nytimes.com, tagesschau.de, bbc.co.uk, ccn. com, ...)
- Direct contacts/expert interviews via telephone, in person etc.

- Internal contacts

Formats (depending on what you produce)

- Text
- images
- Video
- Audio
- Multimedia

Datatypes don't really play a role as DW wouldn't reuse material found in the research, except if coming from own databases (raw video/audio materials etc.)

4) What type of information are you then extracting / filtering / analysing?

Please describe in great detail the content analysis step, focusing in particular on the type of information that is extracted and filtered.

It depends on the story the journalists are after which information is being extracted, but overall its facts, that journalists are looking for.

If you are writing about a current political event it is most likely that you are looking for information about the people and organisations involved. This could be politicians, managers, celebrities or ordinary people.

It could also be background information about what has led to a certain event or situation. A journalist might also be looking for connections between people, networks or patterns of movement and activity.

It might also just be dates, when something took place or where in relation to something current.

The information could be historic or current information, simple quotes from people or just facts concerning the when, where, what, how and why.

The filtering and extraction process is done by the journalist in regards to the topic they are working on. It is most likely to be done manually with the support of computer systems – in order to save the material and keep the relation to its origin. Same goes for **the analysis –** the journalist is most likely to be doing this manually with the support of software (e.g. using excel to structure data and make it more readable).

5) How do you aggregate / summarise / compile / archive the information?

Please describe how do you structure, store, interpret and summarise the extracted information.

The information is structured according to topic and relevance. The journalist would sort the information so it would best support the story he/she is trying to tell.

There is a different variety of tools available on the market to support this process. At DW the editorial departments are using Microsoft office products as well as Open Media and DW's own CMS to produce and publish (text) materials.

There is other software used to produce and curate audio and video like dira, but the publication eventually also goes through the CMS (if not broadcasted through radio or TV channels.)

Deutsche Welle uses its own data archive system to store its produced material.

6) How do you report the outcome of your work to the customer / recipient?

How do you communicate the results to the user? Please describe also the query functions that are available to the user.

The information processed by Deutsche Welle Journalists is published either on the website or through one of the radio or TV stations. The formats comprise of articles, picture galleries, podcasts, online video clips and series, radio and tv shows.

Deutsche Welle also uses social media channels like Facebook and Twitter to reach its audience and offers RSS-subscriptions to keep up to date about new information on its website. Furthermore there are mobile apps to consume DW material, available for all usual operating systems.

C.3 PUC1, commercial media monitoring - Response from pressrelations

1) Describe your typical application scenario(s)

Please each use case partner describe the typical application scenarios in media monitoring (PR), SME internationalisation (PIMEC), (DW).

As a media monitoring company, pressrelations has the following workflow:

We are usually contacted by a client interested in monitoring his own company and brands. In addition, many clients ask us to follow their particular markets which demands monitoring of competitors, products and topics and to include relevant information in a daily or weekly media review.

The first step in our workflow would be to understand the client's monitoring needs and to translate these needs into keywords and keyword combinations.

Another important input we need before starting is the scope of the media set. Different types of media have very different production workflows:

- Print media is integrated either by
  - accessing central databases that offer the full text articles via keyword searches,
  - processing e-papers by identifying articles via full text search and semi-manually dissecting the paper to create articles or
  - purchasing these from third parties and uploading them onto our system.
- For online media, crawling of a predefined set of global news sites is performed by the pressrelations crawler software.
- RTV clips are gathered by purchasing these from specialised vendors, who use proprietary technologies for capturing, speech-to-text, search and audio/video-cutting.
- Information from social media is gathered via the public APIs of the various social media channels (Facebook, Twitter, YouTube, Google+) while weblogs are searched

and included via so-called aggregators (e.g. Google blog search). Forums are tracked either manually or enter our system through an interface with a cooperation partner and are treated the same way as online articles.

Once the keywords and media set are fixed, we define a so called search profile. Data from our own crawlers and social media conversations delivered through APIs are automatically written into the various client search profiles whenever keywords match. The crawlers go through all available sources every 30 minutes.

Normally, a client will want a daily delivery of a media review which is to include all information from all sources. Every morning, an editor will go to the search profile (and possibly sub-profiles) and will be presented with the news items found since the previous day. The news items are displayed with metadata such as media name, date and country. Highlighting of the keywords is included. The editor selects the articles according to his briefing and saves the articles for further processing. Irrelevant articles are discarded.

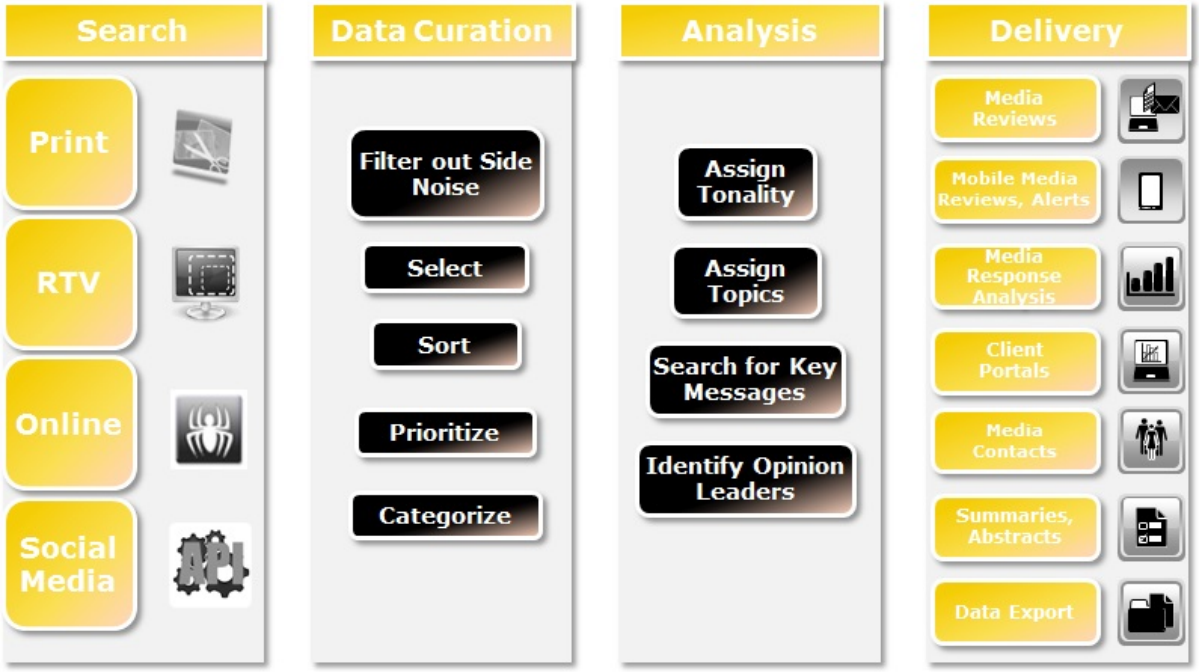
Print articles form third-party suppliers are manually checked for relevance before upload.

During the coding process the editor can open a link to the original source and validate if the text is correct and complete. At any time during processing, the editor can also add a summary or a translation for the article in a separate field. The articles text can also be edited if necessary.

In the coding process the articles is assigned manually to one or more topics/companies/other related codes according to the briefing instructions.

Once coding of all articles is completed, the editor will access NewsRadar, pressrelations' client front end media monitoring portal from where articles can be sorted and media reviews are generated from a choice of several output formats, e.g. HTML, Text, Excel or PDF-Versions. The reports can be forwarded to preconfigured distribution lists. The media review product creation is completed with this step.

Some clients request and commission an in-depth analysis of the monitoring results on top of standard media monitoring. For this, client and media monitoring company agree on a so called analysis codebook. A codebook contains coding parameters such as competitors, brands, main article topics, products etc., but coding will also require judgment over exclusivity or weight, key messages adherence, sentiment, etc. The codebook is always customised according to individual project and client needs. For each news items, all codes will be checked. Assigned codes are stored as code strings. Code strings represent connected codes with regards to content. Information from the code strings is used for reporting of media analysis which is presented to the client as graphs within his individual NewsRadar front end portal.



This chart illustrates pressrelations' services.

2) For each application scenario please describe <u>in detail</u> the different steps and workflows to produce the required results

Please describe in detail all steps in your workflows from e.g. the data collection and cleaning, to the analysis and the presentation of the results to the user.

Data collection: Online articles, blogs and user portals are crawled by pressrelations' crawlers. If a news item from an online source from within the preconfigured media set contains a relevant search term, it will be assigned to the search profile for review. The editor will see the article headline, date, time, name of media, country, number of characters and key word hits, a clipping with all keywords (highlighted) and the link to the original source. If the context of the hit is not clear from the clipping, the editor has the possibility to view the full extracted text or follow the link to the original source. Though it is possible to automate the selection and continue with all articles containing keywords, at this point it is usual for an editor to decide if the context of the article fits the briefing. Each article is either marked as clip or rejected.

All new articles can be clustered for the selection process. The system compares word strings within the article and presents highly similar articles en bloc. This facilitates the selection process.

Once all articles have been reviewed, the selection is saved.

As mentioned, the following social media sources enter our system via API: Facebook, Twitter, YouTube, Google+. The selection process is the same as with other online media.

Print and RTV news are delivered to us by suppliers, researched from third-party databases or read and scanned in house.

For print articles, e-papers and paper clippings are checked for relevance and scanned. The articles enter our system via a clipping-software provided by CCS or a PDF/JPG upload. The text is converted by OCR and written into the text field and PDFs or JPGs are attached. For

our main German print provider we also use an upload function for text files. For these articles, it is possible to add a scan of the original document afterwards.

Most of our RTV providers use speech-to-text technology and some have interfaces with our system. In this case, the editor can go through the same selection process as with online media and is presented with a short text clipping. The article links to a short recording that the editor can use to check the relevance of the news item.

For any media type, the editor has also the possibility to add a news article manually if a relevant article is missing. Necessary fields that need to be filled out in this case are date, media name and headline.

Cleaning: Before encoding, the editor has the possibility to cluster articles and save them as similar or identical. The workload can thus be reduced as assigned codes will be written to all clustered articles. While processing the article, the editor can validate and complete the article text if necessary. This might be the case sometimes with hyperlinked parts of the article or with following pages. Superfluous data may have to be deleted, e.g. if the crawlers have erroneously included advertisement or the like. The editor also needs to copy and paste the author into the corresponding article field.

Analysis: The analysis is a customised part of our service. Essentially, the client can decide on the degree of depth he needs for his analysis. Codebooks containing analysis nodes are created. Common nodes are: company, exclusivity, reputation, topic, products, brands, opinion leaders, press releases, key messages, tonality etc. The analyst can configure each node as optional or mandatory. He can also determine if it may contain one or multiple values.

Every article can be analysed either on the level of the whole article or on the level of singular statement within the article. The analyst can mark a sentence or paragraph and store code information for it. If a code applies, the analyst will tick the corresponding checkbox. For each research object (e.g. company) the analyst creates a code string containing all applicable codes. Example: Company A – topic(s) – products – exclusivity – etc. The number of code strings can differ strongly from project to project. Also a tonality can be applied to the code string with regards to the research object.

Once all articles have been analysed, it is possible to generate queries in the analysis section of our back-office. The analyst can see all codes contained in the analysis nodes and he can make a selection. It is possible to split up the results with codes from a second analysis node or restrict the selection. Also, the analyst can split up or narrow down the results with metadata from the data base such as media source, language, country etc. All queries can be made either on article or statement/code string level. The system will display the results e.g. as stacked-bar charts. These can be saved and annotated. The analyst can download an excel file containing all selected articles with various annotated metadata for each article. He can also click onto the bars in the charts and will be linked to the selected articles which he can then export as HTML, PDF etc.

Presentation: The analyst generates queries in the back-office as described. All or a portion of the saved queries can be included into the NewsRadar frontend application either as graphs or as portlets containing articles. In the NewsRadar client frontend, the client can click on each graph and can use a drill down function that shows first all articles behind the

bar, then links individually to the original article. All charts can be saved and exported as JPG or PNG.

All charts are automatically updated on a daily basis after the editor/analyst has finished the daily procedure.

In another section (the media review section) the application contains a search grid where the client can perform keyword searches within his search profile and use several filters e.g. for media type, country etc. At any point, the client can select articles for export in different formats such as PDF, HTML, Text, Excel, etc.

On top and apart from the frontend application, the analyst may be asked to create annotated PowerPoint presentations. Here, the analysis data is based on the coded articles within the system but creating such a media response analysis within an outside application such as PowerPoint allows for a greater variety of presentation.

3) Which information and data resources do you use?

Please list the exact data sources, data type, encoding types, size etc. of the resources which serve as the data basis for your applications (cf. also Table 1.2 from the DoW, p.19 of 74).

The encoding for the text parts is usually UTF-8.

Online news including social media: Data is extracted by pressrelations' crawlers directly from a news website by crawling the headline and text. Additionally the URL of the source article is saved. Alternatively RSS news feeds are imported if this is possible.

Public APIs from Twitter, Facebook, YouTube and Google+ are also accessed to retrieve information in JSON format.

Daily 500.000 news articles are imported. Approximately 15% are social media news articles from Twitter, Facebook etc.

Print news: Pressrelations scans and clips articles in house with a clipping software provided by CCS.

A XML file with a corresponding PDF file is generated and imported into our database.

PDF files are usually up to 5MB, rarely larger than 10 MB.

Also depending on external suppliers, data can be either texts or PDF/JPG documents with text converted by OCR.

Usually data from an external provider is provided as a ZIP files file that is imported into our system. ZIP files contain a XML file with all articles. Additionally PDF/JPG documents are also sometimes included in the ZIP file.

A ZIP file can be up to 100 MB containing multiple articles. PDF files or JPGs are up to 10 MB large.

Daily about 3200 print articles are imported.

Radio and TV news: Radio and TV articles are imported by accessing our suppliers API. Suppliers provide a short sentence or summary in text form and a link to a short media clip on their website.

If a complete clip is ordered by a client, the media file is uploaded manually into our system by the editor and attached to the articles.

Permitted extensions of data files are: FLV, WMV, MP4, MP3. Data size is usually less than 100 MB, maximum size is 200 MB.

About 250 radio and TV articles are imported daily but mostly without the corresponding media file.

4) What type of information are you then extracting / filtering / analysing?

Please describe in great detail the content analysis step, focusing in particular on the type of information that is extracted and filtered.

For analysis, some metadata are collected for each media and retrieved from our database. These are:

All media: country, language, media type (e.g. daily newspaper (regional or national), popular press, advertising journal)

Online media, blogs, forums: visits

Print: printed edition, distributed edition, print reach

For each news item, we collect the following information within our system in addition:

All media: Date, headline and text, Author, Advertising Value Equivalency (AVE)

Print: Page start , Number of pages, Number of illustrations, Department

RTV: Viewers, Duration (min:sec) (On demand)

Twitter: Followers

At present, it is not possible to trace the following values within our system. Any analysis of these values would depend upon an Excel file:

Facebook: Likes

Google+: +1

YouTube: Views

It is also possible to track backlinks within an article on demand. These links usually need to be checked and verified. Backlinks are used to display connections of articles or tweets etc.

The content analysis of the news is done manually and can vary greatly per client. A standard analysis may contain – but is not limited to – the following analysis nodes:

Company, Exclusivity, Reputation Topic (e.g. Financial/Economic Performance, Strategy & Vision, Management & Leadership, Products & Services, Social Responsibility Topic, Human Resources/Employer Branding), Products, Brands, Opinion Leaders, Press Releases, Communication Messages

Each code string can be combined with tonality. Possible scales are -1 to +1, -2 to +2 or -3 to +3.

5) How do you aggregate / summarise / compile / archive the information?

Please describe how do you structure, store, interpret and summarise the extracted information.

In the backend portal, the analyst defines queries according to the client's information needs. An example:

A client is interested to see how much coverage its divisions generate in the course of time. The analysis codebook therefore contains a node called divisions, listing all relevant business units of the company. The analyst can create a query by selecting all codes in this node and a time frame e.g. "last 90 days". For display, e.g. a weekly consolidation of the results is chosen. A stacked bar chart with a bar for each calendar week, split by divisions is then created within the client front end portal. This chart, based on a defined analysis query is updated automatically within the reading/analysis process. All queries can be stored and exported. If the query has a variable time frame e.g. "this quarter", the same query is available for continual use.

There are multiple variations possible, a few examples: Instead of number of articles, the analyst might also choose printed or distributed edition, visits, Advertising Value Equivalency (AVE) or followers. Instead of splitting up by divisions, displayed news items might be split by media source, language, country, media topic, or media type. It is possible to restrict the results to any code, tonality, media source etc.

Summarisation and interpretation depend completely on the analyst. Usually, the analyst is well informed on the client's requirements and has intimate knowledge of the market to be discussed. Background information on certain topics may need to be researched beforehand. Mostly, the analyst is already involved in the selection process and encoding of the data.

In all the queries that the analyst generates, he searches for significant and interesting peaks. The system offers him the possibility to click on forwards to the original articles. If the analyst notices e.g. a major change in the share of voice the monitored competitors of a client generate compared to the previous quarter, he will look for an event that might explain it. Also, if e.g. the client's brand monitoring depicts very positive or very negative coverage, it might be interesting to look into these news items further. The sentences marked during the encoding process offer significant help in this and are a source of quotes for the report.

6) How do you report the outcome of your work to the customer / recipient?

How do you communicate the results to the user? Please describe also the query functions that are available to the user.

Reporting intervals vary: these can be daily or weekly media reviews in form of an email containing a HTML or PDF document including all relevant articles. Also weekly, monthly, quarterly or yearly analysis or ad-hoc reports for specific events are not uncommon.

If a client has ordered a full analysis report, our analysts usually compile a PowerPoint presentation complete with charts and annotations. This is forwarded via email to the client or uploaded into a documents section of the frontend client portal. The scope and layout of the report needs to be agreed upon with the client beforehand.

On the frontend client portal, the client can perform some queries himself. A search form is available, where key word based searches can be performed. The client can select all topics from the analysis nodes, media specifications (e.g. media source, language, country…), tonalities and a time period. It is also possible to restrict the results to some of these points. Every user can also mark his favourite articles or use folders for an individual structuring of the data. The favourites and folders can be used for the search queries. All self-created queries can be exported to HTML, PDF, etc. or displayed either as tree-map or map.

Preconfigured queries from our backend portal (to which some clients have access) can be integrated and displayed as bar, column, line or pie charts. The client can reduce the number or selected codes himself for these charts. It is also possible to create an excel file for all articles from those charts as well as a PNG image of the chart. Tonality can also be displayed as a barometer gauge.

All charts on our frontend client portal are interactive, giving the user the possibility to see the data base and encoding behind the graphs. Users may also be assigned the rights to change the encoding if necessary. Statements the analysts have marked during the encoding process are displayed under the article clipping and provide a quick overview.

C.4 PUC1, commercial media monitoring - Response from DataScouting

In the following we append the questionnaire response provided by Stavros Vologiannidis (svol@datascouting.com) from DataScoutung (http://www.datascouting.com/).

1) Describe your typical application scenario(s)

DataScouting specialises in research and software development of multimedia information extraction solutions. Our main expertise lies in the area of media monitoring and archiving, providing solutions based on state of the art technologies such as optical character recognition, automatic speech recognition, video feature extraction and retrieval of information from multimedia data.

Our portfolio of expertise also includes:

- Handling of large data sets
- Endowing the material with relevant metadata (Natural Language Processing algorithms-NLP)
- Classification and segregation of processed content (Distributed Parallel Indexing)
- Extraction of personalised actionable information (Machine Learning methods)
- Disperse and exchange results (Open Standards interoperability)

DataScouting invests in big data processing, clustering techniques and web based user interfaces and thus provides both standalone and software as a service services. Certified with TUV Rheinland ISO 9001:2008, DataScouting can help customers address all their media monitoring needs. We provide high quality services, designed for each customer's needs using robust and proven technologies.

The two main products **MediaScouting Print** and **MediaScouting Broadcast** are robust and scalable solutions for monitoring in real time, print, broadcast and Internet media in a global scale.

About MediaScoutingPrint:

- Can be installed either at the client's premises OR can be provided as Software as a service
- Supporting automated and semi-automated lines for digitizing, archiving, retrieving and analysing press media
- Web based user interface for article processing, customer retrieval, advanced management and alerting system
- Available as a standalone platform and as Software as a Service (SaaS), guaranteeing small initial investment

Main features:

- **Scalable-distributed-modular:**Can scale to process hundreds of press publications every day
- **Automatic topic detection and customer assignment**: Each article is assigned to one or more predefined topics/customers using fuzzy keyword matching and lemmatisation techniques.
- **Full text indexing of articles:** Robust and advanced query system with live translations, automated result clustering.
- **Sentiment analysis of articles**: available for specific languages
- **Multilingual**: Supports Optical Character Recognition in 189 languages and guarantees success rates up to 99%
- **Distribution of clippings**: Articles can be distributed in a multitude of ways (for example, fax, email, paper, RSS, API etc.)

About MediaScouting Broadcast:

- Provides a robust and 24/7/365 scalable solution for recording, archiving, indexing, retrieving and analysing multimedia content origination rom broadcast feeds in real time
- A versatile and powerful audio/video stream management and retrieval environment including cutting edge technologies such as Automated Speech Recognition (ASR), metadata parsing from the Electronic Program Guide (EPG) and an advanced query and alert system
- Administrative, customer and broadcasting management interface are accessible via Web Based User Interface (HTML 5 capable browser)
- Web based interface for retrieval management, monitoring and alerting
- Includes recording support for terrestrial and satellite video streams, radio broadcasts via airwaves or internet streams, even video hosting web sites

Main features:

- **Multilingual**: supports Automatic Speech Recognition in many languages, including Arabic, English, French, German, Hebrew, Mandarin Chinese, Norwegian, Russian or Spanish
- **Real time**: can scale to process hundreds of feeds at real time, while storing the highest quality terrestrial, satellite and IP feeds encoded in MPEG4 H.264
- **Versatile multimedia player:** highlights relevant textual content and includes easy video editing controls and downloading/sharing of the user modified streams

- **Robust query system:** automated result clustering and sentiment analysis per query
- **Compression** audiovisual feeds to one fifth of their size without loss of quality
- **Natural language processing** algorithms perform analysis on the textual metadata and expose hidden information in the data such as entities and sentiment
- **Retrieving**: searching for keywords in the title, transcribed text or metadata
- **Optimised search** with advanced parameters such as time period of broadcast, media type or a specific media channel
- **User defined tags** can be created and manually assigned to different clips for categorisation, intra-platform sharing and automatic publishing on any platform
- **Statistical analysis** of traffic usage and customer views. Different categories of information such as viewed streams, tags, media type, time zone broadcasted are used to generate facets
- **Individual clip viewing** via versatile and standards based vide player which allows cropping, downloading and sharing any relevant audio video feed either to the client's website or to social media platforms
- **Dynamic calendaring** for recording streams, assigning priorities

DataScouting also provides **digital convergence services**. DataScouting can help organisations like libraries, archives, museums, public organisations or private company archives, capture, manage, store, share, preserve and deliver information appropriately and responsibly. We provide digitisation services with a combination of consulting, microfilming, digitisation and cataloguing services that increase information availability, correlated metadata, archive wide analysis and "paperless" processes and reduce maintenance costs, access time, organisational resources and physical archive size. DataScouting digitises materials (books, newspapers, magazines, scientific journals, audio-visual material, 3d objects, maps, manuscripts) and indexes all relevant data that is incorporated into web-accessible databases with full-text search capabilities.

2) For each application scenario please describe <u>in detail</u> the different steps and workflows to produce the required results

Please describe in detail all steps in your workflows from e.g. the data collection and cleaning, to the analysis and the presentation of the results to the user.

MediaScouting Print

MediaScouting Print provides a robust and efficient workflow for scanning, archiving, clipping, indexing, and retrieval of print media targeted at companies and organisations providing media monitoring services. It uses an amalgam of **technologies** that include both open source and proprietary software, combining a cutting edge interface and a proven stable platform. It has a modular clustered **workflow**, so it can be simultaneously flexible and expandable. Inputted media can include paper media up to A2 size and any type of electronic document like Portable Document Format.

For a better distribution of human resources and flexible work progress, **article segmentation** and **subject reviewing** were defined as two distinct roles. The digitised media is segmented into relevant articles and preliminary metadata are added to the database. The images are processed with optical character recognition, and the result is **lemmatised** and indexed for **full text search.** Based on the customers requested keywords and subjects, the

articles are grouped into **semantic categories**. Based on the settings of each subject, the article may be published immediately to the customer, or submitted to an article reviewer.

MediaScouting Print **web based user interface** is based on HTML5 and allows not only to quickly deploying to any computer and operating system, but also enables **teleworking** from any internet enabled position.

Publishing options include printed media in custom template with all relevant information, **fax**, **and email**, custom **API** for integration and more importantly on the **Web**. The customer web interface includes all the bells and whistles of a modern platform like 'application-like' look and feel, multiple format output options (JPEG/PDF/DJVU/TEXT/DOC/EMAIL), keyword or full text search, articles categorised by subject, article statistics on all metadata using statistical analysis, and many more.

Administration of MediaScouting Print is on a web based interface that manages user and customer profiles, creates complex subjects from multiple keywords for customers and modifies the lexicon. More importantly, it can output reports and graphs on all aspects of the solution, ranging from user productivity to customer interaction with the platform and search statistics.

MediaScouting Broadcast

The complete software suite MediaScouting Broadcast is a versatile and powerful 24/7/365 **audio/video stream management and retrieval environment** including cutting edge technologies such as **Automatic Speech Recognition** (ASR), subject based categories with **advanced keyword search** and **metadata parsing** from the Electronic Program Guide (EPG). The administrative, customer and broadcasting management interface are accessible via **Web Based User Interface** (HTML 5 capable browser), which can be accessed through Personal Computers as well portable devices such as tablets.

The administration interface is a versatile tool for adding, removing, monitoring and editing the details of the audiovisual streams. The administration framework includes dynamic calendaring for recording streams, assigning priorities on preprocessing and post processing, an interface to monitor customer accounts, viewing of the existing stored streams, as well as a statistical analysis of traffic usage and customer views.

Dynamic calendaring is the starting point of the workflow. Using a web browser, the administrator views and edits information on each stream. This information may include date, time of view or broadcast, priority, EPG, data on TV channels and digital radio stations, other relevant metadata etc. The recording parameters include continuous recording, repeated recording based on a time schedule, daily, weekly or monthly recording, one-off recording or adding manually a stream. Prioritizing the process maximises the utilisation of resources, like time sensitive streams and customers with increased needs. The availability of archived streams depends on the number of streams stored as well as the size of the storage space; storing options include solutions capable of maintaining feeds for one month up to many years.

Customer accounts are created and monitored from the administrative interface, and details like expiry dates and access limits are set. Additionally detailed user statistics are available for reviewing and analysis.

The web based user interface provides an **intuitive** and **effortless experience**. The user has access to the streams available to his access level. All streams **maintain original quality**, but are **optimised for web access** for the best viewing experience. On the initial view of streams, a user can either view a stream from the beginning, or select a word from the recognised speech text to jump to that position and start playback.

Searching for keywords in the title, the text or the details of the available streams, enables the user to easily find any relevant audiovisual streams. The results of the search terms (or query) are shown on the center column with the words of interest highlighted. Advanced search queries can be comprised of Boolean expressions that include “AND”, “OR”, “NOT” and “NEAR” operators while each word is automatically lemmatised. The search can be optimised with advanced parameters such as time period of broadcast, media type or a specific media channel. The parameters of the advanced search of each user can be saved and reloaded at any point, so users can do repeated searches on their topic of interest.Additionally user defined tags can be created and manually assigned to different clips for categorisation according to subject or user criteria like favorites or important.

To refine a clip to the exact content of interest, an intuitive web based cut – join - crop tool for streams is integrated in the platform. All the user needs to do is indicate start and end of the desired stream. The resulting stream of the exact size and content the user wants is downloaded to his computer.All these functions can be performed efficiently using a HTML5 browser with no need to install additional software or to use a specific operating system. Accessibility to the platform from tablets or smart phones is available for the more popular mobile platforms. More importantly, DataScouting is committed to open standards, for compatibility with future technologies and media standards.

The MediaScouting Broadcast platform supports different types of users. The advanced user has access to all streams and can use all features available in the platform. Advanced options include full archive search on any media source at any time. Users that who want access only to relevant streams can use the simpler and easier base user interface. The base user enjoys the benefits of the indexing and categorisation service, either by viewing the automatically categorised results or by having a reviewer examine the results and refine the customer’s subject results. The base user interface provides all the capabilities of an advanced user like custom tags and faceting, but have access only to their assigned streams. No training or technical knowledge is necessary for viewing and using the selected streams.

The monitoring environment of all subsystems of MediaScouting Broadcast is web based. All typical monitoring statistics are available, such as **temperature, CPU and RAM utilisation, HDD space, network throughput and uptime** etc. The modular design of the platform **permits upgrading and inserting new hardware into the workflow**, **seamlessly and easily**, without stopping the service.

3) Which information and data resources do you use?

Please list the exact data sources, data type, encoding types, size etc.of the resources which serve as the data basis for your applications.

MediaScouting Print:

- In house scanning of newspapers and magazines by the client

- Inputted media can include paper media up to A2 size and any type of electronic document like Portable Document Format
- OCR using Linux Abbyy Finereader SDK. Abbyy Finereader is the leader in optical character recognition. Its engine provides outstanding OCR accuracy and currently supports 186 languages.
- Lemmatisation using open source / commercial products: Lemmatisation is especially important in heavily inflected languages such as Greek or Spanish. MediaScouting Print uses either open source stemmers, or commercial products to group together the inflected forms of a word.
- Image processing / typesetting tools: Articles are heavily processed and are automatically added to a template of choice which includes among others the newspaper logo, date, etc.
- HTML5 web based user interfaces: All the web based interfaces are based on HTML5, thus providing cross browser compatibility and a seamless experience.
- High performance text indexing: Each request to the system finishes in msec.

MediaScouting Broadcast:

- Providing a robust and scalable media monitoring platform that includes TV, radio and Internet feeds
- Using state of the art multilingual automatic speech recognition supports recognition of most of the European countries languages and automatic translation of the transcribed clips in real time
- Design and implementation of a storing space of hundreds of terabytes that correspond to hundreds of years of multimedia data encoded in H.264 High 3.1.
- Using HTML5 technologies to create a web based intuitive platform that allows the seamless navigation on TV or radio streams and supports different user roles
- Availability of a documented agile API that complies with international standards (REST)
- Automatic alerting of users through email, SMS or RSS in real time

The process to record and make streams available for viewing to the advanced user interface or the base user interface has the following steps:

• Connecting to the dynamic calendaring system to indicate streams which need to be recorded

• Indication of stream preprocessing and post processing priority

• Selecting archive options for each recording

• Creation of customer accounts based on needs (advanced interface or base interface)

• Advanced stream search with complex Boolean parameters and saving search options

• Assignment of saved advanced search settings to customers

• Creation of tags and assignment to users

• Assignment of tags to streams to be shared to the assigned customers

• Viewing of the statistical analysis of the use of the customers to indicate abuse of the system

4) What type of information are you then extracting / filtering / analysing?

Please describe in great detail the content analysis step, focusing in particular on the type of information that is extracted and filtered.

MediaScouting Print:

- Sentiment analysis of articles: available for English, French, Spanish
- The customer web interface includes, among others, articles categorised by subject and by customer preference ('favourite'), article statistics on all metadata using statistical analysis

MediaScoutingBroadcast:

- **Visualisation of statistical data** which includes comparison analysis (facets)
- The administration framework includes, among others, **statistical analysis of traffic usage and customer views**
- detailed user statistics are available for reviewing and analysis for customer accounts from the administrative interface
- A statistical analysis of the broadcasted streams is created on the fly with the use of **facets**. Different categories of information such as viewed streams, tags, media type, time zone broadcasted (morning, afternoon, etc.) are used in generating the facets. These are visual tools to assess and grade the viewed streams.

5) How do you aggregate / summarise / compile / archive the information?

Please describe how do you structure, store, interpret and summarise the extracted information.

Both MediaScouting Print and Broadcast are scalable platforms that can provide immediate access to hundreds of TB of information. Information is archived in a standard compliant format. Specifically in MediaScouting Print we use:

- JPEG in original scale and dimensions
- PDF/A
- XML describing the structure and text of the scanned images and the success rate of the OCR
- Well defined API for immediate access to information

In MediaScouting Broadcast information is archived in the following formats:

- MPEG4
- XML that describes speaker identification information, spoken text having as metadata the exact ms of each word spoken and the success rate of the speech to text engine

In MediaScouting Print we offer automatic summarisation modules in specific languages like English. The curators can query both MediaScouting Print and Broadcast and have immediate access to information in order to compile aggregate and compile customer reports.

6) How do you report the outcome of your work to the customer / recipient?

How do you communicate the results to the user? Please describe also the query functions that are available to the user.

DataScouting provides premium services and products to media monitoring companies. As technological providers, customers are at the forefront of our business. We first schedule a meeting with a customer to discuss the needs in demand. A presentation of our solutions follows with a debate on all possible features and options. A customer can test our services and provide a feedback, which is then reviewed by our team. After every presentation the customer receives a short questionnaire that will provide our team the necessary information to present the customer an offer for the services he is interested in. An installation and usage training is provided to all customers upon signature of an agreement. Support and maintenance are part of the agreement. Adjustments to products can be made based on customers' requirements. Our expertise is also reflected in our partnerships with high-profiled institutions. Communication can be achieved through different channels like personal meetings, e-mail, by phone of via Skype. Conference calls can be arranged. All information about our services and products are available on our web site with live video presentations. Additional information can be found on our social media channels (Twitter, LinkedIn) and on the FIBEP web site, since being a member of the International Association of Media Monitoring and Analysis Companies.Industry news is available on our blog.

MediaScouting Print and Broadcast, includes an integrated web portal for media monitoring companies to provide to their end clients media information. Currently our portal solutions are focused on presenting the raw information including metadata such as title, sentiment, circulation etc. to the end clients. If a media monitoring company needs reporting capabilities, these can be included and integrated to the both the backend services and the end customer portal.

C.5 PUC2, SMEs internationalisation – Response from PIMEC

1) Describe your typical application scenario(s)

Please each use case partner describe the typical application scenarios in media monitoring (PM), SME internationalisation (PIMEC), DW (?).

The typical application scenario for the internationalisation of SME's would be a company that approaches us with the intent of searching for help on how they should start internationalizing themselves. These SME's have different profiles with interest in different countries; therefore we need to start from scratch with each of them.

In order to choose in which country we should try to introduce their product we need to conduct a market analysis. For that we will need a big amount of information. This information can either be easy to find, or of easy access, or information that might be more difficult to find because of its specification regarding a certain product or because you need to look into really specific websites, or maybe depends on articles or comments from people, etc. Since this information is also necessary for the internationalisation of the SME's, we would use Multisensor's platform to give us the access to it (the type of information is specified in question #3)

2) For each application scenario please describe in detail the different steps and workflows to produce the required results

Please describe in detail all steps in your workflows from e.g. the data collection and cleaning, to the analysis and the presentation of the results to the user.

When the company approaches us asking for help on their internationalisation, we first create a deep internal study on the company itself, their financial capability, an analysis, etc.

When we know what the company can or cannot do in terms of money or resources, then we conduct a study on which countries could be feasible for it. We tend to choose a couple of countries from each region of the world that we find interesting for the company.

Once we have selected the regions that we want to study, we establish a set of criteria that will help us choose whether a country is a potential market or not. The criteria depends on the company and its sector, but can be, for example, proximity of the potential country, the income level of the company, the size of the sector in the potential country, the GDP growth, the knowledge of the sector, the barriers to enter the country, any specific regulations, etc. We weight all these criteria depending on how important we think it is for the company and then we study them in regards of the potential countries giving them numbers from 1 to 5 (1 being the less important).

Once we have done these we will have a couple or three potential countries. These will be the ones that we will study more in depth and we will investigate more about the competition, how we should enter that country, if there is any specific regulation that we need to know about, how is the demand of our product and how the people feels about it,…

The following steps will be to start creating a data base with all potential clients in that country and then, with the help of a junior assistant, contact them. Usually during these we discover new information that we were not able to find before because it was not easy to access to it, like for example specific regulation about the product or that we need to contact an agent, etc.

3) Which information and data resources do you use?

Please list the exact data sources, data type, encoding types, size etc. of the resources which serve as the data basis for your applications (cf. also Table 1.2 from the DoW, p.19 of 74).

As already explained in question 1, to help on the internationalisation of an SME we need to conduct a market analysis. This market analysis has information which is more numerical and general, like how is the country (population, capital, language, currency, religion), social indicators (population density, per capita income, GINI coefficient, life expectancy), GDP structure, economic situation (GDP growth, inflation data, economic policy, fiscal policy details, foreign sector), politic situation (foreign policy), bilateral relations, market size and foreign trade of the country (definition and characteristics and subsector related), quantitative analysis (size of supply, local production, import, export, development of imports, consumption, top brands and marketers, product marketing, containers, price levels, product quality, distribution channel, entry channel, tariffs),…
In order to find this information we can easily go to websites such as:

- http://www.oficinascomerciales.es/icex/cda/controller/pageOfecomes/0,,5280449_5282899_5283038_0_DE,00.html

- http://www.spanische-handelskammer.de/pages/viewfull.asp?CodArt=14
- https://www.cia.gov/library/publications/the-world-factbook/geos/gm.html
- http://www.germany.info/gic/
- http://www.tradingeconomics.com/germany/gdp-per-capita
- http://madb.europa.eu/madb/indexPubli.htm
- http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database
- http://www.agenciatributaria.es/AEAT.internet/Inicio_es_ES/Aduanas_e_Impuestos_Especiales/Aduanas_e_Impuestos_Especiales.shtml

For the search of the rest of the information that we need, which can be a qualitative analysis (marketing and labelling regulations because every product has its regulation, regulation labelling, food law certifications IFS, in some products you have to specify importer, image of the packaging, regulations in terms of recycling and disposal, intake level (Consumption data, yoghurt market shares by region,…), habits, trade name of the product, name or signature and address of the manufacturer, packer or an established trade-in company, list of ingredients, capacity in the container, date of minimum durability, standardisation of packaging container-volume, mention Lot, etc.

We would have to look for it in more specific websites, where we can find PDF forms, or in social media or newspapers where it depends more on emotions and opinions and we could ask MULTISENSOR to find this information in websites such as:

- http://www.ixpos.de/IXPOS/Navigation/EN/Your-business-in-germany/Business-sectors/Consumer-goods/food-and-beverage,did=263444.html
- http://www.gtai.de/GTAI/Content/EN/Invest/_SharedDocs/Downloads/GTAI/Industry-overviews/industry-overview-food-beverage-industry.pdf
- http://www.ifs-certification.com/index.php/en/imprint-left-en/51-global-news/2005-news-2013-10-23-vplf-v2-en
- http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_en.htm
- http://www.spring.gov.sg/archives/ETAC/Documents/Germany.pdf
- http://www.bmelv.de/SharedDocs/Standardartikel/EN/Food/GermanImportconditionsforFood.html
- http://www.dairyreporter.com/Markets/Kosher-dairy-demand-on-the-rise-claims-German-ingredients-supplier
- http://www.toytowngermany.com/lofi/index.php/t291358.html
- http://eu.greekreporter.com/2013/10/01/fake-greek-yoghurt-wins-german-palates/
- http://www.toytowngermany.com/lofi/index.php/t2530.html
- http://europa.eu/legislation_summaries/consumers/product_labelling_and_packaging/l21090_en.htm
- http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31979L0112:EN:NOT

- www.ami.informiert.de
- www.analizacalidad.com/es/contenido/?iddoc=211
- www.lebensmittelzeitung.net

4) What type of information are you then extracting / filtering / analysing?

Please describe in great detail the content analysis step, focusing in particular on the type of information that is extracted and filtered.

The different types of research that need to be used are the description of the country itself (population, GDP structure, social indicators, etc.), applied information (that helps us find an immediate solution for a specific problem), quantitative information (important statistics of the country, etc.) and qualitative (using mainly questions when talking personally with the company).

Some specific sources from where we will find the data necessary are:

- Literature search
  Consulting internal information formation from the company, trade publications, newspapers, magazines, annual reports, and on-line data bases.
- Interview
  Possible candidates for interviews include prospects, customers, suppliers, and other external stakeholders. They can be undertaken one by one or in groups such as trade shows, seminars, and association meetings. We can use either a Structured Interview or an Unstructured Interview.
- Focus group
  This specific method of research can be useful for product development and advertising campaigns.
- Telephone Survey
  This can provide us with valuable information that comes with experience of other companies internationalizing. From companies that contact PIMEC with questions that we might relate to the user case, personal contact with the user case's company itself, etc.
- Email Survey
  Not as effective as the telephone survey but can also provide us interesting information.

5) How do you aggregate / summarise / compile / archive the information?

Please describe how do you structure, store, interpret and summarise the extracted information.

The way we structure all the information that we need, as we have said before, is by saving all the documents with relevant information, we read them though and we filter which of these information will be interesting for the internationalisation of the company and that we should have in mind when doing the sales calls.

6) How do you report the outcome of your work to the customer / recipient?

How do you communicate the results to the user? Please describe also the query functions that are available to the user.

At the end of the study of the countries we create what we call a 'Internationalisation Promotion Plan' where the company has specified all the market analysis that has been done, all the information that we have found and any other information important for the internationalisation of the company.

After that, to keep up with the contact, one of our senior consultants, the one specifically in charge of that company, has a meeting with them once a week for several hours were they go through a rapport of what has been done throughout the past week (calls that has been made, results that we have got, if there is any specific quotation that needs to be done or anything specific that has to be addressed,…); once every three months there is a trimester rapport to have a more general view of how everything is evolving (this has to do with the fact that the companies pay a fee to PIMEC every month, and they can stop paying whenever they want, so it's very important for them to see whether they are having results or not in order to consider to pay or not pay again the fee); there is also a forecast of the future actions that should/are going to be taken,…

C.6 PUC2, SMEs internationalisation – Response from ZEBRA

Answers by:
Fran Querol
International Business Developer
@ ZEBRA design + retail
fran@zebradc.com
+34 681 267 540
www.zebradc.com

1) Describe your typical application scenario

Information regarding your company, which products you export, to which countries you have recently exported and current general information regarding exports.

ZEBRA is an SME that offers integrated services of retail. We are focused on providing firms related to a high-income/luxury sector their tailored project of retail. At ZEBRA we offer the services of design, Guild and install the commercial space for the Brands.

Currently, our main clients are based in Switzerland, France, Italy, Germany, Spain, Rusia and Great Britain.

As part of the strategy of our company, we have created strategic relations with companies from our sector in other regions of the world, to be able to offer our services woldwide. Currently we have partners in: Mexico, Morocco, South Africa, Great Britain, Germany, Rusia, Turkey, Greece, Saudi Arabia, Qatar, Dubai, Xina and Singapore.

In terms of turnover, our International activity represents a 35% of the total turnover. This figure is now growing.

2) For each of the scenarios of application, please describe in detail the different stops and workflows in order to obtain the required results.

We work with what we call collaborators. We base the decision of the country in the needs of our clients and in the data that they provide us such as: tourism index, new flight routes, new airports, GDP per capita, level of high-income population, etc.

3) Which information sources and data you use?

Mainly we take this information from our partners in the country, like our collaborators. The information it is always on the Internet, sector magazines, fairs, etc. We have get information from our clients.

4) What type of information you extract/filter/analyse from these sources?

Mainly we take this information from our partners in the country, like our collaborators. The information it is always on the Internet, sector magazines, fairs, etc. We have get information from our clients.

5) How do you aggregate/summarise/collect/store these information?

We use Excel sheets and the collect it in our internal Server from the company so all the team can access the information.

D Appendix: Initial lists of sources (use-case descriptions)

D.1 PUC1: Media Monitoring – Journalistic Scenario

Journalistic Use Case – Topic: Energy Policy

This list is not conclusive

Source	languages	type/level of data
Major News outlets:		
SZ/ FAZ/ Die Zeit/ Spiegel	de	All articles on ‚Energie'
Le Figaro/ Le Monde/ Liberation	fr	All articles on ‚énergie'
El Mundo/ El Pais	es	All articles on ‚energía'
The Guardian/ The Times/ The daily telegraph	eng	All articles on Energy
Reuters/ AFP/ dpa/ agencia efe	eng/de/fr/es	All articles on Energy
…		
Governmental Authorities		
German Ministry for Energy (and more)	eng/ger/fr	All on energy, press materials, mediathek
Agency for the Cooperation of Energy Regulators	eng	All available items
German Ministry for the Environment	eng/ger	All on climate & energy, video material
French Ministry for Energy (and more)	fr	All items on Energy
European Comission: Energy	eng/de/fr	All available items
Spanish Ministry for Energy (and more)	es/eng	All available items
…		
NGOs/International organisations		
UN Energy Knowledge Network	eng	All items on Energy
International Association of Energy Economics	eng/fr/de/es	All available items
Gesellschaft für Energiewissenschaft und Energiepolitik e. V.	de/eng	All available items
Verein für ökologisch-solidarische	de/eng	All items on energy

Energie- & Weltwirtschaft e.V.		
Climate Action Network Europe	eng	All items on Energy
Öko-Institut e.V.	de/eng	All items on Energy
Bundesverband Erneuerbare Energien (BEE)	de/eng	All items on Energy
…		
Academia		
Regional Center for Energy Policy Research		All available items
Central European University - Center for Climate Change and Sustainable Energy Policy		All available items
Central European University – Energy Policy Research Group		All available items
European Energy Research Alliance	eng	All items on Energy
Florence School of Regulation - Energy	eng	All available items
Prof. Dr. Lorenz Jarass	de/eng	All items on energy
…		
Blogs		
Clean Technica		All items on energy
Council on foreign relations – Energy, Security and Climate		All items on energy
Volker Quaschning	de/eng	All items on Energy/Energie
Smart Grid Watch Blog	eng	All available items
Social Media		
Twitter (Hashtags)	eng/de/fr/es	#energy #renewables #greenenergy #energiewende #EEG
Facebook (keywords)	eng/de/fr/es	Energy, renewable, energiewende
Reddit (keywords)	eng/de/fr/es	Energy, renewable, energiewende

D.2 PUC1: Media Monitoring – Commercial Scenario

Media Monitoring Use Case – Topic: Household Appliances

This list is not conclusive

Source	Languages	Type/Level of Data
Major News Outlets		
The Guardian/ The Times/ The daily telegraph El Mundo/ El Pais Le Figaro/ Le Monde/ Liberation SZ/ FAZ/ Die Zeit/ Spiegel Reuters/ AFP/ dpa/ agencia efe	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
…		
Popular Press		
Brigitte / Bunte / Elle / Elle Decoration / Glamour / InStyle / Jolie / GQ / Ideal Home / Architectural Digest	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
…		
Trade Press		
Elektrojournal / Küchenmagazin / elektronik journal / ERT / Get Connected / Independent Electrical Retailer / appliancist.com	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
…		
Blogs		
decor8blog.com / greenlivingonline.com/blog / diisign.com / blog.hospimedia.fr	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
…		
Social Media		

Twitter (Hashtags)	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
Facebook (keywords)	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
YouTube (keywords)	eng/de/fr/es/bg	All articles on a predefined list of products and brand names, e.g. AEG or Kitchenaid
…		

D.3 PUC2: SMEs internationalisation

This list is not conclusive

Source	languages	type/level of data
Market Analysis		
Spanish Office of Economy and Competitiveness	ESG	General Information about Germany(location, size, climate, demography, society, a little bit of history)
Spanish Office of Economy and Competitiveness	ESG	Practical Information about how to access the market, visas, city websites, bank holidays, banks,...
Central Intelligence Agency – The world factbook	ENG	Economic overview, info about sectors, GDP, unemployment rate, labor force, imports, exports, debt, exchange rates,…
Central Intelligence Agency – The world factbook	ENG	Transnational Issues important to have in mind.
Trading Economics	ENG	Evolution on multiple indicators (Markets, GDP, Labour, Prices, Money, Trade, Government, Business, Consumer, Taxes & Housing)

Germany's website	ENG	Link to pages with interesting information about different matters regarding Germany.
Spanish Office of Economy and Competitiveness	ES	General Information about Germany (location, size, climate, demography, society, a little bit of history)
Centro Alemán de Información para Latinoamérica y España – Spanish Center of Information for Spain and Latin America	ES	Foreign policy
Ministerio de Asuntos Exteriores y Cooperación – Spanish Foreign Minister	ES	Bilateral relations
Federal Ministry for Economic Affairs and Energy	ENG/GER	Quantitative analysis: foreign trade
EUROSTAT	ENG/GER/FR	Quantitative analysis: macroeconomic data
German Trade and Invest	ENG	Segmentation by food and beverage in Germany
Kane Juliane Schröder: Cannibalization on the yoghurt market	ENG	German yoghurt market
Industry Analysis: Competitors	ENG/GER	DANONE (18%) EHRMANN AG (12,8%) MÜLLER
Regulation		
Europa – summaries of EU legislation	ES/ENG/GER/…	Product labeling and packaging
International Dairy Food Association	ENG	European health certification program
German Business Portal	ENG	Overview of market access of food and beverage
IFS	ENG/GER	IFS Food packaging guideline
Summaries of EU legislation	ENG	Labelling, presentation and advertising of foodstuffs
Singapure Government	ENG	General requirements and standards

		for food and agricultural imports into Germany
Blogs		
Banco Sabadell	ESP	General information about economic and finantial issues
Social Media		
Twitter (Hashtags)	ENG/GER/FR/ES	#yoghurt #nonfat #healthy #food #mango #lowfat #fitness #sogood #iogo #breakfast #vanillayogurt #healthybreakfast